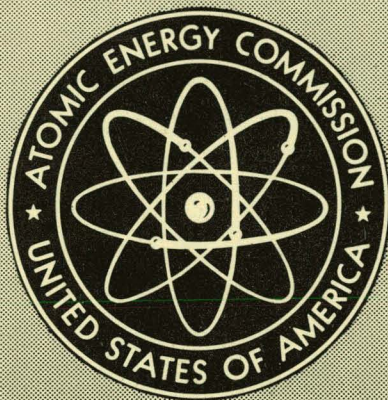


MASTER



UCRL-8816

LECTURES ON PHYSICS, BIOPHYSICS, AND CHEMISTRY FOR HIGH SCHOOL SCIENCE TEACHERS

Given at the Ernest O. Lawrence Radiation Laboratory,
Berkeley, California, June–August 1959

Edited by
Edward C. Calhoon
Paul W. Starring

August 1959

University of California
Berkeley, California

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

This report has been reproduced directly from the best available copy.

Printed in USA. Price \$1.50. Available from the Office of Technical Services, Department of Commerce, Washington 25, D. C.

LECTURES

ON

PHYSICS, BIOPHYSICS, and CHEMISTRY
FOR HIGH SCHOOL SCIENCE TEACHERS

Given at the

Ernest O. Lawrence Radiation Laboratory
Berkeley, California

June - August 1959

Edited by

Edward C. Calhoon and Paul W. Starring

August 1959

LAWRENCE RADIATION LABORATORY

University of California

Berkeley

January 12, 1961

ERRATA

FROM: Technical Information Division

SUBJECT: UCRL-8816, "Lectures on Physics, Biophysics, and Chemistry for High School Science Teachers"

P. 5. Equation (17) should read $\cos \alpha = \frac{c/\eta}{\beta c}$

$$\cos \alpha = \frac{1}{\beta \eta}$$

P. 16. Col. 1, 3rd paragraph, line 3: should read "chance of winning three dollars is 1 in 12"
Same, line 6: should read " $P = 1/4 + 1/12$ "

P. 58. Table E-II: last two columns should read

<u>Mass^b (Mev)</u>	<u>Mean life (sec)</u>
0	stable
0.51	stable
106	2.2×10^{-6}
140	2.6×10^{-8}
135	10^{-16} to 10^{-15}
494	1.2×10^{-8}
498	10^{-10}
498	6×10^{-8}
938	stable
940	1040

P. 58. Table E-II (continued)

Mass ^b (Mev)	Mean life (sec)
1115	2.5×10^{-10}
1192	10^{-19}
1190	0.7×10^{-10}
1196	1.6×10^{-10}
1311	1.5×10^{-10}
1319	1.9×10^{-10}

- P. 71. In Fig. H-5 the departure of the pole pieces from flatness is exaggerated. It should appear qualitatively similar to Fig. H-7 which depicts the opposite kind of departure from flatness.
- P. 72. Col. 1, 2nd paragraph, line 19: should read "and be directed so as to oppose the displacement..."
- P. 72. Col. 2, line 7: should read "V. I. Veksler in Russia and independently by E. M. McMillan,..."
- P. 72. Col. 2, paragraph 3, last line: should read "accelerators."
- P. 72. Col. 2, paragraph 4, first word should read "Since"
- P. 82. Fig. J-1: Replace word "photon" with "relativistic particles."
- P. 97. Fig. L-5: Delete entire sentence beginning "The ellipses are... θ^0)." "
- P. 112. Col. 1, paragraph 3, line 9: should read "a human ever touching the materials."
- P. 115. Name of author of this section should read "H. Wade Patterson"
- P. 116. Col. 1, under Maximum Permissible Exposure, delete entire first two sentences. Replace with
- "The present maximum permissible occupational exposure is age-prorated and is given by $[(A-18)/5]$ rem, where A is the age in years. However, the exposure must not exceed 3 rem in any one-quarter of a year, or 12 rem in any one year."
- P. 116. Col. 1, lines 3-4: should read "sources have a range of a few meters in air. They can penetrate a few..."
- P. 116. Col. 2, lines 1-7: should read "received over the whole body. The maximum permissible exposure to the general population is 0.5 rem per year. The amount of radiation, as measured in rads, to produce a 0.1-rem exposure depends upon the type of radiation. Table P-II lists the equivalent exposures in rads and rem for various radiations."

- P.116. Table P-II, change heading of 2nd column "rep" to "rad"
- P.116. Col. 2, paragraph 1, line 6: should read "In 1928 the maximum"
Same, line 9: delete entire sentence, "It is now set a 5 rem/yr."
- P.116. Col. 2, Table P-III: Heading should read "Effect vs dose for full-body radiation received in a few hours"
- P.116. Col. 2, paragraph 2, line 2: should read "given full-body exposure given in a few hours."
- P.116. Col. 2, paragraph 3, line 6: should read, "not carefully measured. In recent years much care..."
- P.117. Col. 1, line 9: should read "eventually developed. It was possible..."
- P.117. Col. 1, paragraph 3, lines 12-13: should read "It is very difficult to separate effects due to neutrons and gamma rays in the Japanese cases, and it is..."
- P.117. Col. 1, under Shielding, line 6: should read "A useful relation for point sources is..."
- P.117. Col. 2, line 5: should read "whose intensity I_0 at 1 cm is 8400 millirads/hr. A..."
- P.117. Col. 2, equation should read

$$I = \frac{I_0 d_0^2}{d^2}$$

$$= \frac{8400}{100^2}$$

$$I = 0.84 \text{ millirads/hr,}$$

- P.117. Col. 2, paragraph 1, line 2: should read "ing with point neutron source is..."
- P.117. Col. 2, paragraph 2, lines 6-7: should read "outlined in Section C, "Nuclear Physics." The estimated attenuation of shielding is generally increased in practice by..."
- P.140. Col. 2, last paragraph, line 4: should read "cannot routinely detect this narrowing,..."
- P.141. Col. 1, paragraph 3, delete first sentence and replace it with
"We also found that of these three factors, the first two are independent significant indicators of risk. Overweight is found significant but not independent of blood pressure and lipoprotein values. Thus, the overweight risk contribution is primarily due to the association, on the average, between overweight and elevated blood pressure and lipoprotein values."

- P. 141. Col. 2, line 2: should read "red blood cells. Above this we would normally see a clear..."
- P. 141. Col. 2, lines 5-6: should read "we are principally interested in, because it contains the lipoproteins and their constituent lipids. The lipids are extracted from serum by..."
- P. 141. Col. 2, line 9: should read "cholesterol esters, cholesterol, phospholipids, and triglycerides."
- P. 141. Col. 2, line 10: should read "These materials (except cholesterol) are chemical compounds..."
- P. 141. Col. 2, paragraph 3, line 3: should read "aqueous solvent (such as serum)."
- P. 142. Col. 1, line 7: should read "a major stabilizing factor for fats in the blood is..."
- P. 142. Col. 1, paragraph 1, line 4: should read "Salt ions can cruise around..."
- P. 142. Col. 2, legend to Fig. T-6, under (b): should read (b) Electron micrographs of specific lipoprotein fractions. In this report the $S_{f20-400}$ group is referred to as "large" lipoproteins; the S_{f10-20} and S_{f6-8} groups as "intermediate-sized" lipoproteins; and HDL₂₊₃ as "small" lipoproteins. The S_f designations refer to rate of migration in the ultracentrifuge."
- P. 143. Col. 1, line 3: should read "cholesterol, cholesterol esters, and protein are present."
- P. 143. Col. 1, lines 7-9: after "the cholesterol esters" delete hyphen and remainder of sentence to the period.
- P. 143. Col. 1, paragraph 2, line 2: should read "large and intermediate-sized lipoproteins (which transport..."
- P. 143. Col. 2, line 13: should read "fats are associated with elevated concentrations of the..."

LECTURES ON PHYSICS, BIOPHYSICS, and CHEMISTRY
FOR HIGH SCHOOL SCIENCE TEACHERS

Edited by

Edward G. Calhoon and Paul W. Starring

Lawrence Radiation Laboratory
University of California
Berkeley, California

August 1959

PREFACE

. During the summers of 1955, 1956, and 1957 the Lawrence Radiation Laboratory conducted an institute for outstanding high school science teachers. Omitted in 1958, the institute was again held this summer with 18 teachers participating. Each teacher was assigned to one of the research groups in nuclear physics, biophysics, or nuclear chemistry. Combined with this practical experience was a series of lectures by members of the Laboratory staff. These lectures outlined the theories and principles underlying many phases of the research performed here. In some cases new techniques and instruments were described and recent findings were discussed.

The participating teachers also attended a number of lectures on related subjects given on the Berkeley campus as a part of the Summer Institute for Teachers of Science sponsored by the National Science Foundation. The following lectures were included in that group:

Bubble Chambers
High-Energy Particle Accelerators
Elementary Particles
Antiparticles
Origin and Evolution of the Solar System
The Role of Space Satellites in Gathering Astronomical Data
Chromatography
Chemistry and Physics of Virus
Biology of Virus
Radiation and Life in Space

The reports contained in this volume were compiled by the editors from the tape-recorded lecture proceedings. So that the collection would fit into a single volume, many illustrative examples and supporting facts had to be omitted. Consequently, these reports should not be regarded as definitive articles, but rather as summaries or notes. Each lecturer reviewed the material appearing under his name for factual accuracy.

Two exceptions to the above procedure should be mentioned. The report "Photosynthesis" by Dr. John Bassham is a condensation of a paper that he had prepared for the Journal of Chemical Education. The paper "Radiation and Life in Space" by Dr. Cornelius Tobias was presented by him at the 1959 Symposium of the Federation of American Societies for Experimental Biology and is to be published in the Federation Proceedings.

The editors wish to thank Dr. Harry H. Heckman, coordinator for the summer institute program at this Laboratory, for his advice and cooperation in preparing these lecture notes.

Edward Calhoon and Paul Starring

LECTURES ON PHYSICS, BIOPHYSICS, and CHEMISTRY FOR HIGH SCHOOL SCIENCE TEACHERS

Contents

Preface	vii
A. Mathematics Review	
1. Trigonometry	Harry H. Heckman 1
2. Vector Analysis	Roger W. Wallace 6
3. Elements of the Calculus	Harry H. Heckman 10
4. The Statistical Analysis of Data.	Harry H. Heckman 16
B. Atomic Physics (six lectures)	Roger W. Wallace 20
C. Nuclear Physics (six lectures)	Warren F. Stubbins 39
	Robert W. Kenney
D. Solid-State Physics	Roger W. Wallace 53
E. Elementary Particles	Robert L. Thornton 57
F. Antiparticles	Owen Chamberlain 61
G. Design of Experiments	Robert W. Kenney 66
H. High-Energy Particle Accelerators	David L. Judd 70
I. Survey of Particle Detectors.	Roger W. Wallace 75
J. Emulsion As a Particle Detector	Harry H. Heckman 82
K. Counters Used in High-Energy Physics	LeRoy T. Kerth 88
L. Bubble Chambers	James D. Gow 94
M. Computer Programming	John Caris 99
N. Chromatography	Wilson Johnson 106
O. The Transuranium Elements	Richard M. Diamond 110
P. Health Physics (two lectures)	Wade Patterson 115
Q. Photosynthesis	John A. Bassham 119
R. The Chemistry and Physics of Virus	Robley C. Williams 125
S. The Biology of Virus	Gunther S. Stent 136
T. Lipoproteins and Heart Disease	Alexander V. Nichols 140
U. Origin and Evolution of the Solar System	Otto Struve 144
V. The Role of Space Satellites in Gathering Astronomical Data	Otto Struve 146
W. Radiation and Life in Space	Cornelius A. Tobias 149
Acknowledgment	

A-1. TRIGONOMETRY

Harry H. Heckman

In past years, the lecture series included no review of mathematics. There subsequently were instances when a simple mathematical symbol was used which lacked meaning until time had been taken to define it and explain its derivation. This year three lectures will be devoted to reviewing some basic mathematical concepts and notations. These will fall into the categories of trigonometry, vector analysis, and the calculus. In each category the lecturer will give a pointed example of how it is used in the analysis of some physical phenomenon.

When one performs an experiment in physics, measurements are inevitably taken of such quantities as distances and angles. Our first lecture is on trigonometry, that portion of mathematics which deals with angles. If certain parts of a triangle are known, trigonometry makes possible the calculation of other parts.

Let us first consider the properties of a triangle. (Fig. A-1) We label the corners, or vertices, A, B, and C, and the angles α , β , and γ respectively. The first property of the triangle is that the sum of its angles equals 180° :

$$\alpha + \beta + \gamma = 180^\circ. \quad (1)$$

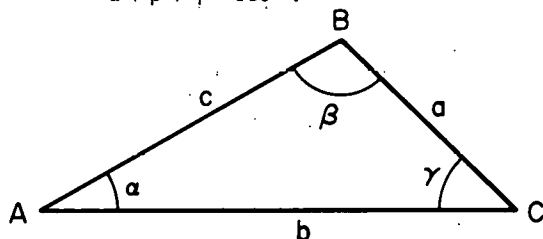


Fig. A-1

This is rather easy to demonstrate. Extend the base b, and by construction draw a line d from A parallel to side a (Fig. A-2). The angle between line d and

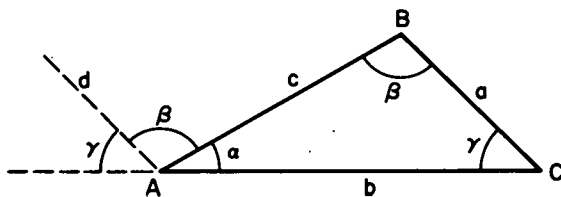


Fig. A-2

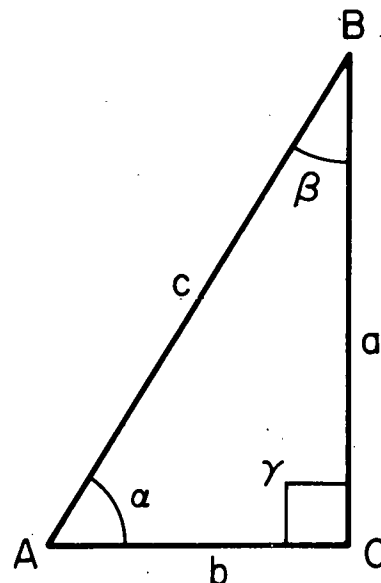
the extension to line b is then equal to γ ; the angle between lines d and c is equal to β . Thus, we see that the angles α , β , and γ add up to make a straight line, or form an angle of 180° .

The Right Triangle

The most important figure to trigonometry is the right triangle. In this triangle (Fig. A-3) one of the angles equals 90° . Since γ is a right angle, the sum of the other two angles is 90° :

$$\alpha + \beta = 90^\circ. \quad (2)$$

Fig. A-3



The Pythagorean theorem describes a fundamental property of the right triangle:

$$a^2 + b^2 = c^2. \quad (3)$$

The truth of this theorem can be demonstrated most simply by drawing a diagram like that shown in Fig. A-4. In this drawing, we have a small square drawn inside a larger square, so that all the a's are equal, all the b's, and all the c's. The area A of the larger square may be found by adding up the areas of its various parts:

$$\begin{aligned} A &= c^2 + ab/2 + ab/2 + ab/2 + ab/2 \\ &= c^2 + 4(ab/2) \end{aligned}$$

$$A = c^2 + 2ab. \quad (4)$$

Another way to find its area is to write:

$$\begin{aligned} A &= (a + b)^2 \\ A &= a^2 + 2ab + b^2. \end{aligned} \quad (5)$$

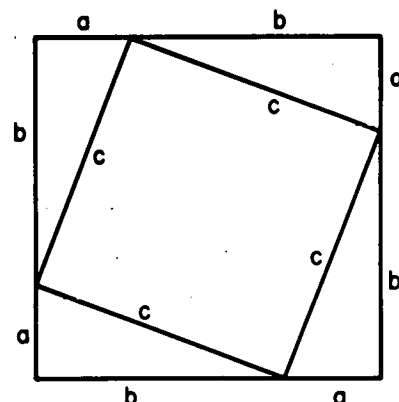


Fig. A-4

Setting the two expressions for A i. e., Eqs. (4) and (5) equal to each other, we obtain

$$a^2 + 2ab + b^2 = c^2 + 2ab.$$

Subtracting $2ab$ from each side yields

$$a^2 + b^2 = c^2.$$

Another property of a right triangle is that of similarity. If one of the angles in a right triangle is equal to an angle in another right triangle, then the two triangles have the same shape, or are said to be similar (Fig. A-5). That is, the lengths of corresponding sides are proportional, and the ratio of any two corresponding sides, say a and a' , is a constant:

$$\frac{a}{a'} = \frac{b}{b'} = \frac{c}{c'} = \text{constant.} \quad (6)$$

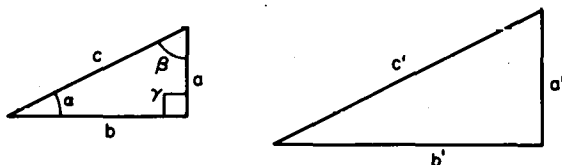


Fig. A-5

If, in addition, the sides b and b' are equal, the triangles are then the same size and shape, and are said to be identical or congruent (Fig. A-6). This notion of similarity and congruency applies to oblique as well as right triangles.

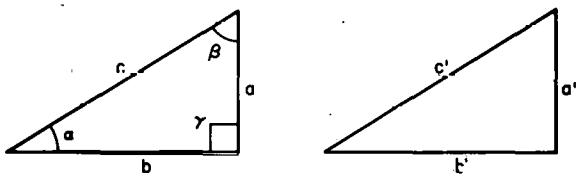


Fig. A-6

Measuring Angles

So far we have discussed angles without describing how they may be measured. To define an angle, we must first draw a circle (Fig. A-7). We next draw a radius and rotate it from Point A to Point B. Angle α may now be measured by the length AB along the circumference. If the length AB just equals the radius, we say that the angle subtended is one "radian." This type of angular measurement is called "radian measurement."* An angle which makes a complete revolution is equal to the circumference, or $2\pi r$. In a unit circle (a circle in which the radius is defined as a unit length), this length is simply 2π . Such an angle equals 2π radians.

*We call α positive when we rotate the radius in the counterclockwise direction, and negative when we rotate it in the clockwise direction.

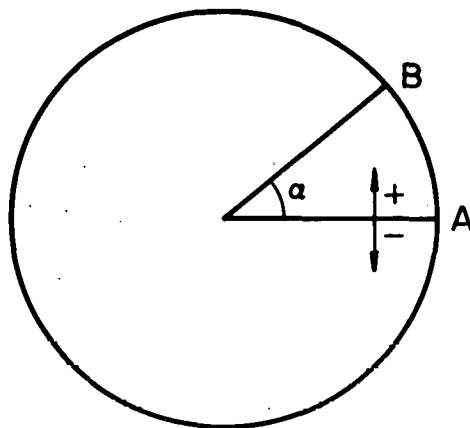


Fig. A-7

An angle which makes one half a revolution (180°) equals $2\pi/2$, or π radians. Table A-I lists the equivalents for other angles.

Table A-I

Equivalence between angles measured in radians and in degrees

Angle (degree)	Angle (radian)
30	$\pi/6$
45	$\pi/4$
57.3	$\pi/\pi=1$
90	$\pi/2$
180	π
270	$3\pi/2$
360	2π
540	3π
720	4π

From Table A-I we see that an angle can sweep through more than one revolution. Note also that one radian equals 57.3° . It is useful to remember this number when converting from radians to degrees.

Keep in mind that any system of measuring angles is purely arbitrary. It would be well if we could construct these units, such as the degree or the radian, a priori by means of straight edge and compass. For 2000 years mathematicians have attempted to do this, but unfortunately it is impossible. The German mathematician and physicist Karl F. Gauss gave proof for this impossibility in 1796.

Trigonometric Ratios

A practical method for measuring an angle is simply to measure a side of a right triangle whose base b is a radius of a unit circle (Fig. A-8). The side a is drawn tangent to the circle at point C. Since we have $b = 1$, we can define the angles α or β simply by measuring either a or c , or by giving a ratio of one of the sides to another. It is convenient to define the ratios given in Table A-2, Column I.

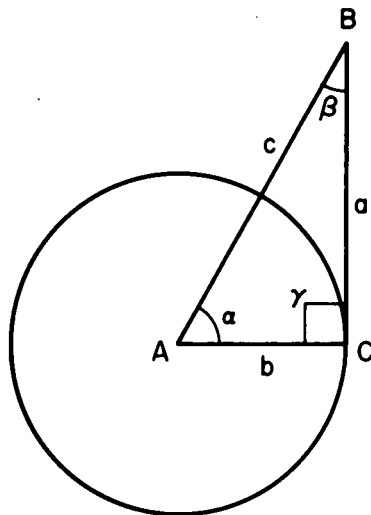


Fig. A-8

Table A-II

Trigonometric functions for a right triangle	
I	II
Tangent (tan) $\alpha = a/b$	$= \cot \beta$
Sine (sin) $\alpha = a/c$	$= \cos \beta$
Secant (sec) $\alpha = c/b$	$= \operatorname{cosec} \beta$
Cotangent (cot) $\alpha = b/a$	$= \tan \beta$
Cosine (cos) $\alpha = c/a$	$= \sin \beta$
Cosecant (cosec) $\alpha = b/c$	$= \sec \beta$

The terms in Column II were deduced from Fig. A-8. Note that the "cofunctions" of the angle α are simply related to the trigonometric functions of angle β , in that $\beta = 90^\circ - \alpha$.

Graphs of the Trigonometric Functions

The next question to arise is, "How does the value of one of these functions change with the angle, i. e., how does the sine of α vary with α ?" This is the function most encountered by students in electronics or those who must otherwise use alternating currents. We draw a unit circle and see how the ratio a/c (i. e. $\sin \alpha$) varies with α (Fig. A-9). We see

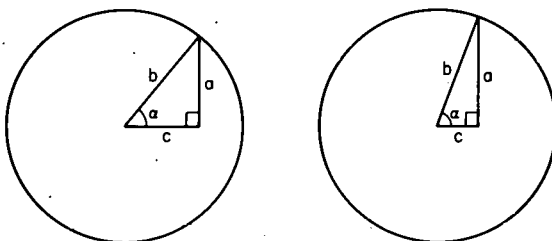


Fig. A-9

that for $\alpha = 0^\circ$, a is 0 and c is 1; therefore, we have $\sin \alpha = 0$. As α increases, the ratio a/c increases. At $\alpha = 90^\circ$, both a and c are 1. As α passes 90° , the ratio a/c decreases, until at $\alpha = 180^\circ$, a/c is zero. As α passes 180° , a becomes

negative (c is always positive, by definition), and the ratio a/c becomes larger. When we plot this, we obtain the curve shown in Fig. A-10.

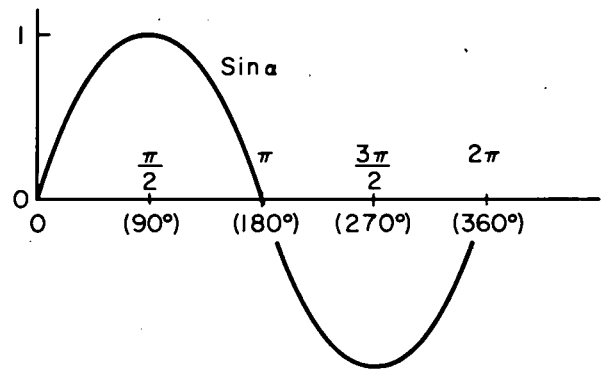


Fig. A-10

The curve of $\cos \alpha$ is shown in Fig. A-11. It should be noted that the curve for $\cos \alpha$ is the same as for $\sin \alpha$, except that it is shifted in phase (i. e., displaced) by 90° .

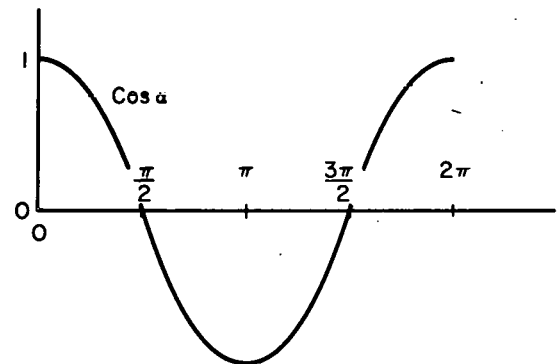


Fig. A-11

We next plot $\tan \alpha$ (i. e., a/b) (Fig. A-12). As α approaches $\pi/2$, a approaches 1 and b approaches 0. Thus, the ratio a/b becomes infinitely large. However, if $\pi/2$ is approached from the other direction (i. e., from 180°), the ratio a/b becomes infinitely large but in a negative sense. Thus, at $\pi/2$ and $3\pi/2$ the curve is discontinuous.

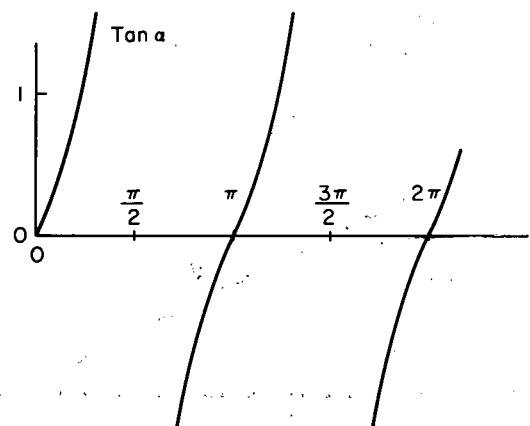


Fig. A-12

Trigonometric Identities

We must consider some of the fundamental identities of the trigonometric functions. These stem from the basic definitions already given. They are of invaluable assistance in the solution of problems, and so have been tabulated and are available as a reference in trigonometry books. The simplest identities are the "inverse identities:"

$$\sin a = 1/\operatorname{cosec} a \quad (7)$$

$$\cos a = 1/\sec a \quad (8)$$

$$\tan a = 1/\cot a. \quad (9)$$

Another simple identity is

$$\sin a/\cos a = \frac{a/c}{b/c} = a/b$$

$$\sin a/\cos a = \tan a. \quad (10)$$

The Pythagorean identities stem from the Pythagorean theorem

$$a^2 + b^2 = c^2. \quad (11)$$

Dividing this by c^2 , we obtain

$$\begin{aligned} a^2/c^2 + b^2/c^2 &= c^2/c^2 \\ (a/c)^2 + (b/c)^2 &= 1 \end{aligned}$$

or

$$\sin^2 a + \cos^2 a = 1. \quad (12)$$

In a similar way we divide Eq. (11) by b^2 to obtain

$$\begin{aligned} a^2/b^2 + b^2/b^2 &= c^2/b^2 \\ \tan^2 a + 1 &= \sec^2 a. \end{aligned} \quad (13)$$

Dividing Eq. (11) by a^2 , we get

$$\begin{aligned} a^2/a^2 + b^2/a^2 &= c^2/a^2 \\ 1 + \cot^2 a &= \operatorname{cosec}^2 a. \end{aligned} \quad (14)$$

The Oblique Triangle

The above are applicable to the right triangle owing to its special properties. It is essential now that we consider the more general case of the oblique triangle (Fig. A-13), in which no angle is 90° . A very beautiful and symmetric relationship exists here which we call the law of sines:

$$\frac{\sin a}{a} = \frac{\sin \beta}{b} = \frac{\sin \gamma}{c}. \quad (15)$$

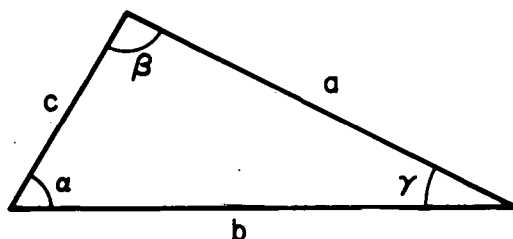


Fig. A-13

This relationship applies to all triangles including right triangles. Also we have the law of cosines:

$$\cos a = \frac{b^2 + c^2 - a^2}{2bc}$$

or

$$a^2 = b^2 + c^2 - 2bc \cos a, \quad (16)$$

which enables one to calculate the length of the third side of a triangle when the length of the other two sides and their included angle are known.

An Application of Trigonometry to High-Energy Physics

A striking example of the use of trigonometry in high-energy physics is in the determination of the velocity of a particle by means of its Cherenkov radiation. (It might be recalled that Cherenkov, a Russian physicist, recently received the Nobel Prize for the discovery and interpretation of this radiation.) Consider a charged particle (e.g. an electron) traveling with a velocity nearly equal to that of light, i.e., at $\beta \approx 1$. The quantity β is defined as the particle velocity v divided by the velocity of light c ($\beta = v/c$). Let the particle enter a transparent material like Lucite, in which light travels at a velocity c_1 . The velocity c_1 is given by

$$c_1 = \frac{c}{\eta}$$

where η is the index of refraction of Lucite ($\eta=1.5$) and c is the velocity of light in a vacuum. Here we have a situation in which the particle is traveling through the Lucite faster than light travels through the same medium. Since the particle is traveling faster than energy can be carried away, a shock wave is set up which is analogous to that created when a jet plane flies faster than the speed of sound in air.

Figure A-14 represents the way in which Cherenkov radiation is formed. At a given instant the particle is at the point P. A certain time earlier it was at P' . The light waves emitted at point P' have traveled outward, as represented by the circle surrounding P' . At a still earlier time the particle was at P'' . The circle around P'' indicates the wave front of light emitted at that point. The waves from successive points P' , P'' , etc., establish a coherent conical wave front which travels in the direction of the dashed arrows. This shock wave is called Cherenkov radiation and is visible as a bluish light.

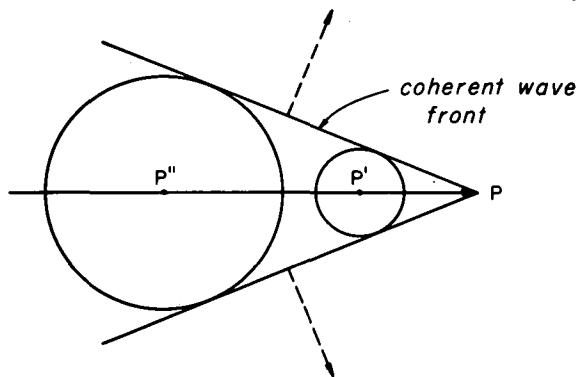


Fig. A-14

The angle α with which the shock wave moves relative to the particle path is a measure of the particle velocity (Fig. A-15). The greater the velocity, the smaller is the angle. In Fig. A-15 the particle is moving to the right at velocity βc ; the wave front is proceeding at a velocity c/η . From our definitions of the trigonometric functions we have

$$\begin{aligned}\cos \alpha &= \frac{c/\eta}{\beta c} \\ \cos \alpha &= \frac{1}{\beta \eta} \\ \text{or} \quad \beta &= \frac{1}{\eta \cos \alpha}\end{aligned}\quad (17)$$

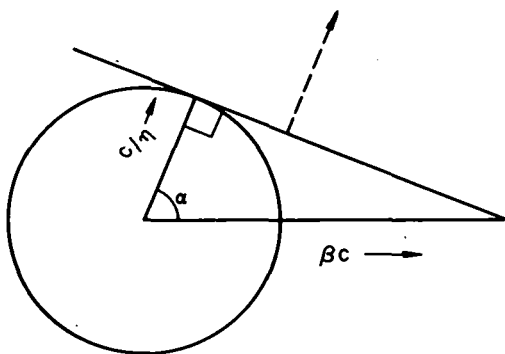


Fig. A-15

Suppose that in an experiment we observe the angle α to be 48° for a particle traveling through Lucite ($\eta = 1.5$). Consulting a table of the trigonometric functions, we find $\cos 48^\circ = 0.669$. Solving Eq. (17), we obtain

$$\beta = \frac{1}{1.5(0.669)}$$

$$\beta \approx 0.98.$$

Thus, the particle is traveling at a velocity of about 98% of the speed of light.

In practice, a series of optical vanes or "Venetian blinds" may be set up so that only the Cherenkov radiation of a given angle α can be observed. The detection equipment then will not respond to particles having greater or smaller velocities.

This is one of the techniques which made possible the discovery of the antiproton (negative proton) at this Laboratory in 1955. A beam from the Bevatron was passed through a number of steering magnets, so that particles reaching the Cherenkov counter had essentially the same momentum (product of mass times velocity). The beam consisted primarily of pi mesons (light unstable particles having about one-seventh the mass of an antiproton) and perhaps 1/100% antiprotons. Since pi mesons are much lighter than antiprotons, they must travel at a greater velocity in order to have the same momentum.

In this experiment the pi mesons had a velocity $\beta \approx 0.99$. The β for antiprotons, if they existed, would be $\beta \approx 0.78$. Thus the antiprotons would give off Cherenkov radiation at an angle α of about 31° , while α for pi mesons would be about 48° . This difference was sufficient for identifying the particles.

Cherenkov counters can be constructed to differentiate between particles having velocities differing by only about 3%. Since the effect is an electromagnetic phenomenon, only charged particles produce Cherenkov radiation. Neutral particles therefore cannot be detected by this method.

A-2. VECTOR ANALYSIS

Roger Wallace

Vector analysis may be viewed as a shorthand notation which is more convenient to use than conventional geometry. It is widely employed by physicists and electrical engineers for describing physical quantities, such as electric or magnetic fields, that have both a direction and a magnitude.

The word "vector" sometimes worries people who think it must stand for an extremely complicated notion. Quite the opposite is true: a vector is simply an arrow. Its length represents the magnitude of a physical quantity, like the number of pounds of force applied to an object; its direction or orientation indicates the direction the force is being applied.

Vectors are most often used in 2-dimensional or 3-dimensional coordinate systems, but they may be used in higher-ordered coordinate systems. For instance, sometimes it is desirable to add a fourth dimension such as time in certain equations.

First, we consider the 2-dimensional case (Fig. A-16). We draw an arrow (vector) from the origin of the coordinate axes to the point (a_x, a_y) . We label the arrow $\vec{a}$ to indicate that it is a vector, not a number. The length of the vector $\vec{a}$ is denoted simply by the letter a , or sometimes by $|\vec{a}|$, which are equivalent expressions. The latter expression ($|\vec{a}|$) reads "the absolute value of $\vec{a}$," and is a number not a vector. We use the Pythagorean theorem to find $|\vec{a}|$:

$$\begin{aligned} |\vec{a}|^2 &= a_x^2 + a_y^2 \\ |\vec{a}| &= \sqrt{a_x^2 + a_y^2} = a. \end{aligned} \quad (1)$$

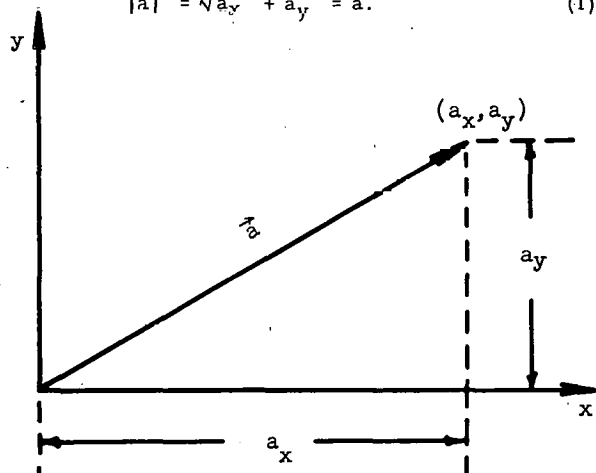


Fig. A-16

We now turn to the 3-dimensional case. (Fig. A-17.) It should be noted that we employ a "right-handed" coordinate system. This means that if the right hand is held so that the fingers point from the positive x axis toward the positive y axis, the thumb points in the direction of the positive z axis. Other conventions are sometimes used in various foreign countries, but the "right-handed" convention is followed in this country.

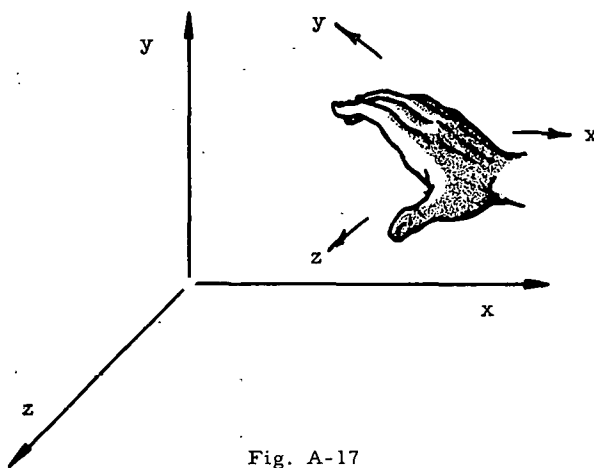


Fig. A-17

We draw a vector from the origin to the point (b_x, b_y, b_z) and call the vector $\vec{b}$ (Fig. A-18). (Any letter may be used; b was chosen arbitrarily). The length b of the vector is

$$b = \sqrt{b_x^2 + b_y^2 + b_z^2}. \quad (2)$$

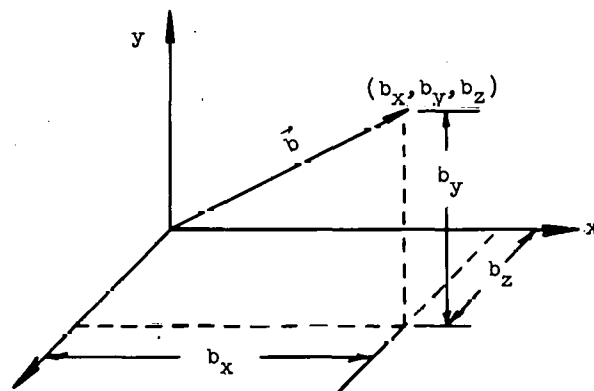


Fig. A-18

One advantage of vector notation should now be apparent - one symbol can stand for a large number of other symbols. For instance, the symbol $\vec{a}$ takes the place of the much longer notation ordinarily used in solid geometry, (a_x, a_y, a_z) . Another convenience of vector analysis is that we can define quite simply certain useful operations. These may involve rather complex ideas, and may be quite lengthy if written in conventional geometrical notation. These include vector addition, subtraction, multiplication, and division.

Vector Addition

Vector addition is illustrated in Fig. A-19. We have a number of vectors $\vec{a}, \vec{b}, \vec{c}, \vec{d}, \vec{e}, \vec{f}$, and $\vec{g}$. The vector sum is indicated by the vector $\vec{m}$:

$$\vec{m} = \vec{a} + \vec{b} + \vec{c} + \vec{d} + \vec{e} + \vec{f} + \vec{g}. \quad (3)$$

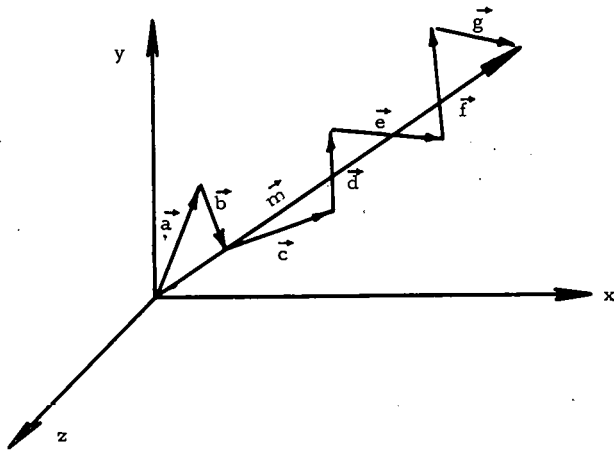


Fig. A-19

In the case shown in Fig. A-19, the length of m is:

$$m = \sqrt{h_x^2 + h_y^2},$$

where

$$h_x = a_x + b_x + c_x + d_x + e_x + f_x + g_x \text{ and}$$

$$h_y = a_y + b_y + c_y + d_y + e_y + f_y + g_y.$$

Vector Multiplication

Two types of vector multiplication are defined: the dot (scalar) product and the cross (vector) product.

Dot Product

The dot product $\vec{a} \cdot \vec{b}$ is defined as

$$\vec{a} \cdot \vec{b} = ab \cos \theta, \quad (4)$$

where θ is the angle between $\vec{a}$ and $\vec{b}$. (Fig. A-20.) It should be noted that the dot product $ab \cos \theta$ is a number or scalar, not a vector.

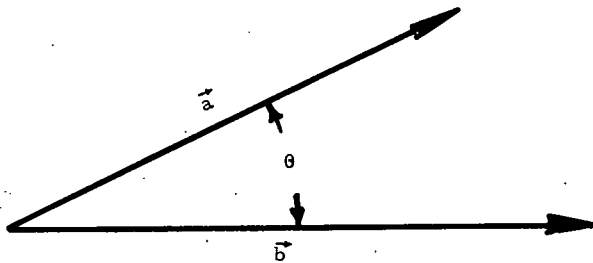


Fig. A-20

Several properties of the dot product are of interest:

a. The dot product commutes:

$$\vec{a} \cdot \vec{b} = \vec{b} \cdot \vec{a}. \quad (5)$$

b. If the dot product is zero, ($\vec{a} \cdot \vec{b} = 0$) then $\vec{a}$ and $\vec{b}$ are perpendicular. On the other hand, if $\vec{a}$ and $\vec{b}$ are perpendicular, their dot product is zero.

c. If the dot product $\vec{a} \cdot \vec{b}$ equals the product ab ($\vec{a} \cdot \vec{b} = ab$), then $\vec{a}$ and $\vec{b}$ are parallel. Conversely, if $\vec{a}$ and $\vec{b}$ are parallel, we have $\vec{a} \cdot \vec{b} = ab$.

d. In terms of the coordinates (a_x, a_y, a_z) and (b_x, b_y, b_z), the dot product is

$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y + a_z b_z. \quad (6)$$

e. The triple dot-product $\vec{a} \cdot \vec{b} \cdot \vec{c}$ is meaningless.

f. The derivative of ($\vec{a} \cdot \vec{b}$) with respect to t is

$$\frac{d}{dt} (\vec{a} \cdot \vec{b}) = \frac{d\vec{a}}{dt} \cdot \vec{b} + \frac{d\vec{b}}{dt} \cdot \vec{a}.$$

It should be noted that this is obtained by the conventional rules of differentiation, as is true of the derivatives of other vector functions.

Cross Product

The cross product $\vec{a} \times \vec{b}$ is defined as

$$\vec{a} \times \vec{b} = \overrightarrow{\text{absine}}, \quad (7)$$

where the quantity $\overrightarrow{\text{absine}}$ is a vector. The cross product has the following properties:

a. The cross product does not commute;

instead we have

$$\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}. \quad (8)$$

b. In terms of the coordinates (a_x, a_y, a_z) and (b_x, b_y, b_z), the cross product is

$$\begin{aligned} \vec{a} \times \vec{b} = & (a_y b_z - a_z b_y) \vec{i} + (a_z b_x - a_x b_z) \vec{j} \\ & + (a_x b_y - a_y b_x) \vec{k}, \end{aligned} \quad (9)$$

where $\vec{i}$, $\vec{j}$, and $\vec{k}$ are unit vectors in the x , y , and z directions respectively (Fig. A-21). Written in matrix notation this is

$$\begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix}$$

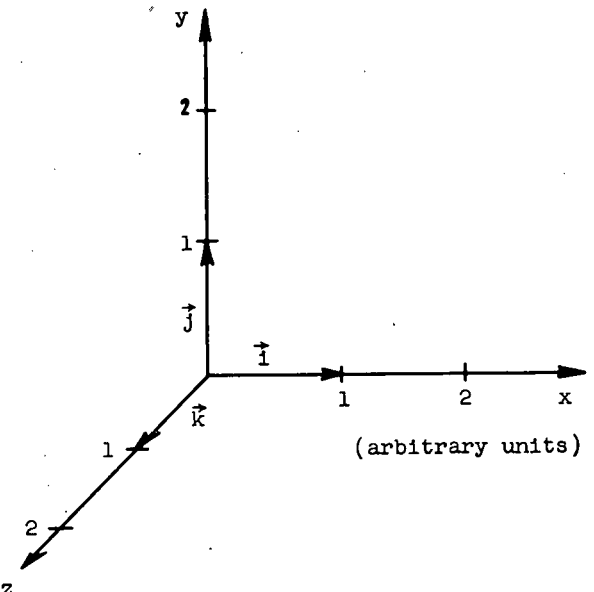


Fig. A-21

Another way to write Eq. (9) is

$$\vec{a} \times \vec{b} = c\vec{i} + d\vec{j} + e\vec{k}, \quad (9a)$$

where we have $c = (a_y b_z - a_z b_y)$, $d = (a_z b_x - a_x b_z)$, and $e = (a_x b_y - a_y b_x)$.

The absolute value of the cross product, $|\vec{a} \times \vec{b}|$, is equal in magnitude to the area of the parallelogram drawn in Fig. A-22.

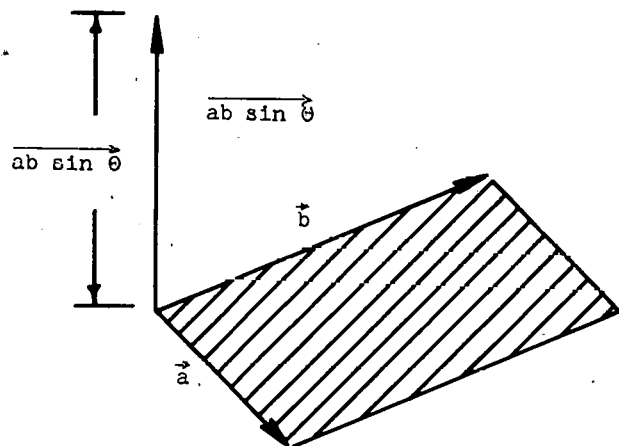


Fig. A-22

The vector $\vec{a} \times \vec{b}$ itself is perpendicular to the plane of $\vec{a}$ and $\vec{b}$.

d. When performing the triple cross-product $\vec{a} \times \vec{b} \times \vec{c}$, one must decide on the order of operations; that is, the product can be formed as either $(\vec{a} \times \vec{b}) \times \vec{c}$ or $\vec{a} \times (\vec{b} \times \vec{c})$, which are not the same. We have

$$\vec{a} \times (\vec{b} \times \vec{c}) = -\vec{a} \times (\vec{c} \times \vec{b}) \quad (10)$$

$$= (\vec{c} \times \vec{b}) \times \vec{a} \quad (11)$$

$$= -(\vec{b} \times \vec{c}) \times \vec{a}. \quad (12)$$

It should be noted that the triple cross product is a vector.

e. The derivative of $[\vec{a} \times (\vec{b} \times \vec{c})]$ with respect to t is given by

$$\begin{aligned} \frac{d}{dt} [\vec{a} \times (\vec{b} \times \vec{c})] &= \frac{d\vec{a}}{dt} \times (\vec{b} \times \vec{c}) + \vec{a} \times \left(\frac{d\vec{b}}{dt} \times \vec{c} \right) \\ &\quad + \vec{a} \times \left(\vec{b} \times \frac{d\vec{c}}{dt} \right). \end{aligned}$$

Mixed dot and cross product

For the mixed dot and cross product, we have the following relationships:

$$\vec{a} \cdot \vec{b} \times \vec{c} = \vec{a} \times \vec{b} \cdot \vec{c} \quad (13)$$

$$= \vec{b} \cdot \vec{c} \times \vec{a} \quad (14)$$

$$= \vec{b} \times \vec{c} \cdot \vec{a} \quad (15)$$

$$= \vec{c} \cdot \vec{a} \times \vec{b} \quad (16)$$

$$= \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix}.$$

The mixed dot and cross product $\vec{a} \times \vec{b} \cdot \vec{c}$ is a number or scalar whose magnitude is the same as the volume of the prism shown in Fig. A-23. The derivative of $(\vec{a} \cdot \vec{b} \times \vec{c})$ with respect to t is given by

$$\begin{aligned} \frac{d}{dt} (\vec{a} \cdot \vec{b} \times \vec{c}) &= \frac{d\vec{a}}{dt} \cdot \vec{b} \times \vec{c} + \vec{a} \cdot \frac{d\vec{b}}{dt} \times \vec{c} \\ &\quad + \vec{a} \cdot \vec{b} \times \frac{d\vec{c}}{dt}. \end{aligned}$$

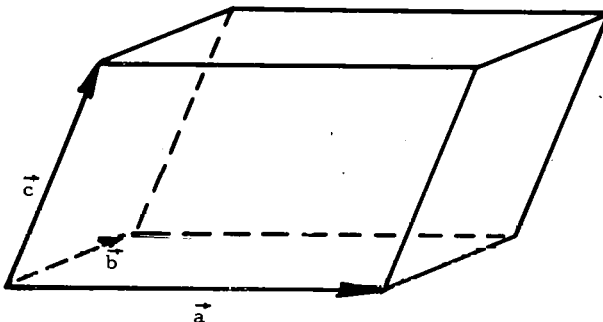


Fig. A-23

Applications of Vector Analysis to Electrical Theory

The most common application of vector analysis is to electrical theory. As an illustration, we will consider the question of the force exerted on a charged particle such as a proton when it is in an electric field. The force $\vec{F}_{\text{elec}}$ is found from this expression

$$\vec{F}_{\text{elec}} = e\vec{E}, \quad (17)$$

where e is the charge of the proton and $\vec{E}$ is the electric-field vector. The force and the electric field are both treated as vectors because a charge is urged in the direction of the field.*

A force $\vec{F}_{\text{mag}}$ is also exerted on a charged particle when it moves in a magnetic field (no force is produced by the magnetic field if the charge is at

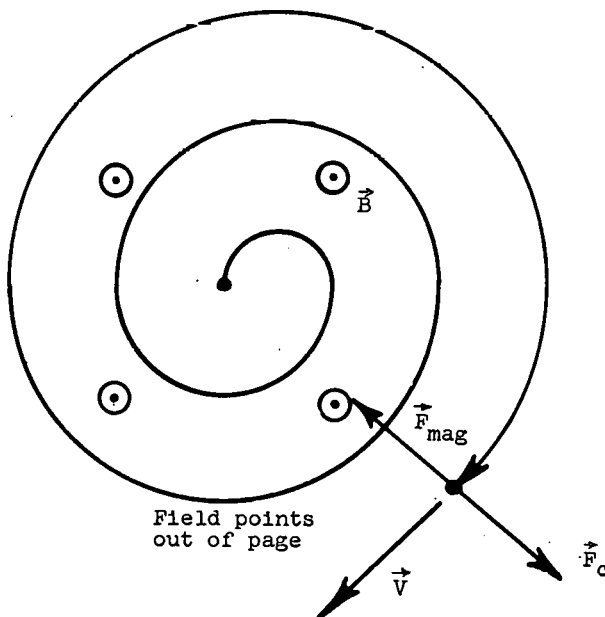


Fig. A-24

* By definition, the direction of an electric field is the direction in which a positive charge is pushed.

rest). The value of $\vec{F}_{\text{mag}}$ is given by this expression:

$$\vec{F}_{\text{mag}} = e(\vec{v} \times \vec{B}), \quad (18)$$

where $\vec{v}$ is the velocity of the charge and $\vec{B}$ is the magnetic-field vector. Note that the term $(\vec{v} \times \vec{B})$ is a cross product, not a dot product.

If a charged particle is moving through both an electric and a magnetic field, the total force $\vec{F}$ exerted on the particle is:

$$\begin{aligned} \vec{F} &= \vec{F}_{\text{elec}} + \vec{F}_{\text{mag}} \\ \vec{F} &= e\vec{E} + e(\vec{v} \times \vec{B}). \end{aligned} \quad (19)$$

An example of a charged particle moving in a magnetic field is the case of a proton being accelerated in a cyclotron (Fig. A-24).

Here the direction of the magnetic field is out of the page. As the proton travels in a clockwise direction,

a force $\vec{F}_{\text{mag}}$ is exerted on it by the magnetic field. This force $\vec{F}_{\text{mag}}$ is just equal to the centrifugal force $\vec{F}_c$, so that the proton travels in a circle (if viewed at a given instant). This circle becomes a spiral as energy is added at each turn of the proton around the circle. The Bevatron, the 184-inch cyclotron, and the 60-inch cyclotron at Berkeley accelerate positively charged particles. Since their magnetic fields are up, the particles are accelerated in a clockwise direction viewed from above. In the electron synchrotron the magnetic field is down. Since electrons carry a negative charge, they too travel in a clockwise direction.

Bibliography

1. H. Margenau and G. M. Murphy, The Mathematics of Physics and Chemistry, Second Edition (Van Nostrand, Princeton, 1956), Ch. IV.

A-3 ELEMENTS OF THE CALCULUS

Harry H. Heckman

Rates of change are encountered each day. In teaching, for example, one speaks of a pupil's progress or his rate of learning, and makes some attempt to measure it from time to time. At the end of a certain period the pupil is given a grade, which is an approximate measure of the total progress made, that is, the total change he has made integrated over a period of time.

Often one speaks of velocity, which is the rate of change of position with respect to time; or of acceleration, which is the rate of change of velocity. To the physicist these are fundamental quantities. If he can measure the velocity of an object, he can determine its acceleration by means of the calculus, which is that branch of mathematics treating rates of change.

Consider a freely falling object such as a person diving from a height of 16 feet into the water. (Fig. A-25) He is timed with a stopwatch

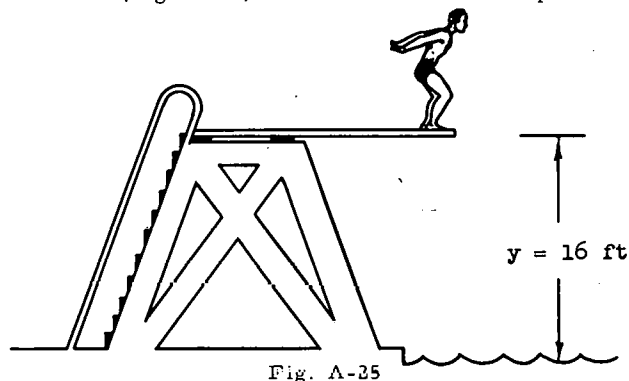


Fig. A-25

and one second elapses before he reaches the water. Therefore, his average velocity is simply 16 ft/sec. It is known, however, that he is not moving at a constant velocity; he is continually being accelerated by gravity.

Suppose now that one wants to find his final velocity, that is, his velocity at the instant he strikes the water. This can be approximated by noting the distance y that he has fallen at a certain instant after he started the dive, say 0.8 sec. If this were done, it would be discovered that in 0.8 sec he traveled 10.24 ft. Thus, his average velocity V_{av} during the final 0.2 sec is

$$\begin{aligned} V_{av} &= \frac{y_{\text{final}} - y_{\text{initial}}}{t_f - t_i} \\ &= \frac{16 - 10.24}{1 - 0.8} \\ &= \frac{5.76}{0.2} \end{aligned}$$

$$V_{av} = 28.8 \text{ ft/sec.}$$

To obtain V_{av} with greater accuracy, one could take successive measurements at times nearer to 1 sec, so that the difference $t_f - t_i$ becomes very small. We call a small interval like this an "increment" in time, and use a symbol such as Δt to describe it. Table A-III summarizes the results.

Table A-III

Determining a final velocity by approximation						
t_f (sec)	t_i (sec)	$t_f - t_i$ (Δt) (sec)	y_f (ft)	y_i (ft)	$y_f - y_i$ (Δy) (ft)	velocity $\frac{y_f - y_i}{t_f - t_i}$ $= (\Delta y / \Delta t)$ (ft/sec)
1	0	1	16	0	16	16
1	0.8	0.2	16	10.24	5.76	28.8
1	0.9	0.1	16	12.96	3.04	30.4
1	0.99	0.01	16	15.6816	0.3184	31.84

From Table A-III it is seen that as Δt becomes smaller, the velocity $(\Delta y / \Delta t)$ approaches 32 ft/sec.

An electronic timing system would allow one to measure the distance the diver falls during extremely short time intervals, so that his velocity could be closely approximated at any instant. The calculus, however, mathematically allows Δt to approach zero, and hence permits one to calculate the instantaneous velocity at any time. Figure A-26 shows the position of any freely falling object with

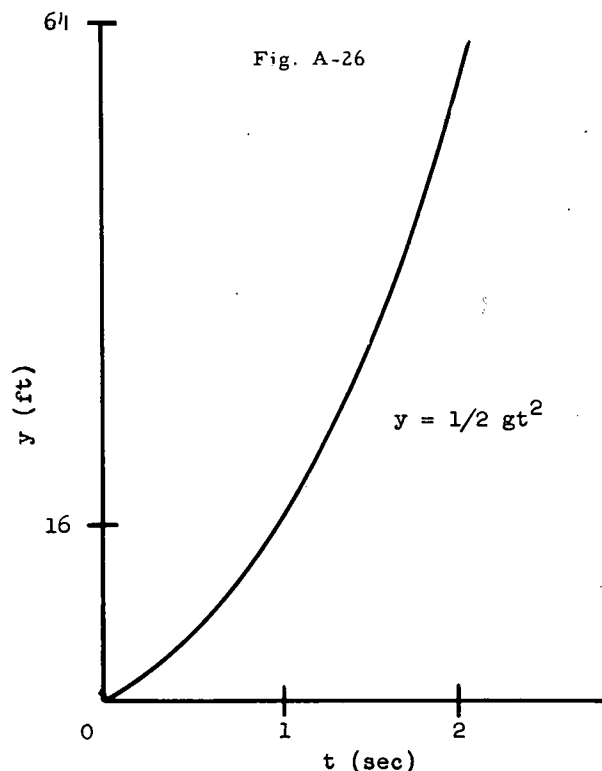


Fig. A-26

respect to time. This graph was obtained by plotting the values of y_i versus t_i from Table A-III. This curve can be expressed by the equation $y = (1/2)gt^2$, where g is the acceleration of gravity ($g = 32 \text{ ft/sec}^2$), a constant. The origin of this equation is apparent from the following relationships:

distance (y)		time
16 ft	$= (32/2) \times 1$	1 sec
64 ft	$= (32/2) \times 4$	2 sec
144 ft	$= (32/2) \times 9$	3 sec.

It should be noted that the average velocity ($V_{av} = 16 \text{ ft/sec}$), calculated above, can be found from Fig. A-26 by drawing a chord from the origin to the curve at the point $t = 1 \text{ sec}$ (Fig. A-27). The slope (y/t) of the chord is the average velocity:

$$V_{av} = 16/1 \text{ ft/sec.}$$

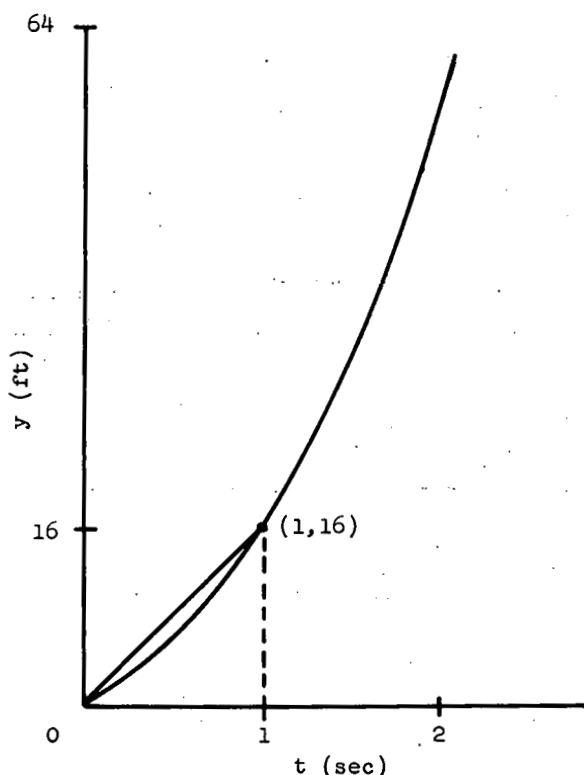


Fig. A-27

The velocity of the diver as he strikes the water can be approximated by drawing successive chords, moving the lower end of the chord along the curve while holding the upper point at 1 sec. We see that as the time interval decreases, the slope of the chord approaches the slope (tangent) of the curve at time $t = 1 \text{ sec}$ (Fig. A-28).

The slope of the tangent at the point $t = 1$ is the instantaneous velocity of the diver as he hits the water.

It is convenient to derive a general mathematical expression for the slope of the tangent to a curve at any point, since the slope represents a velocity (i.e., the rate of change of the curve).

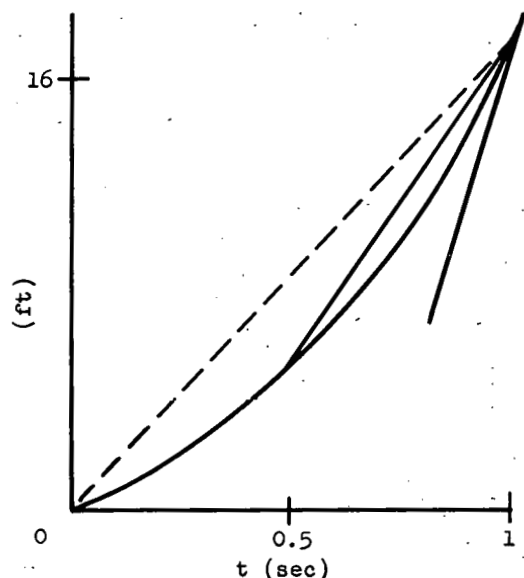


Fig. A-28

Consider the equation

$$y_i = (1/2)gt_i^2$$

Now we ask, "What will the value of y be if we change t slightly?" We write

$$\begin{aligned} y_f &= (1/2)g(t + \Delta t)^2 \\ &= (1/2)g(t^2 + 2t\Delta t + \Delta t^2), \end{aligned}$$

where Δt is the change in time. The slope $\Delta y/\Delta t$ is

$$\begin{aligned} \frac{\Delta y}{\Delta t} &= \frac{y_f - y_i}{\Delta t} \\ &= \frac{(1/2)g(t^2 + 2t\Delta t + \Delta t^2) - 1/2 t^2}{\Delta t} \\ &= \frac{(1/2)g(2t\Delta t + \Delta t^2)}{\Delta t} \\ &= (1/2)g(2t + \Delta t) \\ \frac{\Delta y}{\Delta t} &= gt + (1/2)g\Delta t. \end{aligned}$$

Now if we let Δt approach zero, we obtain the slope of the curve:

$$V = \lim_{\Delta t \rightarrow 0} \frac{\Delta y}{\Delta t} = \lim_{\Delta t \rightarrow 0} \left[gt + (1/2)g\Delta t \right] \quad (1)$$

Since the term $(1/2)g\Delta t$ approaches zero as Δt approaches zero, the velocity at time t is

$$V = gt \quad (2)$$

*The notation $\lim_{\Delta t \rightarrow 0} \frac{\Delta y}{\Delta t}$ reads "the limit of Δy with respect to Δt as Δt approaches zero."

When the velocity curve $V = gt$ is plotted, we obtain a straight line whose slope is g (Fig. A-29).

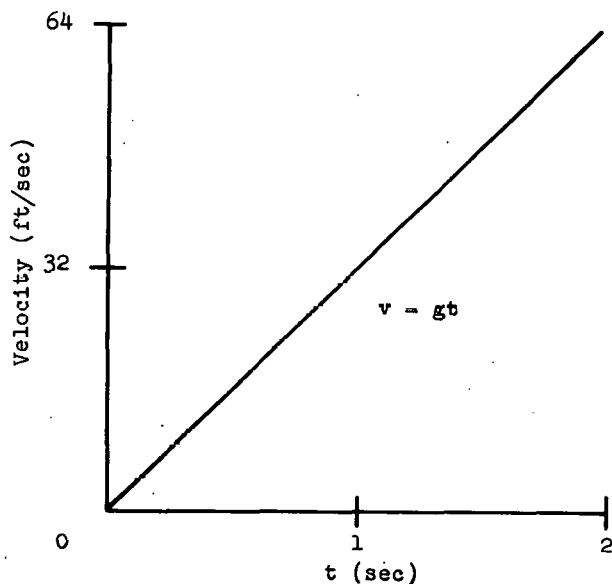


Fig. A-29

Differentiation

What we have done in Eq. (1) is define the "differential," a fundamental definition in the calculus:

$$\frac{dy}{dt} \equiv \lim_{\Delta t \rightarrow 0} \frac{\Delta y}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{f(t + \Delta t) - f(t)}{\Delta t} \quad (3)$$

This applies to any function. The term dy/dt reads "the derivative of y with respect to t ." From this basic definition we evolve a general procedure for evaluating differentials for any function. The process of finding the differential is called "differentiation."

A second derivative of a function can also be obtained. We recall that the first derivative of the function $y = (1/2)gt^2$ is $dy/dt = gt$. From Eq. (2), the second derivative d^2y/dt^2 is

$$\begin{aligned} \frac{d^2y}{dt^2} &= \lim_{\Delta t \rightarrow 0} \frac{g(t + \Delta t) - gt}{\Delta t} \\ &= \frac{g\Delta t}{\Delta t} \\ &= g \end{aligned} \quad (4)$$

since the Δt terms cancel. The second derivative represents the rate of change of velocity, or acceleration. In this case the acceleration is constant (Fig. A-30).

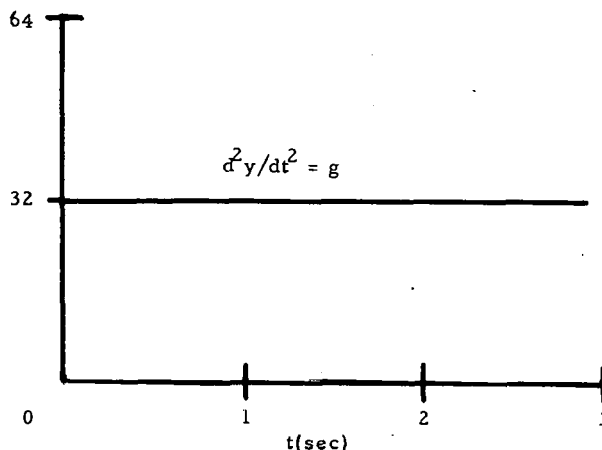


Fig. 30

The third derivative d^3y/dt^3 in this case is zero, since the derivative of any constant is zero. (A constant, by definition, is an entity that does not change with time.)

Table A-IV summarizes the first, second, and third derivatives of the function $y = (1/2)gt^2$.

Table A-IV

Summary of derivatives of the function $y = (1/2)gt^2$			
	Symbol	value	name
first derivative	dy/dt	$= gt$	velocity
second derivative	d^2y/dt^2	$= g$	acceleration
third derivative	d^3y/dt^3	$= 0$	none

It is possible to derive general formulas for the derivative of a function $y = f(x)$. In Table A-V we list some of the more useful of these.

Table A-V

General formulas for the derivative of the functions $y = f(x)$		
Function	First derivative	Second derivative
$y = x^n$	$\frac{d(x^n)}{dx} = nx^{n-1}$	$(n-1)nx^{n-2}$
$y = c^x$	$\frac{d(c^x)}{dx} = c^x \ln c$	$\frac{d^2(c^x)}{dx^2} = c^x (\ln c)^2$
$y = \log_e x$	$\frac{d(\log_e x)}{dx} = \frac{1}{x}$	$\frac{d^2(\log_e x)}{dx^2} = -\frac{1}{x^2}$
$y = \sin x$	$\frac{d(\sin x)}{dx} = \cos x$	$\frac{d^2(\sin x)}{dx^2} = -\sin x$
$y = \cos x$	$\frac{d(\cos x)}{dx} = -\sin x$	$\frac{d^2(\cos x)}{dx^2} = -\cos x$

Two points should be mentioned about using derivatives:

(a) When the first derivative (slope of a curve) is positive, the curve is rising. When the first derivative is negative, the curve is falling (Fig. A-31).

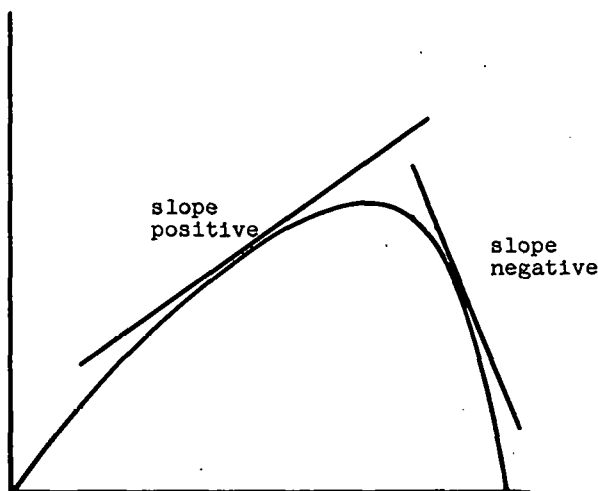


Fig. A-31

(b) When the first derivative (slope) is zero, the curve is at a maximum or a minimum (Fig. A-32). The way to tell whether the point in question is at a maximum or a minimum (without drawing the graph) is to take the second derivative of the function. If the second derivative is negative, the point in question is a maximum; if it is positive, the point is a minimum.

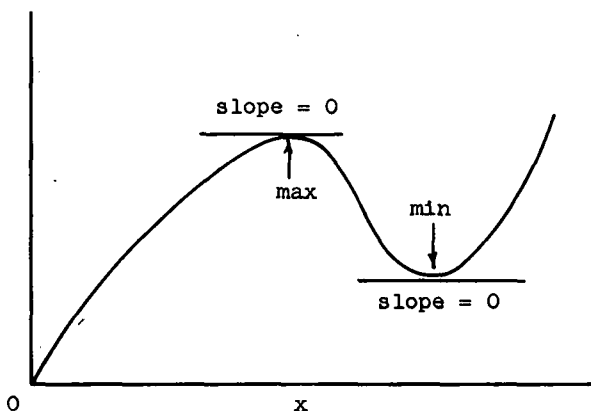


Fig. A-32

Integration

In studying moving objects, physicists often know the equations of motion and laws of force that govern the motion of the object. For example, we have Newton's famous law

$$F = ma. \quad (5)$$

The force F exerted on the object can be measured, as can its mass m . The acceleration a can then easily be derived. Once an expression for the acceleration is obtained, the physicist uses "integration" to find the velocity of the object. Integration is simply the reverse of differentiation. Another "integration" yields an expression for how the position of the object varies with time. In our example above, we took the position of a body as a function of time $[(1/2)gt^2]$ and differentiated it twice to show the motion was that of a body under constant acceleration (g). It would be just as logical if we had started (initially knowing the body was in a force field that

produced a constant acceleration) with $g = d^2y/dt^2$ and asked ourselves: "What is the position of the body at any time?"

To illustrate how integration is performed, suppose we know that the acceleration of a freely falling body is the constant g , and we want to find its velocity V . What function, then, when differentiated, gives g ? The answer is $V = gt$. The way we obtain this is as follows: We know that g is the first derivative of the function being sought. Therefore we write

$$\frac{dv}{dt} = g.$$

We rearrange the terms to obtain

$$dv = g dt.$$

Next we write the symbol $\int$ to indicate that we are to perform an integration. Because g is a constant, it is placed outside the integral sign.

$$\int dv = g \int dt. \quad (6)$$

Performing the integration, we obtain

$$v = gt + \text{constant}. \quad (7)$$

The constant in Eq. (7) arises in every integration. In this case, the constant represents any initial velocity. For instance, in our previous example if the diver had been pushed downward off the diving board with an initial velocity of, say 10 ft/sec, then the constant would be 10 ft/sec.

In Table A-VI we list a number of the more useful general formulas for integrating various functions.

Table A-VI

General formulas for integration		
$\int dy = \int x^m dx$	$y = \frac{x^{m+1}}{m+1}$	$+ c$
$\int dy = \int \sin x dx$	$y = -\cos x$	$+ c$
$\int dy = \int \cos x dx$	$y = \sin x$	$+ c$
$\int dy = \int \frac{dx}{x}$	$y = \log_e x$	$+ c$
$\int dy = \int e^x dx$	$y = e^x$	$+ c$

Application of Calculus to Physics

Escape Velocity from the Earth's Gravitation Field

In building a sun rocket a question that arises is, "What must be the velocity of the rocket in order for it to escape the earth's gravitational field?" In Fig. A-33 we have a rocket of mass m traveling with a velocity V at a distance r from the center of the earth. The radius of the earth is R and its mass is M . Since r is greater than R , the earth's gravitational field acts as though it is concentrated at the center of the earth.

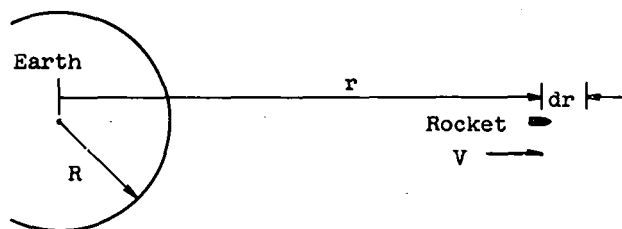


Fig. A-33

The attractive force F between the rocket and the earth is

$$F = -\frac{GmM}{r^2} \quad (8)$$

where G is the universal gravitational constant. The minus sign indicates the force is an attractive one. Let us consider the rocket at distance r from the center of the earth. Then a short time later it is at a distance $r + dr$. The kinetic energy lost, dw , in going through distance dr is given by

$$dw = F dr, \quad (9)$$

where F is the force exerted on the body by the gravitational field. Substituting for F in (Eq. 8), we have

$$dw = -\frac{GmM}{r^2} dr. \quad (10)$$

Integrating this to obtain the total change in the kinetic energy of the rocket moving out to a distance r , we obtain (see Table A-VI)

$$\int dw = -GmM \int \frac{1}{r^2} dr$$

$$w = (1/2) mV^2 = \frac{GmM}{r} + c. \quad (11)$$

To evaluate the constant c in Eq. (10), we know the velocity V will be zero at the maximum radius, $r_{\max}$, the rocket attains before it begins to fall back to earth. Thus, at $V = 0$ we have

$$r = r_{\max}$$

$$(1/2) mV^2 = \frac{GmM}{r} + c$$

$$(1/2) m(0)^2 = \frac{GmM}{r_{\max}} + c$$

$$c = \frac{-GmM}{r_{\max}} \quad (12)$$

The kinetic energy w expended in going from r to $r_{\max}$ is then

$$(1/2) mV^2 = w = \frac{GmM}{r} - \frac{GmM}{r_{\max}} \quad (13)$$

From this we obtain

$$V^2 = 2GM \left(\frac{1}{r} - \frac{1}{r_{\max}} \right) \quad (14)$$

which is the velocity required for a rocket to reach an altitude of $r_{\max}$ when fired from position r . From Eq. (14) we see that the velocity is independent of the mass of the rocket. That is, a 1-pound rocket must travel at the same velocity as a 10-ton rocket to reach $r_{\max}$.

If the rocket is shot off from the surface of the earth (i.e., at $r = R$), and if we want the rocket to escape (i.e., $r = \infty$), we have

$$V^2 = 2GM \left(\frac{1}{R} - \frac{1}{\infty} \right)$$

$$V^2 = \frac{2GM}{R} \quad (15)$$

since $1/\infty$ approaches zero.

On the surface of the earth, the force on an object of mass m is

$$F = mg, \quad (16)$$

where g is the acceleration of gravity. From our general definition of force [Eq. (8)], Eq. (16) is also equivalent to

$$F = mg = \frac{GmM}{R^2}$$

We cancel an m on each side to obtain

$$g = \frac{GM}{R^2} \quad (17)$$

Combining Eq. (17) and Eq. 15, we obtain

$$V^2 = 2gR.$$

Thus, the escape velocity is

$$V = \sqrt{2gR} \quad (18)$$

$$= \sqrt{2(32)4000/5280}$$

$$V = 6.96 \text{ mi/sec.} \quad (19)$$

If a rocket takes off from the earth with a velocity in excess of 6.96 mi/sec it will escape the gravitational field of the earth. An object would have this same velocity (6.96 mi/sec) when it strikes the earth if it were to "fall" in toward the earth from an infinite distance.

Radioactive Decay

Suppose we have a mass of radioactive material containing N atoms. How many atoms will we have after a certain interval of time? We have a detector that counts the atoms as they decay. In a time interval dt we detect dN disintegrations, so that the rate of disintegration equals dN/dt .

It is known from experiments that the disintegration rate dN/dt is proportional to N , the number of atoms present. We write

$$dN/dt = -\lambda N, \quad (20)$$

where λ is a proportionality constant characteristic of a given radioelement. The minus sign indicates that as t increases, N decreases. Rearranging Eq. (20), we obtain

$$dN/N = -\lambda dt, \quad (21)$$

which is in a form that we may now integrate to find an expression which tells us how many atoms N we have after a certain time t . We write

$$\int dN/N = -\lambda \int dt. \quad (22)$$

Performing the integration (see Table A-VI), we find

$$\log_e x = -\lambda t + c, \quad (23)$$

where c is a constant depending upon the initial conditions. At $t = 0$ there has been no decay, that is, there are N_0 atoms to start with, and $c = \log_e N_0$. This means that $\log_e (N/N_0)$ equals $-\lambda t$. We can write this in terms of an exponential:

$$\begin{aligned} N/N_0 &= e^{-\lambda t} \\ N &= N_0 e^{-\lambda t}, \end{aligned} \quad (24)$$

where e is 2.718, the base of the system of natural logarithms.

Equation (24) is the fundamental equation of radioactive decay. Since it describes a statistical process, it does not tell precisely when an individual atom will decay, but merely indicates that for a large group of atoms, a certain fraction will decay (on the average) in a given time. When we plot Eq. (24), we obtain the curve shown in Fig. A-34 (for $\lambda = 1$). At $t = 0$ the fraction N/N_0 is 1; as t increases, the fraction N/N_0 decreases. Here we define the characteristic time τ as $1/\lambda$. We speak of τ as the mean life of radioelement. At one mean life ($\tau = 1$), the fraction of atoms remaining (N/N_0) is 0.37. That is, 37% of the atoms are still present after one mean life. At $\tau = 2$, the fraction N/N_0 is ~ 0.14 , (i.e. $0.37 \times 0.37 = 0.14$) and so on, the curve approaching the x axis asymptotically.

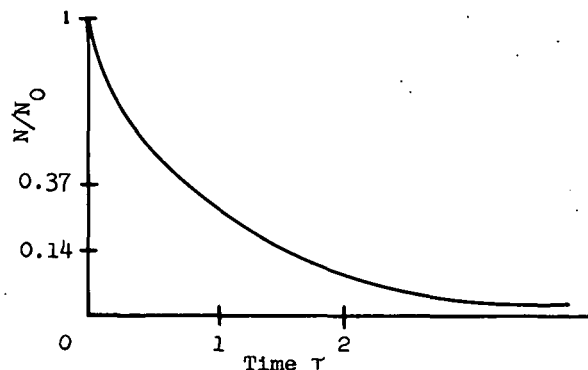


Fig. A-34

Another useful concept is that of half life. The half life of a radioelement equals $0.693/\lambda$. We may describe radioactive decay in terms of the half life $T_{1/2}$, which is related to the mean life τ by the following expression

$$T_{1/2} = 0.693 \tau. \quad (25)$$

Thus, in time $T_{1/2} = 1$ (i.e., in one half life), one half the atoms disintegrate. For instance, if there were 1000 atoms initially, after one half life there would be 500 atoms. After another half life there would be 250 atoms, and so on.

This argument is not valid when there are just a few atoms in a sample. For example, when just one atom is left, the statistical approach no longer holds. It is impossible to say with certainty how long a single atom will live. However, it has a certain probability of living one mean life. The probability is less that it will live two mean lives, and even less that it will live longer. Atoms sometimes do live as many as 20 mean lives, but the probability of this occurring is about one in 500 million.

A-4 THE STATISTICAL ANALYSIS OF DATA

Harry H. Heckman

In an experiment one must always seek to eliminate or minimize sources of error. Data must then be taken in a manner determined to be reliable and the necessary calculations performed correctly. Nevertheless, the results always contain a certain unavoidable error, owing to imperfections in the apparatus or technique. It is therefore necessary to analyze the results statistically in order to determine their accuracy. This lecture will be a review of some of the more useful principles of statistics.

Rules of Probability

First, we define P_A as the probability that event A will occur and P_B as the probability that event B will occur. Then the probability P that either A or B will occur is

$$P = P_A + P_B. \quad (1)$$

Suppose, for example, that on a punch board the chance of winning one dollar is 1 in 4, and the chance of winning three dollars is 1 in 2. The probability of winning either one dollar or three dollars with one punch is then

$$P = 1/4 + 1/2$$

$$P = 1/3.$$

Next, consider a situation where we want to know the probability that two specific events A and B will occur in succession. An example of this is in flipping coins. What are the chances that in two throws heads will come up once and tails once? We know that the probability P_A for heads occurring on a given throw is 1/2, and likewise that the probability P_B that tails will occur is 1/2. Therefore, the probability P that these two events will occur in succession (irrespective of order) is

$$P = P_A P_B \quad (2)$$

$$= (1/2)(1/2)$$

$$P = 1/4.$$

A similar case is where we want to know the probability P that event A will occur twice in succession. This is given by

$$P = (P_A)^2. \quad (3)$$

Thus, the probability P that heads will occur twice in succession is

$$P = (1/2)^2$$

$$P = 1/4.$$

Similarly, the probability P that heads will occur three times in succession is

$$P = (1/2)^3$$

$$P = 1/8.$$

Consider this problem: How large does a group of people have to be in order that there be at least a 50% chance that two of them were born on the same day of the year? The simplest approach is to consider the probability that in a given group of n people the nth person does not have a birthday in common with any of the other (n-1) persons. For instance, in a group of only two people the probability P_2 is 364/365 that the second person does not have the same birthday as the first. In a group of three persons the probability P_3 is 363/365 that the third person does not have the same birthday as either of the first two. The probability P that none of the three has a common birthday is

$$\begin{aligned} P &= P_2 P_3 \\ &= (364/365)(363/365) \end{aligned}$$

$$P = 0.99381.$$

When we extend this for a group of n people we have

$$\begin{aligned} P &= P_2 P_3 P_4 P_5 \dots P_n \\ &= (364/365)(363/365)(362/365) \\ &\quad (361/365) \dots [365-(n-1)/365] \\ P &= \frac{(364)(363)(362)(361) \dots (364-n)}{(365)^{n-1}} \end{aligned}$$

Table A-VII lists values of P for n up to 25. From this table it can be seen that if there are 22 or more people in a group, the probability is greater than 50% that two of them will have a common birthday.

Table A-VII

The probability P that in a group of n people no two people were born on the same day of the year			
n	P	n	P
2	0.99725	14	0.7423
3	0.99381	15	0.7138
4	0.9856	16	0.6845
5	0.9748	17	0.6545
6	0.9614	18	0.6240
7	0.9456	19	0.5932
8	0.9274	20	0.5623
9	0.9071	21	0.5315
10	0.8847	22	0.5009
11	0.8206	23	0.4705
12	0.7959	24	0.4409
13	0.7697	25	0.4119

Frequency Distributions of Random Events

Binomial (Bernoulli) Distribution

In a case where there is a random distribution of events, the probability P_x that a specified event will occur x times in n trials is

$$P_x = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \quad (4)$$

where p is the probability for the event occurring and q is the probability for the event not occurring on a single trial. Since it is evident that the event will or will not occur, we have $p + q = 1$. The symbol $n!$ (called "n factorial") means $1 \times 2 \times 3 \dots n$. For example $3!$ is $1 \times 2 \times 3 = 6$. In Eq. (4), n and x must be integers.

The reason that Eq. (4) is called the binomial distribution equation is that it is the general term in the binomial expansion of the function $(q + p)^n$. We have

$$(q+p)^n = q^n + nq^{n-1}p + \frac{n(n-1)}{2} q^{n-2}p^2 + \dots + p^n \quad (5)$$

$$= \sum_{x=0}^n \frac{n!}{x!(n-x)!} p^x q^{n-x}, \quad (6)$$

where the symbol $\sum_{x=0}^n$ means the sum of the various

terms of the general term $\frac{n!}{x!(n-x)!} p^x q^{n-x}$ for all integral values of x from $x = 0$ to $x = n$. The first term in Eq. (5), namely q^n , is the probability that the specified event will never happen in n trials. The second term, $nq^{n-1}p$, is the probability that it will occur once in n trials. The last term, p^n , is the probability that the specified event will happen on each of the n trials. Similarly, the general term $\frac{n!}{x!(n-x)!} p^x q^{n-x}$ is the probability that the event will occur x times in n trials.

Consider the probability P_x of throwing an ace a given number of times x on n trials with a single die (one of a pair of dice.) The probability P that an ace will be thrown on a single trial is $1/6$. The probability that this will not occur is $5/6$. Therefore, from Eq. (4) the probability P_2 that two aces will be thrown on three trials is

$$P_2 = \frac{3!}{2!(3-2)!} (1/6)^2 (5/6)^{(3-2)}$$

$$P_2 = 5/72.$$

Poisson distribution

The Poisson distribution, an approximation to the binomial distribution, is widely used in physics. An example of its application is in describing radioactive decay. In a given sample of a radioelement the probability that a particular atom will decay is nearly zero, but since there is an enormous collection of atoms, one can determine a probability by counting a disintegration rate. It is in situations like this that the Poisson distribution is used, i.e., where the probability p that a specified event will occur is very small but where the number of trials n is quite large. A further qualification for

the Poisson distribution to be valid is that we must be able to define a true mean value in such a way that the condition

$$m = np \quad (7)$$

is satisfied. By "true mean value" is meant the mathematical mean of a set of numbers. In practice it may not be possible to measure this, but there are methods for estimating it.

The Poisson distribution is given by

$$P_x = \frac{m^x e^{-m}}{x!} \quad (8)$$

P_x is the probability that a specified event will occur x times when the average number of events occurring in a large number of events is m . The symbol e is the base of the natural logarithms ($e = 2.7183$). It should be noted that although x must be an integer, m need not be.

Normal (Gaussian) distribution

The statistical theory of error is based on the assumption that a normal distribution exists. A normal distribution is an approximation to the binomial distribution. It applies when the number of events n is very large. Unlike the previous cases, the normal distribution function is applicable when the observed variable is a continuous function (that is, the observed variable need not be an integer).

The normal distribution is given by

$$P_x^{\Delta x} = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(x-m)^2}{2\sigma^2} \right] \Delta x \quad (9)$$

$P_x^{\Delta x}$ is the probability that a specified event will occur in an interval Δx . The symbol m represents the true mean value, and σ is the standard deviation (see Fig. A-35).

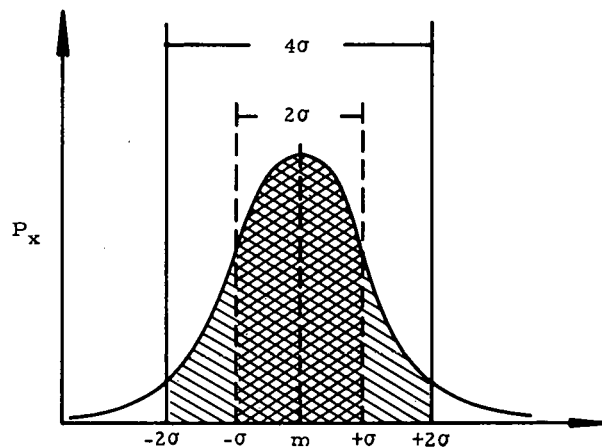


Fig. A-35

The true mean value m is such that the area under the curve (Fig. A-35) is divided into two equal halves. In practice, the value of m is estimated in the following way: a number of measurements n is made of some variable x . We represent these various measurements as $x_1, x_2, x_3, \dots, x_n$. The true mean value m is then approximately equal to $\bar{x}$, where $\bar{x}$ is defined as

$$\bar{x} = \sum_{i=1}^n \frac{X_i}{n} \quad (10)$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Therefore, we write

$$m = \bar{x}, \quad (11)$$

which is not precisely true, but is the best estimate of m that is possible.

The standard deviation σ is a measure of the width of the peak of the normal distribution curve (Fig. A-35). As such, it indicates how well a single measurement has been made. The standard deviation σ is defined by the equation

$$\sigma^2 = \sum_{i=1}^n \frac{(x_i - m)^2}{n} \quad (12)$$

where σ^2 is called the "variance." In practice, σ is found from the expression

$$\sigma^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1} \quad (13)$$

since the value of m cannot be determined empirically. The value of σ as derived from Eq. (13) is

$$\sigma = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}} \quad (14)$$

which we often refer to as the "root-mean-square" of the distribution.

It is found from Eq. (12) that the area under the curve (Fig. A-35) between $\pm\sigma$ is 68.3% of the total area under the curve; the area under the curve between $\pm 2\sigma$ is 95.45% of the total. Thus, the value of σ must vary, depending on the slope of the particular curve. If σ is small, the curve has a sharp peak, indicating that the various measurements do not differ widely from the estimated mean $\bar{x}$.

In an experiment we often want to know the accuracy of the estimated mean $\bar{x}$. Therefore, we define $\sigma(\bar{x})$ as the standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (15)$$

This is the standard deviation of the estimated mean $\bar{x}$. From Eq. (15) we see that the accuracy of $\bar{x}$ does not improve directly with the number of measurements n , but only with the square root of n . For instance, to decrease an error by 50% would require four times as many measurements as already taken.

Propagation of Errors

Consider an experiment in which the result Z depends upon two variables X and Y in the following way:

$$Z = X + Y,$$

where X contains an error $\sigma(X)$ and Y an error $\sigma(Y)$. We want to find the error $\sigma(Z)$ in Z . Therefore, we write

$$\sigma^2(Z) = \sigma^2(X) + \sigma^2(Y), \quad (16)$$

which states that the variance of Z is the sum of the variance of X and the variance of Y . Thus, the errors add quadratically. For example, if $\sigma(X)$ is 1 and $\sigma(Y)$ is also 1, then $\sigma(Z)$ is found from

$$\sigma^2(Z) = 1^2 + 1^2 = 2,$$

which yields

$$\sigma(Z) = 1.4.$$

We see that the absolute error in Z is larger than the error of either X or Y .

Next, consider the case

$$Z = XY.$$

The error in Z is found from

$$\frac{\sigma^2(Z)}{Z^2} = \frac{\sigma^2(X)}{X^2} + \frac{\sigma^2(Y)}{Y^2} \quad (17)$$

Here the fractional errors add quadratically. For $X = 10 \pm 1$, $Y = 20 \pm 1$, and $Z = 200 \pm \sigma(Z)$ we have

$$\frac{\sigma^2(Z)}{(200)^2} = \frac{1^2}{10^2} + \frac{1^2}{20^2}.$$

Solving for $\sigma(Z)$, we obtain

$$\sigma(Z) = 22.36.$$

Thus, a 10% error in X and a 5% error in Y gives a 11.18% error in Z , so that again the error is compounded in the result.

For the case where $Z = X^2$ we have

$$\begin{aligned} \frac{\sigma^2(Z)}{Z^2} &= \frac{4\sigma^2(X)}{X^2} \\ \frac{\sigma(Z)}{Z} &= 2 \frac{\sigma(X)}{X}. \end{aligned} \quad (18)$$

Thus, the fractional error in Z is twice that in X .

A general formula for the propagation of errors in the function $Z = f(X, Y)$ is

$$\sigma^2(Z) = \left(\frac{\partial Z}{\partial X}\right)^2 \sigma^2(X) + \left(\frac{\partial Z}{\partial Y}\right)^2 \sigma^2(Y). \quad (19)$$

The expression $\partial Z / \partial X$ is called the partial derivative of Z with respect to X . It is found by differentiating the function $Z = f(X, Y)$ with respect to X while holding Y constant. The expression $\partial Z / \partial Y$ is found in an analogous fashion by holding X constant.

An Application of Statistical Methods

The acceleration of gravity g may be determined experimentally in many ways. One of the simplest is to measure the period of a pendulum.

The period τ is related to the acceleration of gravity g by the expression

$$\tau = 2\pi \sqrt{l/g},$$

where l is the length of the pendulum arm. Thus, we have

$$g = 4\pi^2 l / \tau^2. \quad (20)$$

An experiment was performed in which the length l was measured as 130.4 cm. Table A-VIII presents the values of τ that were measured on 10 oscillations.

The estimated mean value $\bar{\tau}$ of the period is

$$\begin{aligned} \bar{\tau} &= \sum_{i=1}^n \frac{\tau_i}{n} \\ &= \frac{\sum_{i=1}^{10} \tau_i}{10} \\ \bar{\tau} &= \frac{22.9}{10} = 2.29 \text{ sec.} \end{aligned} \quad (10)$$

Therefore, we obtain

$$\begin{aligned} g &= 4\pi^2 (130.4) / (2.29)^2 \\ g &= 981.67 \text{ cm/sec}^2. \end{aligned}$$

Table A-VIII

Values of the period τ measured in a pendulum experiment			
Trial	τ (Sec)	$\tau_i - \bar{\tau}$	$(\tau_i - \bar{\tau})^2$
1	2.28	0.01	1×10^{-4}
2	2.27	0.02	4
3	2.31	0.02	4
4	2.29	0.00	0
5	2.31	0.02	4
6	2.29	0.00	0
7	2.25	0.04	16
8	2.29	0.00	0
9	2.33	0.04	16
10	2.28	0.01	1
$\sum \frac{\tau_i}{n} = 2.29$			$46 \times 10^{-4} = \sum (\tau_i - \bar{\tau})^2$

The question arises, "How precise is this value for g ?" First we calculate the variance σ^2 :

$$\sigma^2 = \sum_{i=1}^n \frac{(\tau_i - \bar{\tau})^2}{n-1} \quad (13)$$

$$\sigma^2 = \sum_{i=1}^n \frac{(\tau_i - 2.29)^2}{9}$$

When we substitute the ten values for τ from Table A-VIII into the above equation and sum the squares of the deviations, we obtain

$$\sigma^2 = \frac{46 \times 10^{-4}}{9}$$

$$\sigma^2 = 5.111 \times 10^{-4}$$

The standard deviation σ is then 0.0226 sec. This means that in about 68.3% of the cases, the measured value of τ was within 0.0226 sec of the mean value $\bar{\tau}$ of 2.29 sec. From the foregoing we have an indication of the accuracy of a single measurement.

Next we wish to know the precision of the mean value $\bar{\tau}$. The standard error $\sigma_{\bar{\tau}}$ is

$$\begin{aligned} \sigma_{\bar{\tau}} &= \sigma / \sqrt{n} \\ &= 0.0226 / \sqrt{10} \\ \sigma_{\bar{\tau}} &= 0.00715 \text{ sec.} \end{aligned} \quad (15)$$

Therefore, the mean value $\bar{\tau}$ is 2.29 ± 0.00715 . This gives a percentage of error of $100(0.00715/2.29) = 0.312\%$ for $\bar{\tau}$.

From Eq. (18) we recall that in a function of the type $Z = X^2$ the fractional error in Z is twice the fractional error in X^2 :

$$\frac{\sigma(Z)}{Z} = \frac{2\sigma(X)}{X} \quad (18)$$

Equation (20) is of this type, with $g = Z$ and $\bar{\tau} = X$. Therefore, the fractional error in g is twice that in $\bar{\tau}$:

$$\frac{\sigma(g)}{g} = \frac{2\sigma(\bar{\tau})}{\bar{\tau}}$$

The percentage of error in g is then $2 \times 0.312\% = 0.624\%$. The absolute error in g is

$$0.00624 (981.67) = 6.42 \text{ cm/sec}^2.$$

The calculated value of g may thus be written $981.67 \pm 6.42 \text{ cm/sec}^2$. From the size of the absolute error it is apparent that there is little value in retaining the figures to the right of the decimal point, since from this experiment we can say only that g is somewhere between 976 and 988 cm/sec^2 .

B. ATOMIC PHYSICS

Roger W. Wallace

The Chemical Foundations of the Atomic Theory

By 1800 four laws governing the mass and volume relations between chemical compounds had been discovered. These laws led to the atomic and molecular hypotheses long before any experiments were conducted that measured the individual properties of atoms. These laws deal only with enormous numbers of atoms or molecules, and yet their form is a direct consequence of the individuality of atoms and molecules. The laws are:

- a. The law of conservation of mass* - The total mass of a system is not affected by any chemical change within the system.
- b. The law of definite proportions - A particular chemical compound always contains the same elements combined in the same proportions by weight.
- c. The law of multiple proportions - If two elements combine to form more than one compound, the different weights of one that combine with the same weight of the other are in the ratio of small whole numbers.
- d. The law of reciprocal proportions - The weights of two (or more) substances that react separately with identical weights of a third are also the weights that react with each other, or simple multiples of them.

Based on the above laws Dalton, in 1803, introduced the atomic hypothesis which states that

- a. chemical elements are discrete and individual,
- b. all the atoms of an element are the same, and
- c. simple numerical relations exist between the numbers of atoms that combine into molecules.

Soon after this Avogadro introduced the molecular hypothesis in 1811, which states that equal volumes of gases contain equal numbers of molecules.

Somewhat after this Faraday drew the following two conclusions from his experiments with the passage of electricity through solutions:

- a. The weight of an element that is deposited by an electrical current while passing through a solution is proportional to the absolute electrical charge carried by this current through the solution.
- b. The weight of various elements deposited by the same charge is proportional to the chemical equivalent of these elements.

Faraday found that 96,500 amperes flowing for one second deposit one chemical gram equivalent or mole of a monovalent element. From this number and the charge on one electron (which was meas-

ured many years later by Millikan) one can calculate the number of atoms in one mole of an element. This turns out to be 6×10^{23} , which is called Avogadro's number, since he was the first to point out that each mole contains equal numbers of molecules. All of the preceding information, plus a great deal of chemical data (with the exception of the more recently measured Avogadro's number), was available to Mendeléeef in 1869 when he suggested that the chemical elements could best be understood if arranged in a periodic table. He was drawn to this conclusion by the following facts:

- a. The chemical properties of the elements tend to repeat themselves in a periodic fashion when the elements are arranged in the order of their atomic weights.
- b. Similar elements either have approximately the same weight or have regularly increasing weights. (An example of this is the halogens—fluorine, chlorine, and bromine.)
- c. The order of the atomic weights is also the order of the chemical valences. (There are a number of exceptions to this. For example, some elements have two or three valences. However, these exceptions can be explained.)
- d. The common elements have small atomic weights and sharply defined chemical properties.
- e. The magnitudes of the atomic weights determine the chemical characteristics of the elements.

When the elements were arranged in the periodic table it became immediately possible to predict the existence of several elements that were then unknown--all of which have since either been found in nature or have been produced artificially. The table also made it possible to check the accuracy of the atomic weights of adjacent elements and to make excellent predictions about the chemical properties of elements, which were yet to be discovered or were little known.

The preceding material forms the classical bases for the modern atomic theory and was understood and accepted by the latter part of the 19th century. At the close of the 19th century the atomic theory began to develop very rapidly, owing to the discovery within a few years of cathode rays, x-rays, the quantum theory, the special theory of relativity, and the photoelectric effect.

The Electron

Without any attempt to be chronological, we will first consider the discovery of the electron.

Thompson in 1897 used an evacuated tube of the type shown schematically in Fig. B-1 to demonstrate the properties of cathode rays which he eventually called electrons. This tube is very similar to the modern television tube, and all of Thompson's experiments could in principle be performed by anyone with a television set. The negative electrode, or cathode, in the tube emits electrons which are accelerated toward a slit system that has a positive charge on it. Some of the electrons pass through the slits in a narrow beam and strike the far end of the

* This law is now known to be strictly true only if instead of mass we also include energy. As far as masses that can be weighed on a chemical balance are concerned the law is exact.

tube, which has a fluorescent surface. Thompson observed or demonstrated that

- cathode rays travel in straight lines, in the absence of electric or magnetic fields;
- they can penetrate a thin foil;
- they are negatively charged (as shown when they are allowed to charge an electroscope); and
- they carry with them a considerable amount of energy, since they easily heat the surface on which they fall. When he applied a potential across two parallel conducting plates oriented parallel to the beam, he found that the electrons are deflected toward that plate which is positively charged. He also observed that the electrons are deflected on the arc of a circle when they pass through a magnetic field whose lines of force are perpendicular to their direction of motion.

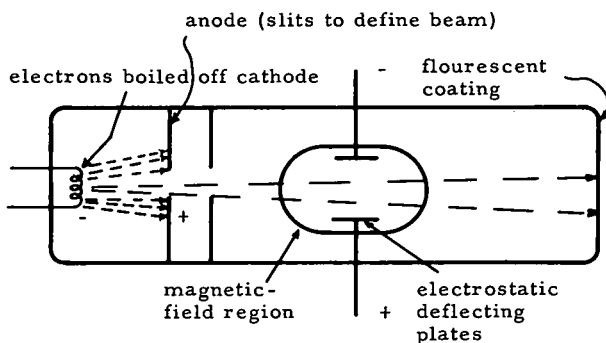


Fig. B-1. Thompson's cathode ray tube.

By passing the cathode rays through a region in which there was an electrical field pulling the cathode rays one way and this was exactly balanced by a magnetic field pulling them in the other direction, he was able to measure their velocity. Since the electrical force eE balances the magnetic force Hev , we have

$$eE = Hev;$$

where e is the charge on the electron, E is the electrostatic field strength, H is the magnetic field strength, and v is the velocity of the electron. Therefore, the velocity v equals the ratio E/H . Thompson found this velocity was typically about 10% of the velocity of light. When he turned off the magnetic field and used only the electrical field he was able to measure the ratio of charge to mass (e/m) for the cathode rays, which turned out to be about 2×10^{17} electrostatic units (esu) per gram (later revised to 5.273×10^{17} esu/g). This was about 1800 times as large as the charge-to-mass ratio of a hydrogen atom. The charge of the electron was not known at the time that Thompson performed his experiment. From this experiment Thompson concluded that the atoms of the cathode were emitting electrons, which are apparently much lighter in weight than a hydrogen atom and which are probably the agent responsible for the conduction of electricity.

In 1909 Millikan succeeded in measuring the charge e on the electron. His apparatus is represented in Fig. B-2. Two horizontal metal plates are contained inside a closed vessel. A fine mist of watch oil is sprayed into the vessel above the plates. Owing to friction in the atomizer, some oil droplets are electrically charged. As a droplet passes through the center hole in the upper plate and settles toward the bottom of the vessel, it is observed with a telescope. The parallel plates are then charged,

the upper plate being made positive. The droplet then moves upward against gravity. By comparing the upward velocity and the downward velocity, and knowing the physical constants involved, Millikan discovered that the various droplets carried integral multiples of a unit charge. This unit charge is called e , and he found its value to be 4.774×10^{-10} esu. Later, this value was revised to 4.80×10^{-10} esu.

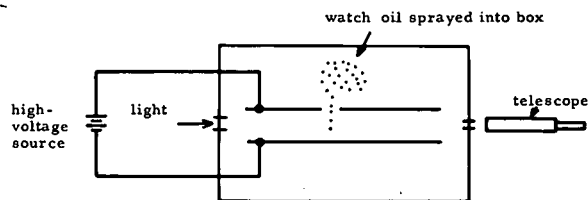


Fig. B-2. Apparatus used by R. A. Millikan to determine e , the charge on an electron.

Since e was now known, the mass m of the electron could be calculated to be 9.108×10^{-28} g. Knowing e also enabled Avogadro's number to be calculated, as well as the approximate mass of the proton, since the proton was known to be about 1800 times heavier than the electron. Millikan's experiment had a number of other ramifications of fundamental value in physics.

The Quantum Theory

The Emission and Absorption of Electromagnetic Radiation

Blackbody radiation

In the latter part of the 19th century and the early part of the 20th century there was much interest in what is called by physicists blackbody radiation. Any body that is not at the same temperature as its surroundings tends to come to the temperature of its surroundings by means of energy transfer between it and its surroundings. This energy transfer takes place either by conduction, convection, or radiation. In the absence of air, radiation is the only means available. Every body above the absolute zero of temperature is continuously receiving and emitting electromagnetic radiation, principally in the infrared region of the electromagnetic spectrum. In order to produce a region which is all at the same temperature and easily comes into equilibrium it is conventional to use a closed box with a small hole in it. This hole when viewed from the outside is said to represent a blackbody, since then the hole appears to be completely black. If the box is heated it eventually begins to emit visible light from the hole, which becomes successively red, orange, and eventually white as the temperature increases. The spectral description of the light emitted from the hole is determined by the temperature of the inner walls of the box. This radiation is said to be the equilibrium radiation emitted by a body at the temperature of the walls.

The spectral distribution of this light (Fig. B-3) was difficult to explain theoretically until Planck in 1901 suggested that electrons can vibrate only at certain frequencies and that electromagnetic radiation must be emitted in discrete quanta rather than continuously as had formerly been thought. When Planck assumed that the energy E of each quantum or photon of electromagnetic radiation is proportional to its frequency ν by the relation

$$E = h\nu, \quad (1)$$

he was able to explain completely the observed spectrum coming from the above-described hole in a box or blackbody. The constant h is called the Planck constant and is 6.62×10^{-27} erg-sec. The frequency ν of the light emitted is given by $\nu = c/\lambda$, where c is the velocity of light (3×10^{10} cm/sec) and λ is the wavelength of the light in cm.

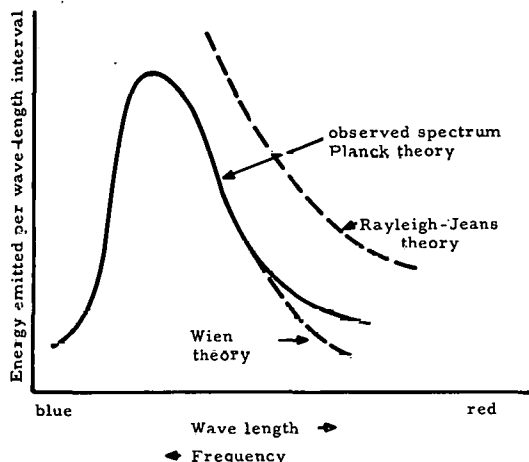


Fig. B-3. Spectrum of blackbody radiation at a given temperature.

This discovery was the beginning of the quantum theory, which later developed into quantum mechanics. The eventual implications of this theory are far-reaching, and have affected every branch of physics, chemistry, and engineering as well as many others.

The photoelectric effect

The explanation of blackbody radiation accounted for the emission of electromagnetic radiation. However, the absorption of electromagnetic radiation was not explained until 1905--by Einstein, who received the Nobel Prize for it. When light falls on an object, electrons are often emitted from its surface. It is found that when the intensity of the light is reduced, the number of electrons emitted is decreased. This was predicted by classical theories. However, an unexpected phenomenon was observed. The energy of the electrons emitted is unaffected by the brightness of the light, but only by its color. This suggests that light probably consists of discrete particles, each having a characteristic energy. Changes in the brightness or intensity of a beam of light affect only the number of such particles present, but not their energy.

As Einstein demonstrated, the explanation for this is that each quantum or photon of light has associated with it an energy $E = h\nu$. This energy is also equal to the kinetic energy $\frac{1}{2}mv^2$ of the electron, plus some small quantity w called the work function (w is the work done by the electron in passing through the surface of the object). We write

$$E = h\nu = \left(\frac{1}{2}\right)mv^2 + w.$$

One photon of light is all that is needed to cause one electron to be emitted. Naturally, if more photons are incident upon the object, more electrons are emitted, but their characteristic energy is not influenced.

Specific Heat of Materials

Another confirmation of Planck's quantum theory was supplied by Einstein, in regard to explaining the specific heat of solids. According to the classical theory of DuLong and Petit, the specific heat of monatomic solids should be $6 \text{ cal/g-mole-}^\circ\text{C}$. Generally, this is true. However, for diamonds the specific heat is only $2 \text{ cal/g-mole-}^\circ\text{C}$. This discrepancy was accounted for by Einstein using the quantum theory. As it turned out, his explanation was not precisely correct, although it was much closer than the classical theory. Debye later provided the correct solution, again based on the quantum theory.

Propagation of Electromagnetic Waves

It can be said that light is emitted and absorbed in quanta, but that in its propagation through space it behaves like a wave motion. There are many optical experiments which clearly demonstrate that light is a wave motion undergoing diffraction, refraction, scattering, interference, and such wave-like behavior. On the other hand, there are many experiments that clearly indicate that light has a corpuscular, or particle, or quantum nature. These two apparently contradictory characteristics of the behavior of light are not actually contradictory when light is viewed from a theoretical standpoint, but are contradictory only when considered by our own limited personal experience. We are quite familiar with the behavior of water waves and somewhat less familiar with the behavior of sound waves, but our detailed experience with light waves is actually quite limited since we never directly observe them traveling along as we do water waves, but only see their result as they strike some object. In general it can be said that we think in terms of classical mechanics, and that we must believe quantum mechanics because of its enormous success in the explanation of experiments, but that we must not expect to have a physical intuition for quantum behavior such as this wave-particle dualism exhibited by light.

The Electromagnetic Spectrum

The electromagnetic spectrum extends over at least sixteen orders of magnitude of wave length that have been investigated by physicists up to the present time (see Table B-I).

Table B-I

Characteristics of electromagnetic radiation	
Type of radiation	Range of wave length (cm)
Gamma rays	10^{-11} to 10^{-9}
X-rays	10^{-9} to 10^{-7}
Ultraviolet light	10^{-7} to 10^{-5}
Visible light	4×10^{-5} to 7×10^{-5}
Infrared light	10^{-4} to 10^{-1}
Short radio waves	10^{-1} to 10^{+5}
Long radio waves	10^{+5} to 10^{+7}

All the above radiations travel at the same velocity and can be thought of as being emitted by accelerated charges. They all behave as photons whose energy is equal to $h\nu$. A naive picture of the emission of electromagnetic radiation can be had by considering that a charged electron is surrounded by an electric field and that if this electron's velocity is changed (i.e. it undergoes an acceleration) some of the electric field surrounding the electron is shaken loose and appears as electromagnetic radiation. This can be thought of as going on in the filament of a light bulb, where the electrons are shaken by the thermal agitation arising from the heating of the filament, or in a radio antenna, where the electrons are forced to move back and forth along the antenna by the voltages impressed on the antenna by the transmitter.

Atomic Structure

The Thompson Atom

The earliest definite theory regarding the internal structure of the atom was that of J. J. Thompson, who in 1898 suggested a model that is sometimes called the "plum-pudding atom." In this analogy the atom might be thought of as a sphere (pudding) inside of which are electrons (raisins). Thompson knew that the positive charges are much heavier than the negative charges and that the latter are responsible for electrical conduction in metals. However, he did not profess to know the exact configuration of these charges within an atom.

The Rutherford Atom

In 1911 Rutherford and Geiger performed an experiment (Fig. B-4) in which they bombarded a thin gold foil with alpha particles, which were known to be the nuclei of helium gas atoms, and which had been discovered by earlier workers in radioactivity. (For our purposes the incident alpha particles can be considered as heavy geometrical points moving at a high velocity through the gold foil.) Most of the alpha particles went through the gold foil almost undeflected. About one in 100,000 were deflected or scattered at large angles, even up to 180° . It could be concluded from this experiment that very little of the foil was a solid barrier to the alpha particles, and that therefore the solid matter in the foil must have an area that is only a fraction of its total area. In addition, the solid matter must be quite heavy, or the large angles of a particle recoil could not be explained. An alpha particle weighing almost 8,000 times as much as an electron could not bounce backward from a collision with an electron any more than a bowling ball would

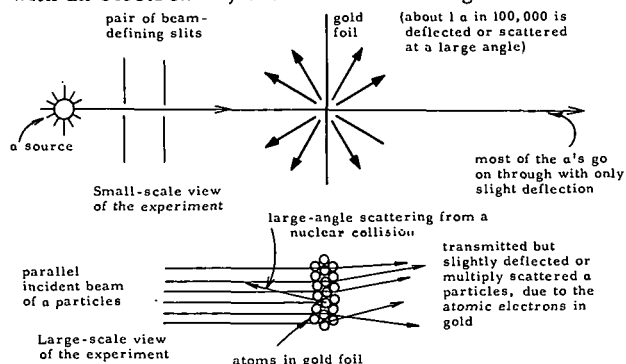


Fig. B-4. Rutherford's scattering experiment.

be expected to bounce back if rolled into a pile of ping-pong balls. Therefore, to account for a bowling ball's bouncing back when it hits a pile of ping-pong balls, there must be an occasional bowling ball hidden among the ping-pong balls in this large-scale analogy. Rutherford concluded that the atom of gold must be largely empty space occupied by electrons, and that the principal mass of the gold was concentrated in a region about 10^{-12} cm in diameter, which is called the nucleus. It had already been shown that the volume available to each gold atom was about 10^{-8} cm in diam., so that the gold must indeed be mostly empty space--as are all solids.

Ever since the days of Newton in the 17th century it had been known that a thin beam of parallel white light when passed through a prism of glass is split up into a variety of colors called a spectrum. This type of spectrum is continuous, one color merging into another. If, however, the light is produced by exciting a gas in an electrical discharge tube, then its spectrum is not continuous (Fig. B-5). Rather, only a few colors appear, as isolated parallel lines. This type of spectrum is called a "line spectrum." The wavelengths of the lines are characteristic of each element contained in the gas. An atom can, in addition to emitting radiation in spectral line, also absorb radiation in a spectral line. This occurs if a continuous spectrum of light passes through a gas. The atoms of the gas then absorb light, leaving breaks in the continuous spectrum at exactly the wavelengths where the atoms would emit light if they were excited. As experimental techniques improved, the variety and complexity of such spectra produced by various incandescent materials rapidly increased. By the beginning of the 20th century many spectra had been carefully measured and the wavelengths and intensities of their spectral lines recorded, but there was no adequate explanation of the structure or origin of these line spectra.

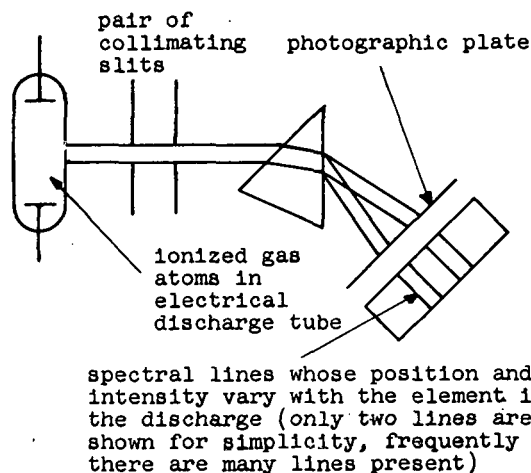


Fig. B-5. Arrangement for producing line spectra.

Rutherford had postulated that the atom consists of electrons circling around a central nucleus in a fashion similar to the planets circulating around the sun. By the electromagnetic theory, an accelerating electron must continually radiate energy. The spectrum would therefore be continuous rather than consisting of lines as is the actual case. A further problem with the Rutherford atom is that in continually radiating energy, the electron would have to spiral rapidly into the nucleus, causing a catastrophe to the atom. Of course this does not occur, since atoms continue to exist for billions of years.

The Bohr Atom

Against this background, Bohr in 1913 made the following suggestions:

a. Contrary to the requirements of electromagnetic theory, an electron does not radiate energy while rotating in a closed orbit. Therefore, its orbit is stable. This is referred to as a "stationary state" of the atom. Several stationary states are possible. The contradiction of classical electromagnetic theory is left as a moot point.

b. The angular momentum of an electron in such an orbit can only have certain discrete values equal to $n\hbar$, where $\hbar$ is the Planck constant h divided by 2π and n is an integer, 1, 2, 3, The angular momentum $n\hbar$ of an electron in an orbit is equal to the linear momentum of the electron (its mass times its velocity) multiplied by the radius of the orbit. This requirement is expressed as

$$mvr = \frac{nh}{2\pi} = n\hbar. \quad (2)$$

This "quantization" of angular momentum actually gives each electron an energy that is characteristic of its orbit, and the particular orbit occupied by an electron is determined by its principal quantum number n .

c. The spectral lines observed in the light emitted by an excited element are a result of the fact that light is emitted or absorbed as an electron makes a discrete jump from one of these orbits to another, and emits or absorbs a definite amount of energy ΔE , which is the difference in the energy of the electron in each orbit. This change in energy ΔE is equal to $h\nu$, where ν is the frequency of the corresponding spectral line ($\nu = c/\lambda$).

Consider an atom of atomic number Z (Fig. B-6). For an electron orbit to be stable, the electrical attraction of the electron toward the nucleus, Ze^2/r^2 , must be balanced by the centrifugal force, mv^2/r . We write

$$\frac{mv^2}{r} = \frac{Ze^2}{r^2} \quad (3)$$

To find the radius r of a Bohr orbit, we must combine Eqs. (2) and (3). First, we multiply Eq. (3) by r , yielding

$$mv^2 = \frac{Ze^2}{r} \quad (4)$$

Next, we square Eq. (2) to obtain

$$\begin{aligned} m^2 v^2 r^2 &= n^2 \hbar^2 \\ r^2 &= \frac{n^2 \hbar^2}{m^2 v^2} \end{aligned} \quad (5)$$

Substituting Eq. (4) into Eq. (5) yields

$$r^2 = \frac{n^2 \hbar^2 r}{mZe^2} \quad (6)$$

Dividing Eq. (6) by r gives

$$r = \frac{1}{mZ} \left(\frac{n\hbar}{e} \right)^2 \quad (7)$$

as the radius of a Bohr orbit.

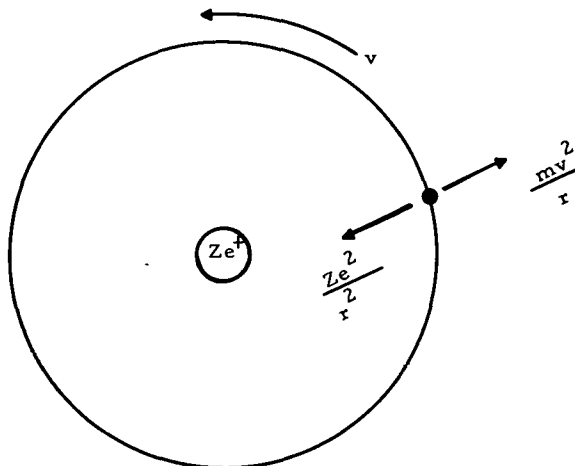


Fig. B-6. Force diagram of an electron revolving about a nucleus.

The energy W of an electron in such a Bohr orbit is equal to the sum of its kinetic energy T and potential energy V :

$$W = T + V, \quad (8)$$

where $T = 1/2 mv^2 = Ze^2/2r$, and $V = -Ze^2/r$. Therefore, we obtain

$$\begin{aligned} W &= \frac{Ze^2}{2r} - \frac{Ze^2}{r} \\ W &= -\frac{Ze^2}{2r}, \end{aligned} \quad (9)$$

the negative sign indicating that the electron is bound in its orbit.

We now wish to find the energy difference ΔE between two different orbits. We write

$$\Delta E = -\frac{Ze^2}{2} \left(\frac{1}{r_1} - \frac{1}{r_2} \right), \quad (10)$$

where r_1 is the radius of the first orbit and r_2 is the radius of the second orbit, as given by Eq. (7). Performing the two substitutions of Eq. (7) into Eq. (10) and rearranging terms, we obtain

$$\Delta E = \frac{me^4 Z^2}{2\hbar^2} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right). \quad (11)$$

The expression $me^4/2\hbar^2$ is defined as the Rydberg constant R , which has the value $109,000 \text{ cm}^{-1}$. Therefore, we write

$$\Delta E = RZ^2 \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right). \quad (12)$$

The value ΔE represents the energy of a photon of light emitted by an excited atom. When a photon is emitted, the electron is said to have jumped from an orbit of higher energy to an orbit of lower energy. This is illustrated for a hydrogen

atom in Fig. B-7. The maximum energy that can be radiated in a hydrogen atom is 13.5 electron volts. This occurs when an electron falls from an outermost orbit (where n is a large number) down to the "ground state" (where n is one). One atom can make only one jump or transition at a time. The many different spectral lines observed are the results of many atoms making a great variety of jumps in a short time.

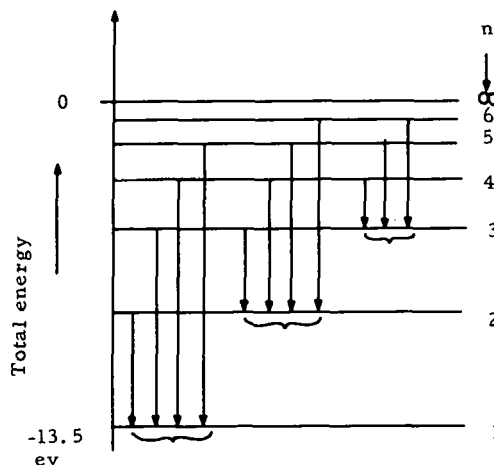


Fig. B-7. Energy-level jumps for an electron in a hydrogen atom.

The symbol n in the above considerations is called the principal quantum number, and each of the integral values that it can assume (1, 2, 3, ...) defines a possible circular orbit for an electron in a hydrogen-like atom.

The Bohr hypothesis predicts the experimental values for the lines in the hydrogen spectrum to about 1 part in 100,000. This close agreement represented a victory for the Bohr theory, even though it contained an apparent contradiction. That is, Bohr used ordinary mechanics to describe the dynamical equilibrium of the orbits, yet he rejected classical electromagnetic theory in explaining the energy jumps by means of the relation $\Delta E = h\nu$. Moreover, while this theory accounted for the important features of the hydrogen spectrum, it did not explain many of the fine details of this and other spectra.

The Bohr-Sommerfeld Atom

In order to make an adequate explanation of the observed complexity of the spectrum, Sommerfeld in 1916 introduced three other quantum numbers. The quantum number ℓ [which can take on the values 0, 1, 2, ... ($n-1$)] is called the orbital angular momentum quantum number. It is associated with a family of elliptical orbits, all of which have the same energy as a particular circular orbit defined by n . For $\ell = 0$, the orbit is circular; for $\ell = 1$, it has a slight eccentricity; for $\ell = 2$ the eccentricity is larger; etc. Another quantum number that must be introduced is s , the spin of an electron; s can take on the values $\pm 1/2$. Each electron can be thought of as having its own angular momentum due to its spin on its own axis, in much the same manner as the earth, which spins on its axis and at the same time travels in its orbit. Another quantum number, m , called the magnetic quantum number, can take on all the integral values between $+\ell$ and $-\ell$.

The magnetic quantum number is the component of ℓ in the direction of the z axis, which in turn is always parallel to the magnetic field.

These four quantum numbers, n , ℓ , s , and m , by assuming values consistent with the selection rules mentioned (such as that s is allowed to be only $+1/2$ or $-1/2$) determine the orbits that an electron can take around a nucleus. For example, for $n = 1$, ℓ and m are both 0, but s can be $+1/2$ or $-1/2$. This means that two electrons can circulate around a nucleus in an orbit for $n=1$, as defined by the equation for the radius above. One of these electrons has its spin pointed up and the other has its spin pointed down. It is not possible for any other electrons to circulate in this particular orbit, owing to the Pauli exclusion principle. This principle states that it is impossible for any two electrons in the same atom to have their four quantum numbers n , ℓ , m , and s identical. That is, only one electron can occupy each possible energy state, as defined by these numbers.

The atoms that are described for $n = 1$ are hydrogen and helium. When n is set equal to two, eight other orbits are possible, and similarly for larger values of n . Table B-II and its accompanying schematic drawings show the situation in the first few elements of the periodic system.

Table B-II

quantum nos.					orbit name
Z	n	ℓ	m	s	
1	1	0	0	$+1/2$	
				$-1/2$	
2	1	0	0	$+1/2$	
				$-1/2$	
3	1	0	0	$+1/2$	
				$-1/2$	
	2	0	0	$+1/2$	
4	1	0	0	$+1/2$	
				$-1/2$	
	2	0	0	$+1/2$	
				$-1/2$	
5	1	0	0	$+1/2$	
				$-1/2$	
	2	0	0	$+1/2$	
				$-1/2$	
	2	1	0	$+1/2$	
6	1	0	0	$+1/2$	
				$-1/2$	
	2	0	0	$+1/2$	
				$-1/2$	
	2	1	0	$+1/2$	
				$-1/2$	

The de Broglie Wave Hypothesis

In 1925 de Broglie suggested that matter can be described in terms of waves. This is a

counterpart to the particle-wave dualism of light and applies to all elementary "solid" particles, such as electrons and protons. The truth of this hypothesis was demonstrated by Davison and Germer. They showed that electron beams are diffracted by crystals in the same way as x-rays.

According to de Broglie, an electron cannot be considered a point source, but must be viewed as a standing wave around its orbit (Fig. B-8). A moving particle has associated with it a wave length λ given by

$$\lambda = h/mv \quad (12)$$

where h is the Planck constant and mv is the linear momentum of the particle. If Eq. (13) is substituted into Eq. (2), it is found that there are n such wave lengths of the electron around the circumference of an orbit of radius r :

$$n = 2\pi r/\lambda \quad (14)$$

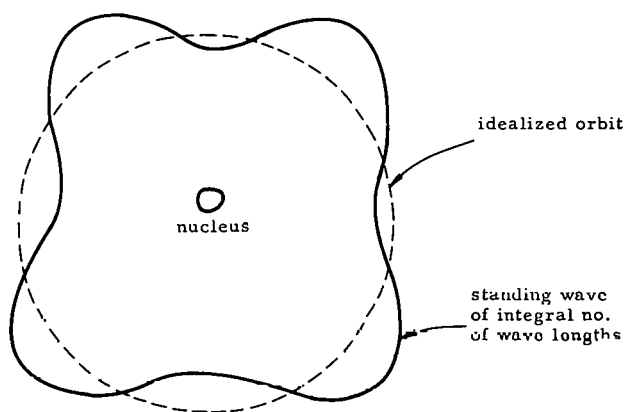


Fig. B-8. The de Broglie atom for $n = 4$.

Thus, we should not think of electrons as particles traveling in orbits. Moreover, in calculating a Bohr orbit by Eq. (7) we should not think of r as describing the position of an electron. Rather, r is a radius about which a standing wave is positioned.

Nevertheless, an electron does behave like a particle (or corpuscle) in many situations, for instance in its motion through electric and magnetic fields. As Bohr later showed, in a single experiment it is impossible to prove that matter has both wave and particle properties. Rather, these are complementary descriptions.

The Heisenberg Uncertainty Principle

In 1927 Heisenberg announced his famous uncertainty principle. He recognized that the wave-particle duality of matter is only one facet of a general law of nature — the concept of a particle cannot be safely extrapolated from a large scale picture to atomic dimensions. According to the uncertainty principle, one can specify the position of an electron to any desired accuracy, but only at the expense of precision in stating its momentum. If Δx is the uncertainty in the position of the electron and Δp is the uncertainty in its momentum, then the product $\Delta x \Delta p$ exceeds or equals some finite value, $\hbar$. We write

$$\Delta x \Delta p \geq \hbar \quad (15)$$

It must be emphasized that this effect is due not to any limitation in one's measuring instruments, but rather to an inherent property of nature, namely that in observing an object it is necessarily disturbed. The constant $\hbar$ is so small that in practice this lower limit to measurement is rarely reached. However, the theoretical implications of the uncertainty principle are far-reaching.

Quantum Mechanics

Because of the uncertainty principle, classical mechanics cannot be used when dealing with atomic particles. Instead, one must use quantum mechanics, introduced by Schroedinger in 1926. In quantum mechanics one does not attempt to state the exact position and momentum of an electron at a given instant, but only assigns a statistical probability for finding it at a given position about the nucleus. This approach avoids the apparent contradictions of the Bohr system. Moreover, the quantum numbers n , l , and m no longer seem arbitrary, and the selection rules governing the electron states are derived rather than imposed.

In spite of having achieved a more satisfactory and rigorous description of atomic structure, quantum mechanics offers no model which can be readily visualized. The Bohr orbits of an electron are now much more complicated three-dimensional probability functions and are referred to as "energy levels." In the Bohr theory the emission of a spectral line was said to accompany a jump of an electron from one orbit to another. In quantum mechanics one thinks of a transition from one energy level to another as causing spectral lines. The concept of an orbit has given way to that of an "orbital," which represents the probability of finding an electron at a given position about the nucleus. However, the term "orbit" is still often used for convenience.

It will be recalled that in the Bohr-Sommerfeld theory that circular orbits are obtained for $l = 0$. For this situation, quantum mechanics states that the probability is greatest that the electron is at the distance from the nucleus represented by the corresponding Bohr radius r [from Eq. (7)]. However, there is a smaller probability that it will be at other values of the radius.

This can be illustrated by the analogy of a particle in a 2-dimensional box (see Fig. B-9). The potential at the edges of the box is infinite, so that the particle cannot get out of the box. However, it is free to roll about the floor of the box without any force being required to produce this motion. There is no friction in quantum mechanics.

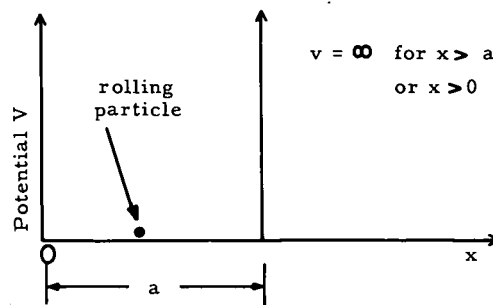


Fig. B-9. Particle in a quantum-mechanical box.

According to classical theory, the probability for the particle being at one point in the box is the same as for another point. This would be true on a large scale, such as for a real ball rolling about in a real box (ignoring friction), but is not true on the atomic scale.

These probability curves are shown in Fig. B-10. The probability $P(x)$ that the electron will be at a distance x from the nucleus is given by the expression

$$P(x) = \psi^2(x), \quad (16)$$

where the wave function $\psi(x)$ is defined as

$$\psi(x) = \sqrt{\frac{n}{a}} \sin \frac{n\pi x}{a} \quad (17)$$

In Eq. (17) n is the principal quantum number and a is the distance across the box and x is the position of the particle in the box.

It should be evident that the probability curve for $n = 1$ is identical to the fundamental mode of a vibrating string. The first harmonic is the same as the curve for $n = 2$, the second harmonic for $n = 3$, and so on. For $n = \infty$ the probability curve would be simply a horizontal line, which of course represents the classical solution for a ball rolling in an ordinary box.

It should be pointed out that as n increases, the energy E associated with a higher level also increases, according to the expression

$$E = \frac{n^2 h^2}{2ma^2} \quad (18)$$

where h is the Planck constant, m is the mass of the particle, and a is the size of the box. From Eq. (18) it can be seen that if n is quite large, say 10^{20} , then the energy would be on the order of that of a billiard ball. Thus, we see how the classical solution is a limiting case of the quantum mechanical solution.

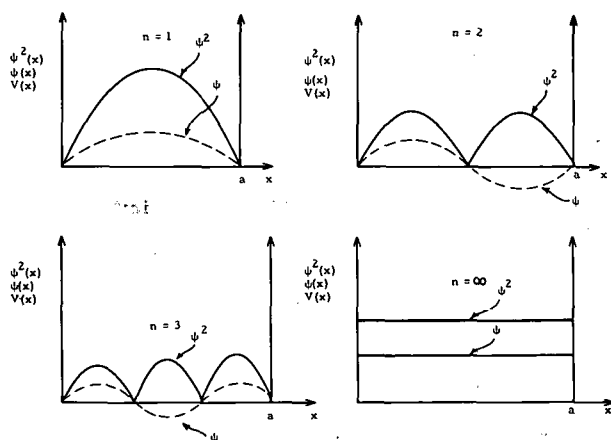


Fig. B-10. Probability ψ^2 of finding an electron at a distance x from the nucleus, where a is the atomic radius.

So far we have discussed only the probability $P(x) = \psi^2(x)$ of finding the electron at a certain distance x from the nucleus. Now we wish to consider its orientation in a 3-dimensional frame.

Instead of using x, y, z coordinates, it is simpler to employ polar coordinates (Fig. B-11). The angle in the xy plane from the x axis to the point lying under the point in question is called ϕ , and is the azimuth of the point. The angle down from the positive z axis is called θ , and is similar to the latitude of a point on the earth, except that the zero point is on the z axis (pole) rather than on the xy plane (equator). The distance from the origin to the point is denoted by r , and is simply the radius. Thus, we have the three variables r, θ , and ϕ , for which we shall find separate solutions.

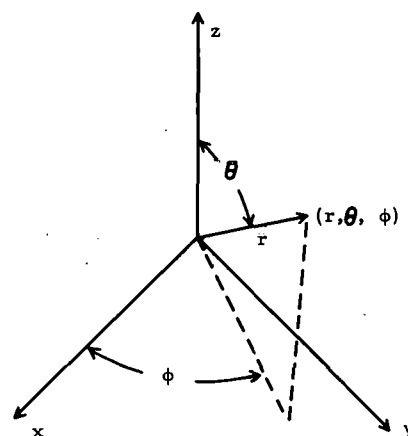


Fig. B-11. Polar coordinates.

Figure B-12 shows the probability $P(\theta)$ of the electron being at a certain angle θ down from the positive z axis. The probability $P(\theta)$ is related to the orbital angular momentum quantum number l and to the magnetic quantum number m . We find that there are zones of high and low probability, depending on the value of l .

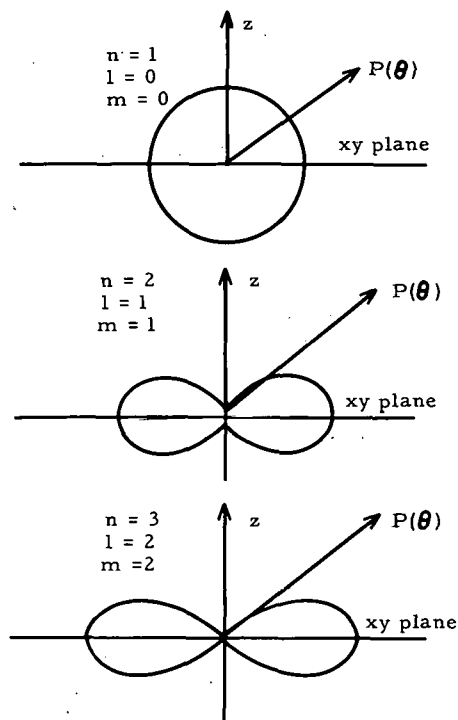


Fig. B-12. Polar graphs of the function $P_{l,m}(\theta)$

Figure B-13 shows the probability $P(\phi) \times P(\theta)$ of the electron being at a certain orientation around a nucleus.

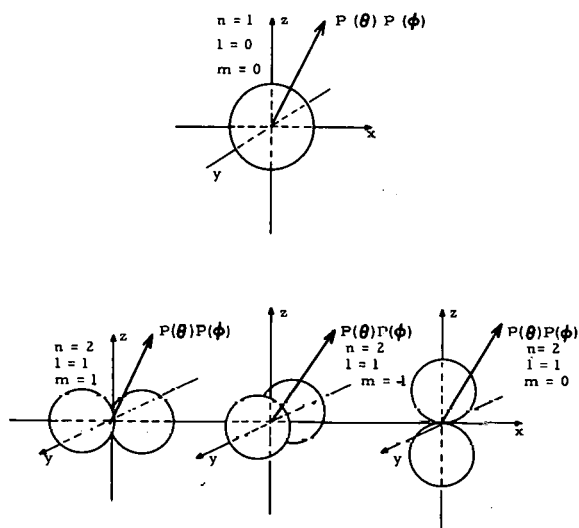


Fig. B-13. Polar representation of the values of the angular probability functions.

The probability $P(r)$ of the electron being at a distance r from the nucleus is shown in Fig. B-14.

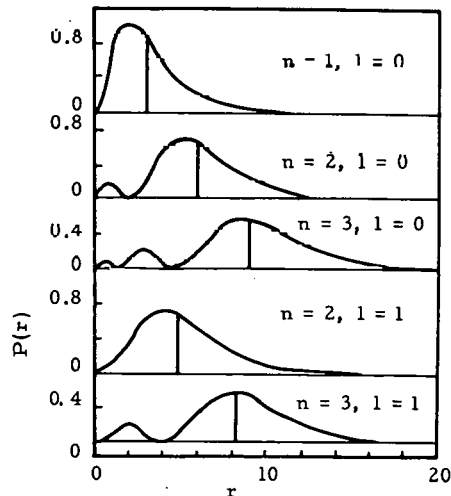


Fig. B-14. Electron distribution functions for the hydrogen atom.

Conclusion

The quantum theory has made it possible to explain not only the atomic structure of all the elements but also the spectral distribution of the light that they emit or absorb. Moreover, this theory can account for their chemical behavior in compounds or as ions, and for their physical behavior in solids, liquids, or gases. The complete elaboration of this picture is the entire subject of atomic physics.

X-rays

Roentgen

In 1895 Roentgen discovered what he called x-rays, which he produced by having cathode rays strike a target. (See Fig. B-15.) When the x-rays that are emitted strike a screen coated with a barium salt, the coating fluoresces, allowing the x-rays to be observed. Roentgen found that x-rays can penetrate an opaque solid, although some of the x-rays are absorbed. He found that the x-rays blacken a photographic plate and cause a gas to conduct electricity, and that the x-rays travel in a straight line and are not influenced by electric or magnetic fields. It is interesting to note that within a very few weeks from the time that Roentgen discovered x-rays they were being used all over the world for medical diagnosis in much the same way as they are used today. Later workers found that x-rays could indeed be reflected, refracted, and diffracted just as light can be. X-rays differ from visible light only in that they have much shorter wavelengths.

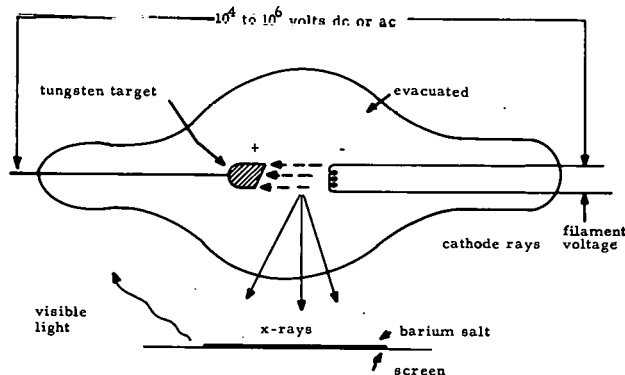


Fig. B-15. An x-ray tube.

Classical electrodynamics shows that a charged particle emits radiation when it is accelerated or decelerated. The sudden stopping of an electron when it strikes a solid is the usual method for producing x-rays. X-rays differ in their penetrating ability, depending upon their particular energy. The energy E of any one photon of x-rays is equal to $h\nu$, where $\nu = c/\lambda$, as is the case with light. When a beam of x-rays from a heterogeneous or nonmonochromatic source strikes a solid object, five different types of radiation are observed emerging from the far side, as shown in Fig. B-16.

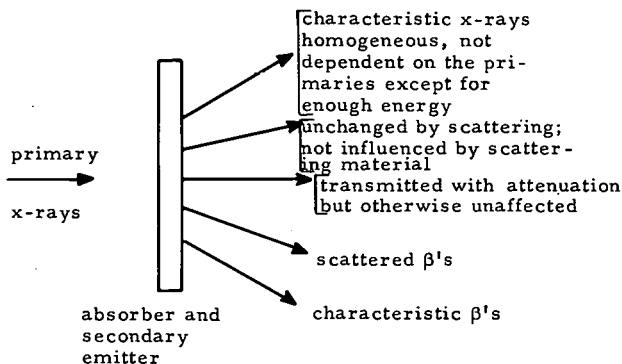


Fig. B-16. Types of radiations that emerge from a material struck on the opposite side by a monoenergetic x-ray beam.

The characteristic x-rays are homogeneous, monoenergetic x-rays emitted by atoms in the absorber undergoing specific energy-level transitions. These x-rays do not depend on the primaries in any way except that the primary x-ray beam must contain x-rays whose energies are greater than a certain minimum necessary to excite the characteristic x-rays from the absorber. The scattered x-rays are merely primary x-rays which have been deflected but are not otherwise influenced by the scattering material. The transmitted x-rays are unaffected by the absorber. We will not discuss the two types of beta rays, other than to say that they are electrons which have been knocked out of the absorber.

Scattering of X-rays

Atoms in an absorber struck by x-rays are slightly accelerated, but the electrons in the absorber undergo much larger accelerations and are by far the most important source of reradiated x-rays. The electrons absorb energy from the x-rays and then, in being brought to rest in the absorber, reradiate this energy or scatter it. In light elements struck by moderately hard x-rays the incident and scattered x-rays differ only in their direction of travel. By classical electrodynamics it is possible to calculate the total energy scattered by a free electron from a plane parallel beam of incident x-rays, as given by the expression

$$I = \frac{8}{3} \pi \left(\frac{e^2}{mc^2} \right)^2 I_0$$

$$I = \frac{8}{3} I_0 \pi r_0^2, \quad (19)$$

where I_0 is the incident energy per cm^2 and I is the energy scattered out of each cm^2 of beam. The expression (e^2/mc^2) is called the classical radius r_0 of the electron and is equal to 2.8×10^{-13} cm, when the values of the electronic charge, electronic mass, and the velocity of light are substituted.

Each electron has an effective area or "Thompson cross section" which it exhibits to an incident beam of x-rays. This cross section equals the ratio I/I_0 , which we denote by the symbol σ_T . We have

$$\sigma_T = \frac{I}{I_0} \quad (20)$$

$$= \frac{8}{3} \pi r_0^2$$

$$= 6.6 \times 10^{-25} \text{ cm}^2$$

$$\sigma_T = 0.66 \text{ barn}, \quad (21)$$

the classical cross section for Thompson scattering by a free electron. Any x-ray striking this area will be scattered from the incident beam. If we consider a thickness of material, dt , containing n atoms per cm^3 , and each atom containing Z electrons, then the fractional attenuation of an incident beam of x-rays in passing through the thickness dt is given by

$$\frac{dI}{I} = -n Z \sigma dt. \quad (22)$$

If we wish to find the fractional loss of intensity as an x-ray beam passes through a thick absorber made up of many little layers each dt cm thick, we integrate the above expression, getting

$$I = I_0 e^{-nZ\sigma t} \quad (23)$$

This expression appears in many different physical situations and is characteristic of any process whereby an incident beam or stream of particles is attenuated "exponentially," as is the case with visible light, x-rays, infrared light, ultraviolet light, radio waves, and with particles undergoing nuclear collisions. This expression is frequently written

$$I = I_0 e^{-\mu t} \quad (24)$$

or

$$I = I_0 e^{-t/\lambda}, \quad (25)$$

where μ is the linear absorption coefficient and λ is the mean free path. If Eq. (21) is solved for Z , we obtain

$$Z = \frac{1}{n\sigma t} \ln \left[\frac{I_0}{I} \right] \quad (26)$$

By substituting the calculated value of σ from Eq. (21) and the measured value of I/I_0 into Eq. (26), we can determine Z for an element that has been used as an absorber. This makes it possible to verify experimentally the atomic number of an element. By x-ray measurements of many elements it has been found that $nZ\sigma$ is given by

$$nZ\sigma \approx 0.2\rho, \quad (27)$$

where ρ is the density of the element. From Eq. (27) and a knowledge of Avagadro's number we find that the atomic number Z is approximately $1/2$ the atomic weight A when the atomic weight of hydrogen is taken to be 1. We write

$$Z/A \approx 1/2. \quad (28)$$

Diffraction of X-rays

In 1912 von Laue showed that crystals act as diffraction gratings for x-rays. In order for a wave motion to be diffracted by a periodic structure it is necessary that the wave length be somewhat shorter than the spacing of the structure, but not very much shorter. X-rays have wave lengths from about 10^{-8} to 10^{-9} cm, and crystals have spacings between their atomic planes of 10^{-7} to 10^{-8} cm.

If monochromatic x-rays fall on a crystal "lattice" at an angle θ to the surface, they are reflected from each of many parallel layers of atoms. These reflected beams are parallel, and if the angle is chosen in such a way that the x-rays from each successive layer traverse an extra distance equal to one wavelength, then the reflected beams of x-rays will add "constructively," and the resulting reflected beam will be quite intense compared with the reflected beam observed at slightly different angles. The values of θ at which this particular "constructive interference" occurs are derived from an examination of Fig. B-17 which shows the extra path length traversed by the right-hand reflected beam (which must be equal to an integral number of wave lengths, $n\lambda$). We have

$$n\lambda = 2d \sin \theta, \quad (29)$$

which is called the "Bragg condition." With a particular crystal, d is a constant. As n changes by integral values from 1 to 2 to 3, etc., the incident x-rays are reflected by angles θ_n , which may be found from the expression

$$\theta_n = \sin^{-1} \left(\frac{n\lambda}{2d} \right) \quad (30)$$

These are called respectively first-order, second-order, etc., reflections.

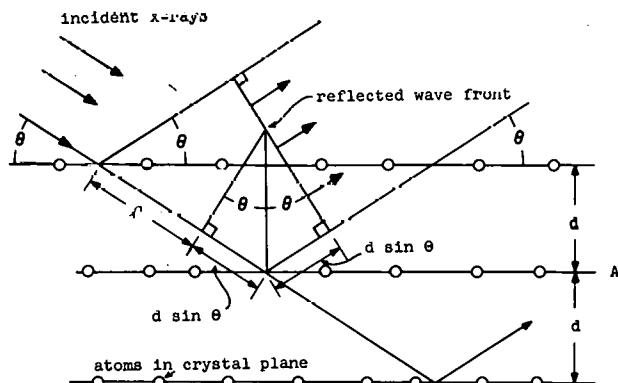


Fig. B-17. Reflection of x-rays from a crystal lattice. Note: $2d \sin \theta$ is the extra distance traveled by the beam reflected from Plane A compared with the distance traveled by the beam reflected from the top plane of atoms, and is equal to $n\lambda$.

X-ray Spectroscopy

A simple spectrometer can be made using a crystal in the manner just described. Such a spectrometer is shown in Fig. B-18.

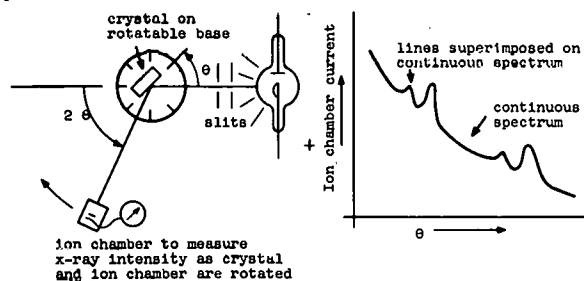


Fig. B-18. Crystal spectrometer. As the crystal rotates through the angle θ , the ion chamber must rotate through the angle 2θ about the same axis.

The X-rays coming from the tube consist of a continuous spectrum. The lines in the spectrum change as the target material changes. The lines move to different angles but do not change in their relation to one another as the crystal is changed, since the spacing d of the crystal planes is different for different crystals.

In 1913-1914 Moseley made x-ray-tube targets of a great variety of elements. He measured the spectra arising from each and noted that there is an orderly shift of the characteristic lines in the x-ray

spectrum from element to element. He concluded that

- there must be a physical quantity that increases from one element to the next in the periodic system (this is Z , the atomic number);
- this quantity must be the total charge on the nucleus or the number of electrons around the nucleus;
- the atomic weight increases by about 2 from one element to the next;
- since we have $Z = 1/2A$, Z must increase by 1 unit from one element to the next;
- the number of the place in Mendeleef's chart is the number of charges on a nucleus; and
- the order of the atomic numbers Z is the same as the order of the atomic weights (except where it was already known that the weights disagreed with the order of the chemical properties).

Futhermore, Moseley was able to verify the vacancies left in the table by Mendeléef for missing elements.

Relation of the Photo Effect to X-Rays

The photoelectric effect is described by the equation

$$E = h\nu - w, \quad (31)$$

where E is the energy of the photoelectron, $h\nu$ is the energy of the incident photon, and w is the work function of the surface of the metal through which the electron must escape.

The energy of an electron striking a target and thus producing x-rays is eV , where V is the potential difference between the cathode and the target. The maximum energy $E_{\max}$ of x-ray photons that are thus emitted is

$$E_{\max} = h\nu_{\max} = eV,$$

where e is the charge on the electron. In all x-ray cases w , the work function of the metal surface, is $\ll h\nu$. Therefore, the energy of a photoelectron emitted by an x-ray is approximately the same as the minimum energy of an electron that could have produced that x-ray.

X-ray Spectra

The several features of x-ray spectra are:

- There is a continuous (or "Bremsstrahlung") spectrum due to electrons being decelerated by the target without exciting any particular atomic level.
- There are line spectra due to a K or an L electron being ejected and an outer electron then falling into this vacancy. In a heavy target these photons have much more energy than the hydrogen spectra had when an electron fell into the K orbit of hydrogen. This can be seen from the energy equation for an L-to-K jump:

$$\Delta E = \frac{mZ^2e^4}{2\hbar^2} \left(\frac{1}{1^2} - \frac{1}{2^2} \right)$$

$$\Delta E = \frac{3}{8} \frac{me^4}{\hbar^2} (Z-1)^2. \quad (11)$$

The term $(Z-1)$ is the effective charge of the nucleus with one K electron circulating around it as seen

by an electron jumping into this K orbit and thus filling the K orbit. The large value of $(Z-1)^2$ accounts for the high energy of x-ray lines compared to optical lines.

c. There are no x-ray absorption lines because the orbits for some distance outside the K and L shells are filled, and the only place a K or L electron can be knocked by an x-ray is into the continuous or unbound energy region above the zero of energy in the energy-level diagram (Fig. B-19).

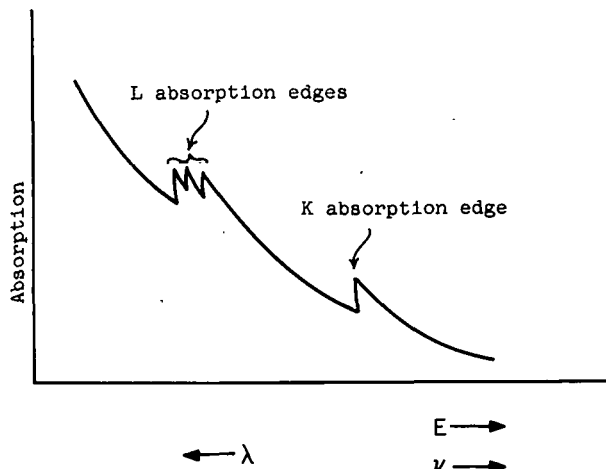


Fig. B-19. Energy-level diagram.

d. There are "absorption edges" in the x-ray absorption spectra, which occur at the energies where a K or L electron can be knocked completely out of an atom.

The Special Theory of Relativity

Einstein, in the year 1905, proposed what is now called the special theory of relativity. Although it is commonly thought that relativity has overthrown many principles of classical physics, this is not true. Relativity merely extends and refines what was previously known, or treats subjects left untouched by earlier theories.

The word "special" refers to the limitation of the theory to problems involving unaccelerated rectilinear motion between two observers. Where there is an acceleration between two observers or where there is motion in a curved path the general theory of relativity applies. For instance, problems concerned with gravitational fields are treated only by the general theory, however there are some problems involving force, acceleration, or curved paths can be treated by the special theory of relativity.

Origin of the Special Theory of Relativity

The special theory of relativity grew out of certain dilemmas encountered by scientists at the end of the nineteenth century. To them it was inconceivable that light could vibrate and be propagated except in some medium. Yet it was known that light passes through space that is devoid of any obvious matter. It was therefore suggested that some weightless gelatin-like material called "ether" must exist. According to this theory, the ether pervades all of space and is stationary in an absolute sense. The earth and other objects move through the ether without disturbing it. Yet when one applies the

equations for the velocity of sound in a solid to finding the velocity of light in ether, some very remarkable results appear. The ether not only is weightless, it is also more rigid than steel. This, as well as other disturbing factors, led physicists to devise experiments that would test the ether theory. The one which has had the most profound significance is the Michelson-Morley experiment, first performed in 1881.

In their experiment, Michelson and Morley sought to measure the velocity at which the earth travels through the presumably stationary ether. (See Fig. B-20.) Suppose that a light is first aimed in the direction the earth travels around the sun. Classically the absolute velocity of this light beam should be $c + v$, where c is the velocity of light (3×10^{10} cm/sec). Next the light source is rotated 90° . The absolute velocity of the beam should then be just c . By means of an interferometer, Michelson and Morley made these measurements. To their astonishment they found $c + v = c$. This equation can be true only if v is zero, that is, only if the earth travels through the ether at zero velocity.

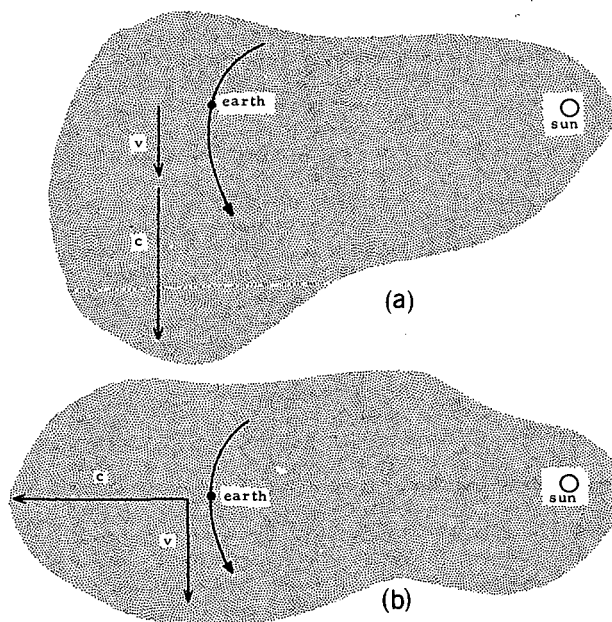


Fig. B-20. The Michelson-Morley experiment. In (a) the velocity of light c is measured parallel to the direction of the earth, and in (b) perpendicular to it.

This surprising result can be interpreted in one of two ways:

- The earth carries the ether along with it,
- There is no ether.

Statement (a) is conceivable but hardly plausible. It can be imagined that the earth could drag the ether along with it in the immediate vicinity of the earth. But it is fantastic that the earth should carry with it all the ether in the entire universe. This would be an unthinkable egotism on our part comparable to believing that our planet is the center of the universe.

As a consequence of this experiment and others the ether theory was gradually abandoned. In its place arose the theory of relativity. Although it

is still referred to as a "theory," relativity has been confirmed in so many ways that its status is that of physical law. Table B-III summarizes the results of the principal experiments that confirm or refute the various theories regarding space, time, and simultaneity. While the other theories are seen in many cases to be contradictory to fact, the special theory of relativity is in complete agreement.

Table B-III

Experimental evidence of the special theory of relativity											
Experiments		Light-propagation experiments						Experiments from other fields			
		Michelson-Morley	Fizeau convection coefficient	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley
Theory		Michelson-Morley	Fizeau convection coefficient	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley
		Michelson-Morley	Fizeau convection coefficient	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley	Michelson-Morley
Ether Theories	Stationary ether, no contraction	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Lorentz contr. and stationary ether	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Ether attached to ponderable bodies	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Original source	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Source Theories	Ballistic	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	New source	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Special theory of relativity		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The experiments outlined above present evidence that:

- (1). The presence of an ether, either stationary or convectively carried, cannot be established.
- (2). Modification of electrodynamics by the emission type theory is untenable.

Theory	Emission theory	Classical ether theory	Special theory relativity
Properties			
Reference System	No reference system.	Stationary ether as reference system.	No reference system.
Velocity Dependence of Light	Velocity of light depends on motion of source.	Velocity of light is independent of motion of source.	Velocity of light is independent of motion of source.
Space-time Connection	Space and time are independent.	Space and time are independent.	Space and time are interdependent.

What the Theory Is

The two principal concepts of special relativity may be summarized as

- a. all motion is relative, and
- b. the velocity of light is constant to all observers.

The first (a) means that it is impossible to measure or detect the unaccelerated translatory motion of an object through free space. This is because there are no unambiguous reference points from which to take measurements. Therefore, when two objects are moving relative to each other at a constant velocity one cannot say that one is moving and one is stationary. Thus, the observer is forced to conclude that physical laws are independent of the velocity of the system in which events are described. If they were not, their difference would define an absolute velocity.

These ideas were obvious to physicists as early as Newton and are implicit in all classical thinking. However, little emphasis was placed on them because few situations were encountered where

there was no frame of reference from which to measure. This is not true in modern physics.

The second major principle (b) of relativity, that the velocity of light is a constant to all observers, has broadened the vistas of physics as no other single postulate has ever done.

It was this aspect of relativity which was implied by the Michelson-Morley experiment. Many experiments conducted since that time have demonstrated this notion to be an experimental fact. Nonetheless, it runs grossly contrary to our classically conditioned intuition.

Consider the following situation, which is very disturbing to a classical thinker. (See Fig. B-21.) Two rocket ships are approaching, each with one-half the velocity of light relative to an observer midway between them. One rocket ship carries a light which it directs toward the other one. The second rocket has a device for measuring the velocity of this light. A classical thinker would suppose that the velocity of the light as measured by the second rocket ship would be $c + 1/2(c) + 1/2(c) = 2c$. However, he would be wrong. The measured velocity would be c no matter what the relative velocity might be between the two rockets.

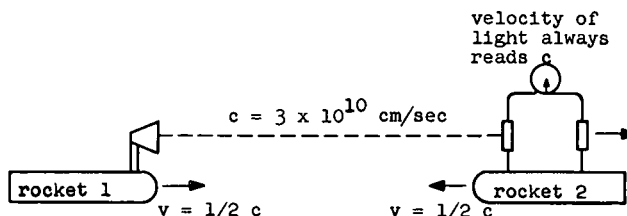


Fig. B-21. "Thought experiment" in which two rocket ships approach each other.

The fact that the velocity of light is a constant to all observers has forced us to reexamine our concept of simultaneity. The notion of a "universal" time scale is no longer tenable. This can be demonstrated in the following situation. Imagine a light pulse starting from a point P. This event is seen by observers stationed in two reference frames. One frame contains P at the origin, whereas the other frame moves relative to P with a velocity v . At the start of a light pulse the origins of the two frames coincide. Since the light has the velocity c relative to both observers independent of their relative velocity v , then both observers must see the light wave propagating as a spherical wave centered at their respective origins. This is in gross disagreement with our classical intuition. Classically only the "stationary" observer would see a spherical wave centered at his origin.

We therefore must abandon the concept of a universal time scale as not corresponding to reality. In its place we must formulate a mechanism whereby simultaneity can be established in a given frame. This mechanism must be such that a measurement of the velocity of light in the particular frame using its time and distance scale always gives c . Thus, the only way to define simultaneity is in terms of the velocity of light. The way in which we must define simultaneity is as follows. Two instants of time t_1 and t_2 observed at two points x_1 and x_2 in a particular frame are simultaneous if

- a. a light wave emitted at the geometrical midpoint between x_1 and x_2 arrives at the time

- t_1 at x_1 and at the time t_2 at x_2 , or if
 b. a light wave emitted at x_1 at time t_1 arrives at x_2 at a time

$$t_2 = t_1 + \frac{x_2 - x_1}{c}$$

Statement (a) assures that a light pulse emitted at the origin will reach all equidistant points simultaneously and that the wave surface is therefore a sphere in that reference frame. Simultaneity of two events at two spatially separated points therefore has no significance independent of the frame. The relation of the time intervals observed by two different frames is dependent on the distance between the events.

We must therefore transform the coordinates x, y, z, t in frame S into coordinates x', y', z', t' in frame S' . Such a transformation must remain linear to assure mathematical equivalence of all points in space and time. However, the spatial and temporal coordinates need not transform independently. This transformation must obey the postulates of special relativity. The following "Gedanken experiments"* illustrate how this is done.

Experiment I - Comparison of Parallel Measuring Sticks Oriented Perpendicular to their Direction of Relative Motion

In Fig. B-22 there are two frames S and S' moving relative to each other. Prior to the experiment the two systems were brought to rest and the lengths of two measuring sticks (one belonging to each) were adjusted to be equal. It is assumed explicitly that the properties of a given body of specified structure are independent of its past history when observed in a frame where that body is at rest. Such a frame is called a "proper frame," and the length of a rod is called its "proper length" when measured in a proper frame.

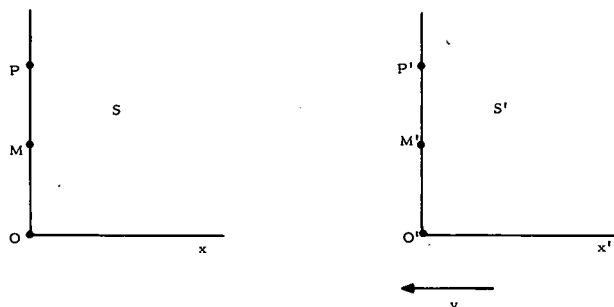


Fig. B-22. "Thought experiment" in which S' moves relative to frame S with a velocity v as measured in S .

Let the two systems S and S' approach each other so that the midpoints M and M' coincide. At the instant when O and P coincide with the y' axis, light pulses are emitted from O and P . Since OM' remains equal to PM' during the motion, O and P appear to cross the y' axis simultaneously in both systems.

It may therefore be concluded that along a direction perpendicular to the direction of relative motion, simultaneity means the same to both systems. Both observers can compare the positions of the end markers at the time of crossover and arrive at the same result, since the time of obser-

vation for both ends is defined identically in both systems. Hence both observers conclude that either $OP \geq O'P'$, or $O'P' \geq OP$. Since both systems are fully equivalent as to their state of motion, an asymmetric solution would provide a means of determining absolute velocity. Thus, the relationships $OP > O'P'$ and $O'P' > OP$ are ruled out by the postulates of special relativity. We therefore have

$$OP = O'P'$$

or, in the general case

$$y = y'.$$

Similarly, we have

$$z = z'.$$

Experiment II - Comparison of Clock Rates

In comparing clock rates between systems, we discover a very astonishing phenomenon: a clock in a moving system runs slow compared to a clock in a system at rest! Consider two frames S and S' (Fig. B-23). Suppose that we are in frame S , which we consider to be stationary. Frame S' is moving relative to our own at a high velocity v . This velocity is nearly equal to that of light, so that the effect we are describing will be noticeable. We have one clock located at the origin of S and another at some point on the x axis. A third clock moves with S' and is located at the origin of S' . At time $t = 0$ the three clocks were synchronized.

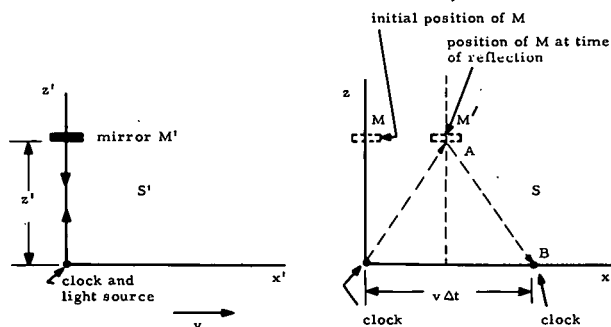


Fig. B-23. "Thought experiment" in which frame S' moves relative to frame S with a velocity v .

At time $t = 0$ the frames S and S' are on top of each other. At that instant a light pulse is emitted from the origin of S' . A light wave travels along the z' axis to a mirror M' located at point z' . It is then reflected back to the origin, where the time interval $\Delta t'$ is recorded by the clock. This interval $\Delta t'$ is related to the distance $2z'$ traveled by the light wave by the expression

$$\Delta t' = 2z'/c. \quad (32)$$

The situation as observed in the stationary frame S is quite different. Before the light pulse is reflected back toward the x axis, the mirror M' has traveled a certain distance in the x direction, to Point A. The light wave therefore strikes the mirror at some angle. The light wave is then reflected from the mirror and strikes the x axis at Point B, where the second clock is located.

The time required for the wave to traverse this path is Δt , as measured by the two clocks in

*Thought experiments.

S. The length of the path is $c \Delta t$, i. e., the velocity of the wave multiplied by the time interval.

The path length of the wave can also be found from the geometry of the triangle. The distance from the origin to Pt. B is $v\Delta t$. The distance from the origin to Pt. A is $\sqrt{(z')^2 + (v\Delta t/2)^2}$, where the relation $z = z'$ has been used. The path length is twice this. Therefore, we equate the two expressions for the path length:

$$c\Delta t = 2\sqrt{(z')^2 + (v\Delta t/2)^2}.$$

Solving this equation for t , we obtain

$$\Delta t = \frac{2\sqrt{(z')^2 + (v\Delta t/2)^2}}{c}. \quad (33)$$

In order to generalize this experiment we should eliminate z' from Eqs. (32) and (33). To do this we substitute the value for z' from Eq. (32) into Eq. (33):

$$c(\Delta t) = 2\sqrt{\left(\frac{c\Delta t'}{2}\right)^2 + \frac{v\Delta t^2}{2}}.$$

Solving for $\Delta t'$, we obtain

$$\Delta t' = \Delta t \sqrt{1 - \frac{v^2}{c^2}}.$$

Defining the ratio v/c as β , we have

$$\Delta t' = \Delta t \sqrt{1 - \beta^2}. \quad (34)$$

The quantity $1/\sqrt{1-\beta^2}$ (which is always > 1) is called the "time dilation" and is referred to by the symbol γ . Thus, we may write

$$\Delta t = \gamma \Delta t'. \quad (35)$$

From Eq. (35) we see that the (not-proper) time interval Δt measured in S by two clocks not at the same place is longer than the (proper) time interval $\Delta t'$ measured in S' by one clock, since γ is always less than unity. From this an observer in the stationary system S would conclude that a clock in the moving system runs slow. To an observer in the moving system S' , the two clocks in S would appear out of synchronism. This is reasonable because simultaneity between two events displaced along the x axis is not independent of the frame.

The apparent asymmetry of Eq. (35) causes no paradox, since this is not a symmetrical situation. The time interval $\Delta t'$ is a "proper time interval" between two events because it is measured between two events occurring at the same place in the S' frame. The interval Δt , on the other hand, is not a proper interval, since it is measured by two clocks at different places. A proper time interval is a definite function of the physical nature of the clock. For example, a particular radioactive decay constant is a constant in a frame where time intervals are observed at a single point, i. e., in a frame where the "clock" is at rest. However, if the clock is not at rest, the decay constant is changed so that the radioactive material lives longer than an observer in a stationary frame would expect.

Another consequence of time dilation is illustrated by the famous "twin paradox."^{*} Consider two twins. One remains on the earth while the other twin makes a voyage into outer space at a very high velocity. Because his biological clock runs slow, the traveling twin does not age as fast as the one at home. When he returns to earth he finds that he is younger than his brother!

As an example, suppose that a rocket ship capable of traveling at a velocity of $\beta = 0.993$ makes a trip to Arcturus, a star some 40 light years away. The time Δt for this journey, as we would measure it on earth is

$$\Delta t = \frac{40}{0.993}$$

$$\Delta t = 40.3 \text{ years.}$$

The time dilation γ at a velocity of $\beta = 0.993$ is

$$\gamma = \frac{1}{1 - (0.993)^2}.$$

$$\gamma = 10.07.$$

Now if one twin goes on this imaginary flight he will return in $2 \times 40.3 = 80.6$ years, according to our clock. However, he will have aged only $2 \times 3.8 = 7.6$ years. The twin who remained at home will thus be $80.6 - 7.6 = 73$ years older than his brother.

It may appear that the above argument is specious. According to the first principle of relativity, all motion is relative. How then can one determine which twin has traveled and which one stayed at home (and therefore aged the more), since absolute motion cannot be detected? The fallacy in this line of reasoning is that this situation is not symmetrical. The twin who remained on earth while his brother traveled had no difficulty knowing whether or not he was traveling. It was his brother who experienced acceleration inside a rocket ship while taking off from earth and traveling relative to the stars and other mass in the universe.

Theoretically, if a rocket traveled exactly at the speed of light its occupants would remain perpetually young. However, it should be noted that we are assuming that biological ageing is influenced by time dilation. This has never been proven because no living organism has ever traveled at velocities high enough to make the relativistic effect noticeable. It is evident from the expression

$$\Delta t' = \Delta t \sqrt{1 - v^2/c^2}$$

that where v is much less than c , $\Delta t'$ is practically the same as Δt . Only when v approaches c do Δt and $\Delta t'$ differ remarkably. Thus at low velocities no difference is detectable.

Therefore, the time $\Delta t'$ for this journey, as measured by a clock traveling with the rocket, is

$$\Delta t' = \frac{40.3}{10.7}$$

$$\Delta t' = 3.8 \text{ years.}$$

^{*} See E. M. McMillan, The Clock Paradox and Space Travel, Science 126, 381-384 (1957); and C. G. Darwin, The Clock Paradox in Relativity, Nature 180, 976-977 (1957).

It is impossible, even theoretically, to build a rocket capable of attaining the velocity of light. This is because the mass of an object increases without bound as it approaches the velocity of light. (Fig. B-24) To accelerate a rocket ship to the speed of light would require an infinite amount of energy. Not even our entire Federal budget could support such an undertaking.

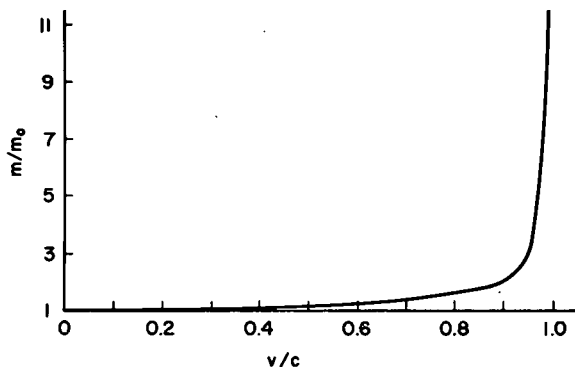


Fig. B-24. Increase in mass of an object as it approaches the velocity of light c ; m is the total mass, m_0 the rest mass, and v the velocity of the object.

The case is different for an extremely light object such as an atomic particle. With modern high-energy accelerators it is possible to obtain particles having velocities as great as 99.999% the speed of light. It is interesting to observe that the mass of these high-energy particles actually does increase, exactly as predicted by the special theory of relativity. (Refer to the section "High-Energy Particle Accelerators" for a discussion of how this relativistic mass increase of particles affects the design of a cyclotron.)

Time dilation is also observed in connection with cosmic rays. At an altitude of, say, 10,000 feet one counts a certain number of μ mesons, which travel almost at the velocity of light. At sea level one again counts the number of μ mesons, which of course is less, owing to the normal decay of many or the loss of others by interaction with air molecules. After making corrections for the expected total loss one finds that many more μ mesons exist at sea level than can be accounted for. The only explanation consistent with the facts is that the mean life of those high-velocity mesons has been prolonged by the time dilation. That is, the atomic clock carried by the high-velocity mesons runs slow, allowing them to live longer than expected. (The same phenomenon is observed in connection with particles from high-energy accelerators. (Refer to the section "Bubble Chambers" for a brief account of how time dilation of decaying particles affected the design of the 72-inch bubble chamber, causing it to be made elliptical in shape.)

Experiment III - Comparison of Lengths Parallel to the Direction of Motion

The results of the Michelson-Morley experiment can be explained in yet another way. In that experiment, it will be recalled, the velocity of light was first measured parallel to the direction of motion of the earth, and then perpendicular to it. The velocity of light was found always to be c . This effect can be explained by what is termed the Lorentz contraction.

Consider two systems, S and S' as shown in Fig. B-25. S' is moving at a velocity v relative to S . As measured in S' , a rod moving with S' has a length x' . That is, its proper length is x' . A light source at S'_0 at one end of the rod sends a pulse to mirror M'_0 at the other end, where it is reflected back to the source. The time interval $\Delta t'$ between the line of emission and the time of arrival is measured by a clock at S'_0 . Note that $\Delta t'$ is the proper time interval, being observable with a single clock at one point. Evidently we have

$$\Delta t' = 2x'/c. \quad (36)$$

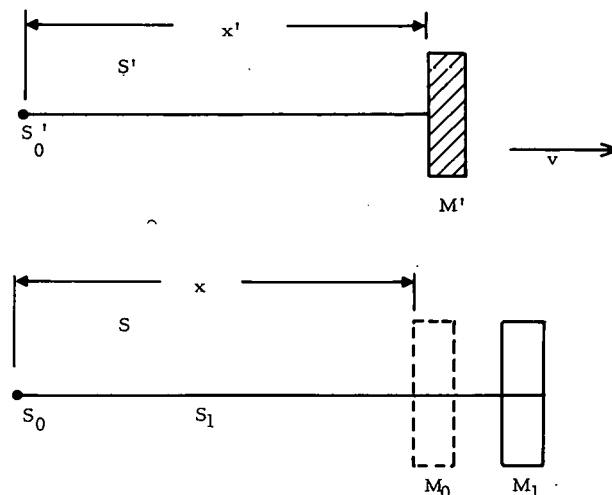


Fig. B-25. "Thought experiment" in which frame S' moves relative to frame S with a velocity v .

In S_0 these same events appear to be more complicated. At the time of emission the source S'_0 was at S_0 and the mirror M' at M_0 . (This statement has unique meaning only when referred to S' .) By the time of reflection the mirror has moved to M_1 . By the time the pulse returns to the source, S'_0 has moved to S_1 . The time interval Δt is therefore measured by two clocks, one at S_0 and the other at S_1 . Thus, Δt is not a proper interval [and therefore Eq. (34) applies to this situation].

The time required for the light to reach the mirror at M_1 after being emitted is $\overline{S_0 M_1}/c$. Thus, the distance $\overline{S_0 M_1}$ is

$$\overline{S_0 M_1} = x + v(\overline{S_0 M_1}/c), \quad (37)$$

where x is defined as the distance $\overline{S_0 M_0}$. Solving Eq. (37) for $\overline{S_0 M_1}$, we obtain

$$\overline{S_0 M_1} = \frac{x}{1 - v/c}$$

$$\overline{S_0 M_1} = \frac{x}{1 - \beta} \quad (38)$$

Similarly, the time required for the light to travel from the mirror at M_1 back to S_1 is $\overline{M_1 S_1}/c$. The distance $\overline{M_1 S_1}$ is

$$\overline{M_1 S_1} = x - v(\overline{M_1 S_1}/c)$$

$$\overline{M_1 S_1} = \frac{x}{1 + v/c}$$

$$\overline{M_1 S_1} = \frac{x}{1 + \beta} \quad (39)$$

The total time interval Δt is

$$\Delta t = \frac{\overline{S_0 M_1} + \overline{M_1 S_1}}{c}$$

$$\Delta t = \frac{2x}{c(1 - \beta^2)} \quad (40)$$

To find the relationship between x and x' we substitute Eqs. (36) and (40) into Eq. (41):

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - \beta^2}} \quad (41)$$

$$\frac{2x}{c(1 - \beta^2)} = \frac{2x'/c}{\sqrt{1 - \beta^2}}$$

$$x = x' \sqrt{1 - \beta^2} \quad (42)$$

This relation, called the Lorentz contraction, is also asymmetrical in x and x' , since it gives the relations between measurements of a proper length x' (at rest in S') and an improper length x (not at rest in S). The length x in S is definable only on assumption of the constancy of the velocity of light.

From Eq. (42) it is evident that a moving object appears contracted along the direction of its motion. In the Michelson-Morley experiment this effect accounts for the fact that the velocity of light is found to be independent of the direction of motion of the earth. The explanation for this is that the interferometer itself appears to be shrunk along the direction of motion when viewed from the "stationary ether" if there were such a thing. Because the velocity of the earth is small compared to that of light, the percentage of difference is small. However, even if the difference were quite large it could not be measured with a meter stick, because the meter stick itself would undergo the same effect. Since we are moving in the same reference frame as the meter stick, we could not see that it had shrunk when aligned with the direction of motion. However, it would be apparent to an observer in the same reference frame as the sun. (It should be noted that the Lorentz contraction does not entail a physical rearrangement of matter; it is a sort of optical illusion that is nonetheless quite real to an outside observer.)

Consider, for example, an electron at rest. (See Fig. B-26.) Its electrostatic equipotential lines may be represented by a family of concentric spheres. As it begins to move, we see these potential surfaces contract in the direction of motion, so that one of them looks like an ellipsoid. When the electron is traveling at the velocity of light, its electric field appears to have the shape of a flat plate.

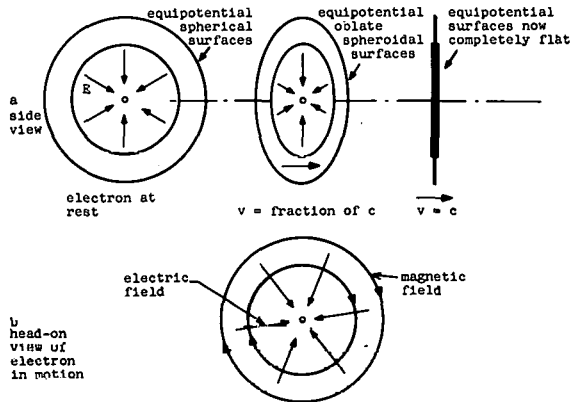


Fig. B-26. Representation of an electron at rest and in motion.

It is this apparent foreshortening of the electron's field when moving that explains how its electric field and the magnetic field resulting from the motion form a pattern like a spiderweb when viewed head on. (Fig. B-26). A complete discussion of this phenomenon is beyond the scope of this lecture, but may be found in textbooks on electrodynamic theory. An electron at rest has a spherically symmetrical electric field, but at rest, the electron has no magnetic field associated with it. As it moves, the component electric field vectors are shortened along the direction of motion, so that now its electric field is elliptically shaped. In addition a magnetic field is created as a result of the motion of the electron's charge. The charge appears like a current and every current has an associated magnetic field. When the electron approaches the velocity of light, the electric field approaches a disc shape with all of the electric field lines lying radially in the disc.

If a moving electron is decelerated or collides with some object and stops, this electromagnetic disc-shaped field pattern must nevertheless continue in its original direction of motion since it has momentum of its own. The form in which it continues is as a photon of light (Fig. B-27). The electric and magnetic field vectors (E and B) are perpendicular to each other. Physical optic experiments on light generally affect only the electric vector E .

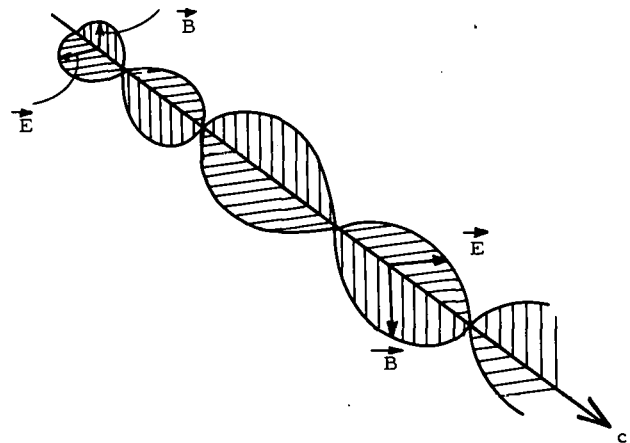


Fig. B-27. Schematic representation of a photon of light, which consists of an electric-field vector E perpendicular to a magnetic-field vector B . Photons are plane-polarized.

X-rays are formed by the process just described. The procedure involves allowing a beam of high-energy electrons to strike a heavy metal target. The electrons are stopped and the x-rays proceed on. Not every electron gives off an x-ray; in fact, fewer than 1% do. The probability for this occurring is directly related to the degree of flattening of the electrons electromagnetic field. Thus, the higher the velocity and consequently the energy of the electron, the greater is the probability of its giving off an x-ray when it strikes a metal target such as tungsten.

Experiment IV - The Synchronization of Clocks

In Experiment II the statement was made that to a moving observer two clocks separated a certain distance in a stationary frame would appear out of synchronism. This statement was made without calculating explicitly the error in synchronization δ . We will now consider this situation in more detail.

Consider two clocks synchronized in a stationary frame S and located a distance x apart (Fig. B-28). In a moving frame S' there is a single clock which records the time t_0' and t_1' as it passes each of the two clocks in S . The corresponding times in S are recorded by the two clocks in S and are denoted by t_0 and t_1 .

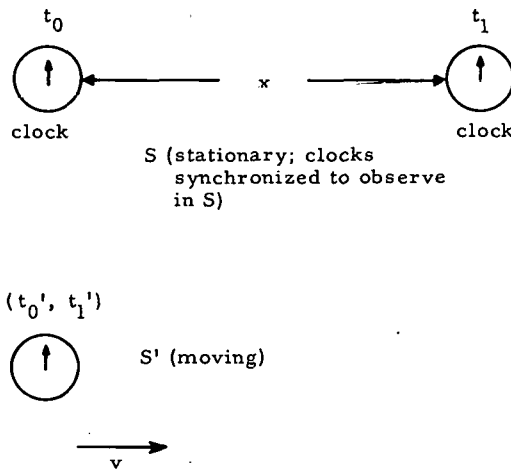


Fig. B-28. "Thought experiment" in which frame S' moves relative to frame S at a velocity v and the synchronization of the two clocks in S is checked by S' .

Because the difference $t_1' - t_0'$ is a proper time interval in S' , we can apply Eq. (41) (in reverse), giving

$$t_1 - t_0 = \frac{t_1' - t_0'}{\sqrt{1 - \beta^2}} \quad (43)$$

The observer in S' could apply the dilation equation (41) to the rates of the individual clocks at the ends of x , but would conclude that they are out of step by an amount δ :

$$t_1' - t_0' = \frac{t_1 - t_0 + \delta}{\sqrt{1 - \beta^2}} \quad (44)$$

Both observers have to obtain the same value for the relative velocity v between the frames, since if one obtained either a larger or smaller value, the frames would not be equivalent. Hence, we have

$$t_1 - t_0 = x/v \quad (45)$$

and

$$t_1' - t_0' = x'/v \quad (46)$$

However, since x is a proper length in S , we obtain

$$x' = x\sqrt{1 - \beta^2} \quad (42)$$

Combining Eqs. (44), (45), (46), and (42), we have

$$\delta = -\frac{x\beta^2}{v} \quad (47)$$

The negative sign indicates that in the opinion of S' the leading clock (t_0) in S should have indicated a larger time, $*$ in order to make the elapsed time between t_0 and t_1 as observed by S' smaller.

Summary of Experiments

These four experiments have thus demonstrated four kinematic relations:

- Distances transverse to the direction of motion are invariant.
- A time interval Δt measured in a frame moving with velocity $\pm v$ relative to a frame in which the time interval $\Delta \tau$ between two events is proper (i.e., the two events occur at one place) is given by

$$\Delta t = \frac{\Delta \tau}{\sqrt{1 - \beta^2}} \quad (48)$$

- The length Δx of a rod measured in a frame moving with velocity $\pm v$ relative to a frame in which the rod is at rest and has the proper length $\Delta \lambda$ is given by

$$\Delta x = \Delta \lambda \sqrt{1 - \beta^2} \quad (49)$$

- Two clocks, synchronous in a given frame, and separated by a distance $\Delta \lambda$ in that frame, appear to be out of synchronism as observed by a frame moving with a relative velocity $-v$ to the clock frame, by an amount given by

$$\delta = -\frac{\Delta \lambda v}{c^2} \quad (50)$$

The Lorentz Transform

The above thought experiments have led to kinematic relations that will now be combined into a general relation between the time and space coordinates of a particular event as observed from inertial frames in relative motion. This relation is called the Lorentz Transform, and is given by the relations

$$x = \frac{x' + vt'}{\sqrt{1 - \beta^2}}, \quad y = y', \quad z = z',$$

*The clock that is ahead (i.e., met first) is behind (in time).

$$t = \frac{t' + \frac{x'y}{c}}{\sqrt{1 - \beta^2}},$$

where $\beta = v/c$.

So far we have dealt with kinematics only, no forces or masses being involved. In order to preserve the conservation of momentum we are forced by the above relations and others for the transformation of velocities to conclude that a mass moving relative to us must be increased according to the relation

$$m = \frac{m_0}{\sqrt{1 - \beta^2}},$$

where m_0 is the rest mass (the mass we would measure if the body were brought to rest).

Another direct result of the Lorentz transformation and the conservation of momentum is that the total energy of a body--that is, the sum of its "rest energy" and kinetic energy--is

$$E = mc^2,$$

where the rest energy is m_0c^2 . It is this energy $E = mc^2$ that is conserved in any interaction. It is thus necessary to realize that each erg of energy has associated with it $1/c^2$ gram of mass. The chemical law of the conservation of mass and the First Law of Thermodynamics (the law of the conservation of energy of which we have not spoken) thus become one law.

Bibliography

1. E. N. da C. Andrade, The Birth of the Nuclear Atom, *Scientific American*, 94ff. (Nov. 1956).
2. M. Born, Atomic Physics (Hafner, New York, 1946). (Very good, not elementary, covers many subjects in modern manner)
3. T. B. Brown, Foundations of Modern Physics, Second Edition (Wiley, New York, 1948).
4. K. K. Darrow, The Quantum Theory, *Scientific American*, 47ff. (Mar. 1952).
5. S. Dushman, Fundamentals of Atomic Physics (McGraw-Hill, New York, 1951).
6. W. Finkelburg, Atomic Physics (McGraw-Hill, New York, 1950). (Advanced but qualitative)
7. G. Gamow, The Exclusion Principle, *Scientific American*, 74-86 (July 1959).
8. V. L. Ginzburg, Artificial Satellites and the Theory of Relativity, *Scientific American*, 149-160 (May 1959).
9. S. Glasstone, Sourcebook on Atomic Energy, Second Edition (Van Nostrand, Princeton, 1958), Ch. I.
10. A. E. Green, Nuclear Physics (McGraw-Hill, New York, 1955).
11. D. Halliday, Introductory Nuclear Physics, Second Edition (Wiley, New York, 1955). (Very good.)
12. G. P. Harnwell, Principles of Electricity and Electromagnetism, Second Edition (McGraw-Hill, New York, 1949).
13. G. P. Harnwell and J. J. Livingood, Experimental Atomic Physics (McGraw-Hill, New York 1933).
14. G. Joos, Theoretical Physics (Hafner, New York, 1947). (Touches on a great variety of subjects)
15. I. Kaplan, Nuclear Physics (Addison-Wesley, Reading, Mass., 1955).
16. F. K. Richtmyer, E. H. Kennard, and T. Lauritsen, Introduction to Modern Physics (McGraw-Hill, New York, 1942). (Best for relativity)
17. L. N. Ridenour, Modern Physics for the Engineer (McGraw-Hill, New York, 1954).
18. A. E. Ruark and H. G. Urey, Atoms, Molecules and Quanta (McGraw-Hill, New York, 1930).
19. H. Semat, Atomic Physics (Rinehart, New York, 1946). (Elementary)
20. J. C. Slater, Modern Physics (McGraw-Hill, New York, 1955).
21. R. L. Sproull, Modern Physics: A Textbook for Engineers (Wiley, New York, 1956). (Well written, very good, very modern.)
22. R. C. Tolman, Relativity, Thermodynamics, and Cosmology (Clarendon, Oxford, 1934). (High level, but very good, a classic.)

C. NUCLEAR PHYSICS

Warren Fenton Stubbins and Robert W. Kenney

Units and Description of the Nucleus

There are several systems of units in use in the sciences. They are the cgs, mks, and fps. Nuclear physicists use almost exclusively the cgs (centimeter-gram-second) system. The electrical and magnetic fields and the strengths of charges are expressed in three systems of units also. These are the emu, esu, and practical systems. In computations, the units must be adjusted to agree, i. e., be in one system only. (See Table C-I.)

Table C-I

Review (and extension) of units				
unit	mks	cgs	fps	atomic
(l) length	meter	cm	foot	atomic
(m) mass	kilogram	gram	pound	mass unit
(t) time	second	second	second	= amu
(F) force	newton	dyne	pound	
(E) work	joule	erg (dyne-cm)	foot-pound	electron volt
power	watt (joules/sec)	watt	horse-power	

We shall encounter a number of natural constants and useful values of certain quantities, some of which arise by definition. A tabulation of a number of these is given in Table C-II.

Table C-II

Approximate values for some physical constants			
Quantity	Symbol	Value	
Velocity of light	c	$=$	3×10^{10} cm/sec
Electronic charge	e	$=$	4.803×10^{-10} esu
Planck's constant	h	$=$	6.625×10^{-27} erg sec
Avogadro's constant N_0		$=$	6.02×10^{23} per gram mole
Rest mass of electron	m	$=$	9.1×10^{-28} g = 0.51 Mev
1 atomic mass unit (amu) =	$1/N_0$	$=$	1.66×10^{-24} g = 931 Mev
Ratio mass electron/proton		$=$	1/1836
One electron volt	1 ev =		1.602×10^{-12} erg
One million ev = 1 Mev,		$=$	1.602×10^{-6} erg

Einstein revealed much of the nature of nuclear physics in his study of special relativity, and the direct application of this theory is necessary

in order to permit analysis and understanding of the observations of nuclear systems. Thus we note the following definitions of the energy content of a system and other quantities. We accept the premise that mass and energy are equivalent and are connected by the relations

$$E = \gamma m_0 c^2 = m_0 c^2 / \sqrt{1-\beta^2} = m_0 c^2 / \sqrt{1-v^2/c^2} \\ = T + m_0 c^2, \quad (1)$$

where E is the total energy of the system; m_0 is the rest mass of the particle and v is its velocity; c is the velocity of light in a vacuum, a universal constant; $\beta = v/c$, the ratio of the particle velocity to the velocity of light; γ is defined by the above equation; and T is the kinetic energy. When the velocity of the particle is small compared with the velocity of light, Eq. (1) reduces to the classical, i. e., non-relativistic, expressions of our daily experience. For example, the value obtained when the velocity is zero is the rest mass in energy terms. The difference between the energy and the rest energy is the energy of motion, or what is called kinetic energy. Thus we define kinetic energy as the total energy less the rest energy. For a slow particle, $\beta \ll 1$, the kinetic energy by our definition becomes the classical

$$T = 1/2 m_0 v^2. \quad (2)$$

However, if the particle is not slow, a serious error results if one uses the classical expression for kinetic energy. One may decide to consider a particle as relativistic when this error exceeds an arbitrarily chosen limit. As an example, for a particle with a velocity 1/10th the velocity of light, the total energy is about 0.5% greater than the rest energy. This increase is the kinetic energy of the particle. As the velocity increases the energy increases, and as the velocity nears the velocity of light the total energy increases without bound. For a particle with the velocity of light its total energy would be infinite; thus we conclude that this velocity is not attainable by particles. In high-energy accelerators the energy of the particles may increase higher and higher, but the velocity only approaches the speed of light.

The linear momentum of a particle is defined as

$$p = m_0 v \frac{1}{\sqrt{1-v^2/c^2}} = m_0 \gamma v, \quad (3)$$

and a useful relation between the momentum, rest energy, and total energy is

$$E^2 = m_0^2 c^4 + p^2 c^2. \quad (4)$$

The simplest nuclei have masses of about 10^{-24} gram. An atomic mass unit -- a useful unit of the masses of nuclei and atoms -- is defined as 1/16 the mass of that isotope of oxygen having 16 nucleons. By this choice, 1 amu equals 1.6599×10^{-24} gram; the masses of all other atoms are very close to integers when expressed in atomic mass units. In energy units, 1 amu is equivalent to 931

Mev. The atomic masses are determined by comparison in mass spectrographs, by dividing the total mass of a sample of a substance by the number of atoms in it, or by other means. A relation between the mass of a substance and the number of atoms in it is given by Avogadro's number (6.02×10^{23}), whose numerical value is the reciprocal of the atomic mass unit.

The atomic size may be learned by x-ray examination of crystals, and nuclear sizes by scattering high-energy neutrons off the nucleus or by studying the theory of alpha decay. Atoms have a radius of about 10^{-8} cm and their nuclei about 10^{-13} cm. An interesting and revealing relation between the radius of a nucleus and the number of nucleons making up the nucleus is

$$r = r_0 A^{1/3}, \quad (5)$$

where A is the number of nucleons and $r_0 = 1.3 \times 10^{-13}$ cm = 1.3 fermi. A picture of the distribution of nuclei in the most dense material, a crystal, may be seen by expanding the nuclear size from its small actual size to the size of a baseball and allowing atomic dimensions to increase similarly. We thus find a 3-in. ball hung in space about every 2 miles in each direction. This wide spacing may make the diffusion of neutrons through matter a little more plausible, but because of the number of nuclei in matter (as reflected by Avogadro's number) a neutron can travel only a few centimeters through matter before a nuclear reaction occurs.

The density of nuclear matter may be calculated by dividing the mass of a nucleus by its volume. This is seen to be approximately 10^{14} grams per cubic centimeter, or about one billion tons per cubic inch.

It is necessary to define the terms "system" and "state of a system." For our purposes a system consists of the atomic and nuclear particles of interest prior to and after the nuclear process. These particles may be few (or even only one) before, and few or many afterwards. For example, a system of a decaying alpha emitter is initially the single nucleus, and afterwards the alpha particle and the recoiling residual nucleus. It may be initially a neutron and a uranium nucleus, and afterwards a number of free neutrons and several fission fragments. The state of a system is the description of the condition of each member of the system, including the energy, momentum, type of particle, charge, and other parameters.

An atom consists of a nucleus and a system of negative electrons about it. The un-ionized atom is neutral, i.e., the number of electrons is equal to the number of positive charges in the nucleus. An ionized atom is one from which one or more electrons have been removed, and is thus no longer neutral. The chemical properties of the atom are determined by the electrons surrounding the nucleus. The positive charge on the nucleus is equal to the sum of the charges on the protons contained in it. The other nuclear constituent is the neutrons. The protons and neutrons are both called nucleons, and their total number is called the mass number and is designated by A . The number of protons -- equal to the number of electrons -- is called the atomic number and is designated by Z . The difference $A - Z$ equals the number of neutrons and is called N ; i.e., $N = A - Z$.

By the definition of the atomic mass unit, the actual mass of the nucleus (a particular isotope) is very close to the numerical value of A when the mass is expressed in mass units. For example, the mass of Al^{27} is 26.9899 amu. This is the number of nucleons in the aluminum nucleus. We know that aluminum has 13 electrons about it, so that the number of protons, Z , is 13. Thus, there are $27 - 13 = 14$ neutrons in the nucleus. We may denote this by $^{13}\text{Al}^{27}$, but we should observe that the 13 is redundant because it and the chemical name "aluminum" both mean the same. Thus we may not wish to write both.

Any nucleus is completely identified by specifying its A and Z . Some nuclei (A, Z) are stable and found in nature, while others are radioactive, having various types of radiation associated with them. A radioactive nucleus is one that changes from one state to another spontaneously. The time required for half of the nuclei of a specific kind in any sample to decay is called the half life for that kind of nucleus. This time may be small or large. If it is so large as to be unmeasurable -- say, something like 10^{27} years -- we then say that the nucleus is stable; otherwise the nucleus is radioactive.

Any chemical element (i.e., of a given Z) may have several different values of A . These together are the isotopes of the chemical element; for example, the isotopes of carbon include C^{11} , C^{12} , C^{13} , and C^{14} . Whereas the isotopes of a particular element are chemically indistinguishable, for all practical purposes, the isotopes differ greatly in the following respects:

- a. They have different masses.
- b. They vary in stability.
 1. If they are unstable, the characteristic radiations differ.
 2. They have different probabilities of nuclear reactions (we say simply that their cross sections differ).
- c. The abundances of the stable isotopes found in nature differ.

For the radioactive isotopes, differences between the types and characteristics of emitted radiation permit each isotope to be identified, e.g. the half life and energy of radiation.

Those isotopes of different chemical elements which have the same mass number A are called isobars. Isobars can be separated chemically and they each have different nuclear properties. Some examples of isobars are: Be^{10} ($Z=4$, $N=6$), B^{10} ($Z=5$, $N=5$), and C^{10} ($Z=6$, $N=4$).

Conservation Laws

A conservation law is a statement which says that some property or quantity characteristic of a given system remains unchanged when the state of the system is altered in some way. In nuclear reactions the following properties are among those that are always conserved, i.e., the amount present before and after a reaction is the same:

- a. Energy. In applying this conservation law one must bear in mind the mass-energy equivalence given by $E = T + m_0c^2$ for each particle, where

E is the total energy, T is the kinetic energy, m_0 is the rest mass of the particle, and c is the velocity of light. The sum of total energies of all particles in a system is conserved in any change of state.

b. Electric charge. The total electric charge in a system is the same both before and after a nuclear interaction.

c. Linear momentum. Linear momentum p is defined as the relativistic mass m times the velocity v of the particle: $p = mv$. This expression is equal to $\gamma m_0 v$, where γ is the relativistic factor given by

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

Both before and after nuclear interaction, the state of a system has the same linear momentum, although it may be distributed differently among the various particles.

d. Angular momentum. For a particle of relativistic mass m rotating at a radius r about some point, its angular momentum is mvr, where v is the linear velocity of the particle. The angular momentum of a system is conserved in any interaction.

e. The number of nucleons (baryon number). Nucleons are each assigned a value of +1, and anti-nucleons a value of -1. For example, consider the reaction

$$\pi + p \rightarrow \pi + p + (p + \bar{p})$$

and the baryon numbers

$$0 + 1 = 0 + 1 + (1 - 1)$$

$$+1 = +1.$$

The pion, π , not being a nucleon, is assigned a value of 0 (see the lecture "Elementary Particles"). The proton, p, is given a value of +1 and the anti-proton, $\bar{p}$, a value of -1. Since the total value on the left side of the equation equals that on the right, the number of nucleons is conserved.

f. Lepton number. This conservation law, as well as the one immediately above, are discussed in the lecture "Elementary Particles." Also discussed in that lecture are three other properties of nuclei that are usually, but not always, conserved in interactions. They are:

g. Parity,

h. Isotopic spin, and

i. Strangeness.

A particular nuclear reaction may be prohibited because one or more of these parameters are not conserved. "Selection rules" apply to portions of a system and give the allowable changes in some of the properties listed above during changes in state of the system; e.g., during certain radioactive decays, the total angular momentum of the nucleus (exclusive of the emitted radiation) may change by 0 or 1 unit. The system as a whole still conserves total angular momentum. Nuclear-reaction equations can be balanced similarly to chemical-reaction equations by requiring charge, number of nucleons, etc., to be conserved.

Stability of Nuclear Systems

A nucleus is composed of $N = (A - Z)$ neutrons and Z protons. The configuration of a stable nucleus must be more stable than any other arrangement of the same number of nucleons. Otherwise, the nucleus will decay into the more stable arrangement (provided that no other restriction is violated in the decay, e.g. no selection rule is violated).

The energy holding the nucleus together is called the binding energy. We find that the nucleus is less massive than the sum of its separate parts. We call the energy equivalent of this mass difference the total binding energy, TBE. Its value is given by

$$\text{TBE} = (AM_P + NM_N - M_{A,Z})c^2, \quad (6)$$

where M_P is the mass of a proton, M_N is the mass of a neutron, and $M_{A,Z}$ is the mass of the nucleus. The greater the TBE, the more stable is the nucleus. Thus, for a given A, the less massive the nucleus the more stable it is.

As an example, consider a proton and a neutron at rest and far enough apart so that they do not interact. Their total rest-mass energy is $M_P c^2 + M_N c^2 = 938 + 940 = 1878$ Mev. If the neutron and proton then approach each other closely, so that they come within the range of the nuclear force, then they radiate an energy of about 2 Mev in combining to form a deuteron. The rest-mass energy of the deuteron M_D is

$$\begin{aligned} M_D c^2 &= M_P c^2 + M_N c^2 - \text{TBE} \\ &= 938 + 940 - 2 \end{aligned}$$

$$M_D c^2 = 1876 \text{ Mev.}$$

Thus, the deuteron is $2 \text{ Mev}/c^2$ lighter than the sum of the masses of its separate constituents.

Another concept of interest is the binding energy per nucleon. The average value of this may be found by dividing the total binding energy by the number of nucleons, A:

$$\frac{\text{TBE}}{A} = \frac{ZM_P c^2 + NM_N c^2 - M_{A,Z} c^2}{A}. \quad (7)$$

Figure C-1 shows the ratio TBE/A plotted against the mass number A. The heaviest elements have about 7.5 Mev TBE/nucleon, and the middle ones have about 8.5 Mev TBE/nucleon.

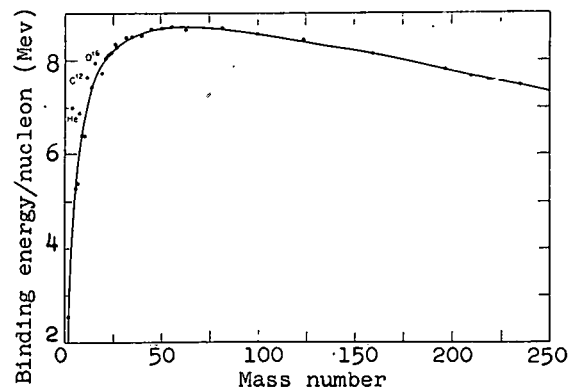


Fig. C-1. Total binding energy per nucleon vs the mass number.

If uranium fissions and becomes two middle elements, the energy available per nucleon is roughly $8.5 - 7.5 = 1$ Mev. the total energy released in the fission of uranium is about 235 time this, or about 200 Mev. If hydrogen is combined to make helium, energy is also released, about 7 Mev per nucleon. Thus, we see that the fusion of hydrogen is about seven times as productive of energy for each nucleon involved compared to the fission of uranium. For this reason a great deal of scientific effort is being expended in attempting to produce a controlled thermonuclear reaction, which could provide a nearly limitless source of power for peaceful uses.

We still must ask how we know a nuclear system is stable and under what conditions a nuclear reaction will take place. A necessary but not sufficient condition for a nuclear system to change its state spontaneously is that the energy E_f of the final state be less than the energy E_i of the initial state. We refer to the energy released or absorbed in a nuclear reaction by the symbol Q , which is given by

$$Q = E_i - E_f. \quad (8)$$

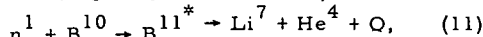
If energy is released, then Q is positive and the system is unstable. If this is the case, a nucleus can spontaneously change its state. This type of reaction is termed exoergic. On the other hand, if energy must be supplied to the system in order to make the reaction proceed, Q is negative and the system is stable. This reaction type is called endoergic. However, selection rules govern the way this happens and may prohibit change. As a rule of thumb, the greater the energy Q , the more likely it is that the change will happen. That is, we expect a shorter half life for large- Q values than for small.

If the mass (energy) of a nucleus is larger than the masses of the parts into which it could divide, we suspect that it will transfer to the lower energy state by dividing. On the other hand, if it is less massive than any combination of parts into which it may be divided, it is stable. Its stability is determined by a simple calculation of the masses involved. If we subtract from the mass of a nucleus the masses of all products of a particular division, we learn whether the initial or final state is the more massive. The calculation is as follows:

$$\frac{Q}{c^2} = M_{A,Z} - (M_{A_1 Z_1} + M_{A_2 Z_2} + \dots), \quad (10)$$

$$\frac{Q}{c^2} = M_{\text{initial}} - M_{\text{final}} \quad (9)$$

where $M_{A,Z}$ and $M_{A_1 Z_1}$, $M_{A_2 Z_2}$, etc., represent the masses of the initial nucleus and product nuclei, respectively. Consider a reaction in which a very slow neutron impinges upon a stationary B^{10} nucleus:



where n^1 is a neutron, B^{10} is a boron isotope, and B^{11*} is an intermediate nucleus (called the compound nucleus) which exists for a very short time. Lithium and helium are the final products. We have

$$\begin{aligned} Q/c^2 &= M_i - M_f \\ &= (1.00893 + 10.01618) \\ &\quad - (7.01822 + 4.00390) \\ &= + 0.00299 \text{ amu} \\ Q &= + 2.79 \text{ Mev,} \end{aligned}$$

since 1 amu equals 931 Mev. Because Q is positive, energy is released in the reaction. Therefore, the reaction proceeds spontaneously. The energy released appears as kinetic energy of the Li and He. This exoergic reaction is used for detecting neutrons, which cannot be detected directly because they carry no charge (see the lecture "Survey of Particle Detectors").

A neutron detector might consist of a proportional counter containing $B^{10}F_3$ gas. An entering neutron causes the reaction described by Eq. (11) to proceed. The reaction products, being charged, are then readily detected by the counter. (See the lecture "Counters used in High-Energy Physics" for further exposition.)

Nuclear Systematics

The known nuclei display many of their properties to good advantage when arranged in the form of the chart shown in Fig. C-2. This chart is divided into a large number of squares, with Z (the number of protons) plotted vertically and N (the number of neutrons) plotted horizontally. Each nuclide is then placed in the appropriate square according to its composition. It is evident that the nuclides arrange themselves about a diagonal line drawn across the chart. Those nuclides that are nearest this line are the most stable; those that are off this line tend to adjust their ratio of protons and neutrons by radioactive decay so as to move toward the line, thus becoming stable. Those above the line decay in a

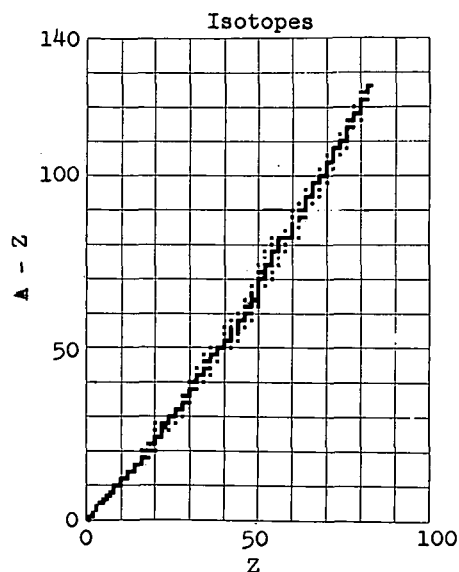


Fig. C-2. Number of neutrons vs the number of protons by stable nuclei.

manner which gives fewer neutrons and more protons to the final nucleus. This process is negative beta decay. Those below the line exhibit positive beta decay. The increased number of neutrons in the heavier elements is expected because of the relatively greater Coulomb repulsion between the protons in the heavier nuclei. A lower-energy state (for a heavy nucleus) can be achieved with fewer protons than neutrons.

We recall that the relation between the radius of a nucleus and the number of nucleons in the nucleus is given by

$$r = r_0 A^{1/3}. \quad (5)$$

This tells us that nuclear forces are short-range forces, by the following argument: The volume of a nucleus, is given by

$$V = (4/3) \pi r^3 = (4/3) \pi r_0^3 A. \quad (12)$$

The volume is proportional to the mass number (the number of nucleons). If each nucleon were attracted by each other one in the nucleus, we would expect the density of the nucleus to increase as the number of nucleons increases, since each additional one would be even more tightly bound. Because this is not observed we believe the nuclear forces to be short-ranged and the interaction roughly limited to adjacent nucleons.

When the Coulomb repulsion between protons has been accounted for in the nucleus, the binding between one proton and another proton is equal to that between two neutrons or between a neutron and a proton. Nuclear forces are therefore charge-independent. Three types of forces are known: gravitational, electrical, and specifically nuclear. The nucleus cannot be bound by only gravitational and electrical forces as we know them at long range.

The nature of the nuclear force is represented in Fig. C-3. It is evident that when two nucleons approach one another within a distance of about 1/2 fermi, the repulsion between them becomes very large. The greatest attraction occurs at a separation of about 1 to 2 fermi, and the attraction diminishes quite rapidly at smaller separations and finally becomes a strong repulsion at the "core."

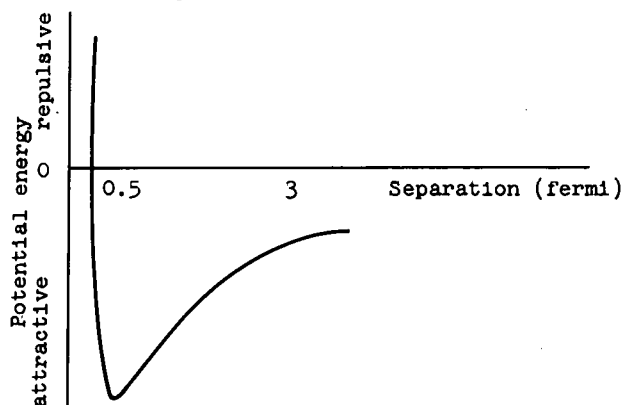


Fig. C-3. The nuclear force between two nucleons as a function of their separation, expressed in fermis (1 fermi = 10^{-13} cm).

No one has yet formulated an equation that adequately describes the nuclear force in all its details. However, it is known that this force is made up of several components, the details of which depend upon the particular nuclear model one discusses.

A nucleon is surrounded by a meson field through which it interacts with other nucleons, producing the nuclear force. This interaction via an intermediary pion is analogous to two charged particles interacting with each other by virtue of their electromagnetic fields. The π meson carries

the field from nucleon to nucleon in a fashion comparable to a photon propagating the electromagnetic field. For nucleons on the surface of the nucleus, the number of adjacent nucleons is fewer and they are less tightly bound. Minimizing the number of such nucleons leads to a more tightly bound nucleus (in the lowest energy state). This leads to a spherical nucleus.

Few nuclei are exactly spherical; some are prolate ellipsoids (footballs) and some are oblate ellipsoids (pocket watches). A nonspherical shape of a nucleus (and its electric charge) gives rise to an electric quadrupole moment. The nuclide with the largest quadrupole moment, Q , is Lu^{176} , for which $Q = +7 \times 10^{-24} \text{ cm}^2$. This represents a departure from the spherical of about 18%. The nuclide with the largest negative quadrupole moment is I^{127} , with a Q of $-0.6 \times 10^{-24} \text{ cm}^2$. This is about a 3% departure from spherical. When Q is positive the nucleus is a prolate ellipsoid (football); when it is negative the nucleus is an oblate ellipsoid (pocket watch); and when it is zero the nucleus is spherical. Most nuclei have positive quadrupole moments, but the departure from the spherical is never great.

It is also instructive to group the stable nuclei found in nature according to whether they have an even or an odd number of protons or neutrons (Table C-III). From Table C-III it is evident that those nuclei that have an even number of both protons and neutrons are by far the most abundant (stable). Next in abundance are those having an even number of protons but an odd number of neutrons. Those nuclei having both odd are the least stable. (However, the number of odd-odd unstable nuclides is quite large.) From the foregoing we conclude that pairing of either protons or of neutrons increases the stability of the nucleus.

Table C-III

Classification of nuclei as to whether they have an even or an odd number of protons (Z) or neutrons (N)		
Group	Designation	Number of stable nuclei
Z even, N even	even-even	163
Z even, N odd	even-odd	56
Z odd, N even	odd-even	50
Z odd, N odd	odd-odd	4

It has been found that none of the even-even stable nuclei possess spin (angular momentum). However, the other stable nuclei do have spin. The latter also have associated with them magnetic moments, by virtue of their rotating charges.

Another observation concerns the relative stability of some nuclei with respect to others. Those nuclei which have 2, 8, 20, 50, 82, or 126 neutrons, or 2, 8, 20, 50, or 82 protons are more stable (more abundant and have more isotopes) and are more tightly bound. This suggests that there is a shell structure of nucleons in the nucleus analogous to the shells of electrons about the nucleus.

In performing experiments it is convenient to have a formula that gives the mass of a nucleus very precisely, so that values may then be substituted

into the energy-balance equation to find whether or not a certain reaction will proceed. The semiempirical mass formula, which fills this need, is

$$M_{A,Z} = M_P Z + M_N N - a_1 A + a_2 A^{2/3} + a_3 \frac{(A/2 - Z)^2}{A} + a_4 \frac{Z^2}{A^{1/3}}, \quad (13)$$

where we have the following correction factors:

($a_1 A$) for the bulk heat of condensation (binding energy),

($a_2 A^{2/3}$) for the decreased binding at the nuclear surface,

$\left[a_3 \frac{(A/2 - Z)^2}{A} \right]$ for the unpaired nucleons in the nucleus, and

($a_4 Z^2$) for the mutual Coulomb repulsion of protons.

In Eq. (13) the terms a_1 , a_2 , a_3 , and a_4 are constants. Equation (13) can be simplified by combining the various constants. When this is done, we obtain

$$M_{A,Z} = 0.99391 A - 0.00085 Z + 0.014 A^{2/3} + 0.083 \frac{(A/2 - Z)^2}{A} + 0.000627(Z^2/A^{1/3}) + \delta, \quad (14)$$

where the term δ depends on the pairing of nucleons in the nucleus in question (see Table C-IV).

Table IV

Values of δ to be used in Eq. (14)

N	Z	δ
even	even	$-0.036 A^{-3/4}$
odd	even or odd	0
odd	odd	$+0.036 A^{-3/4}$

Radioactivity

The discovery of radioactivity and the subsequent experiments showed that the rates of decay and the types of radiation are independent of the atomic conditions in which the nucleus lives. Pressure, temperature, and chemical combination do not alter radioactive characteristics. This is a clue that radioactivity is a nuclear, not an atomic property. The rate of radioactive decay is proportional to the number of atoms of a particular species, since each nucleus decays independently of every other nucleus. These facts may be combined in the following equation, which applies when one has a very large number of nuclei:

$$dN/dt = -\lambda N, \quad (15)$$

where N is the number of nuclei of the specific type present at the time t , dN/dt is the rate of their radioactive decay, and λ is a proportionality constant called the decay constant. The dimensions of λ are sec^{-1} . Equation (15) may be integrated to give

$$N = N_0 e^{-\lambda t}, \quad (16)$$

where N_0 is the number present at the time $t = 0$ and e is the base of natural logarithms ($e = 2.718$). (See the lecture "Elements of the Calculus.")

The half life $T_{1/2}$ is defined as the time required for one half the nuclei present at $t = 0$ to decay. The half life is found by setting N equal to $1/2 N_0$, which gives

$$T_{1/2} = 0.693/\lambda. \quad (17)$$

Another useful concept is that of the characteristic time τ , which is usually referred to as the mean life of an unstable nucleus. By definition, the mean life is

$$\tau = 1/\lambda. \quad (18)$$

The fraction of nuclei remaining after one mean life ($\tau = 1$) is $1/e = 0.37$ (see Fig. C-4). At $\tau = 2$ the fraction remaining is $1/e^2 = 0.14$, etc. A more convenient presentation of the data obtained in this manner is obtained by plotting the logarithm of N/N_0 vs time. For a simple radioactive decay the data fall on a straight line (Fig. C-5). One usually plots the data on semilog paper directly.

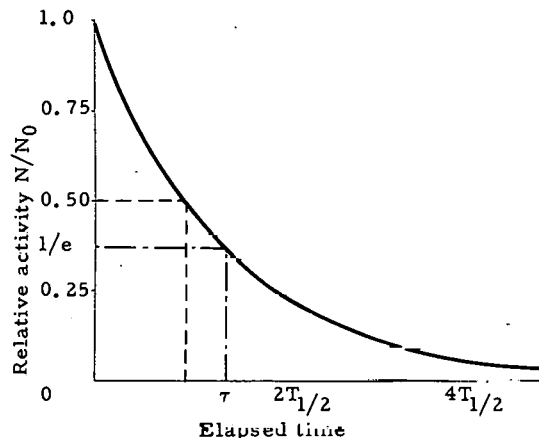


Fig. C-4. Relative activities vs time in units of half life.

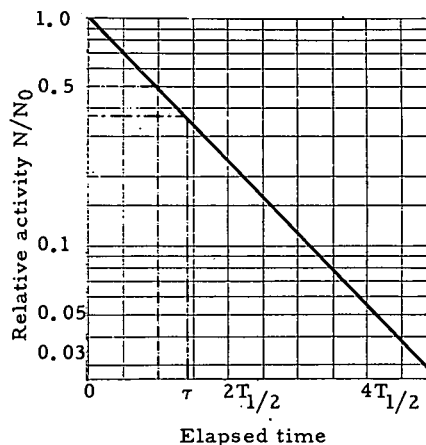


Fig. C-5. Relative activities vs time in units of half life.

If two radioactive species are present in the sample being measured, a plot of the measured radioactivity on semilog paper gives a curve (Curve a in Fig. C-6). Curve a represents the composite activity of the two species. It is evident that curve a is asymptotic to a straight line, i.e., to curve b, which represents the activity of one of the species present (the long-lived one). By subtracting curve b from curve a, one obtains curve c, which represents the other species (short-lived one) present. In this way, the decay constants for each of the two species may be obtained.

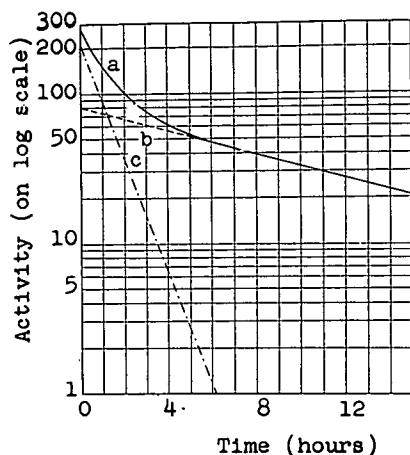


Fig. C-6. Activities vs time for two radioactive species in the sample.

When observing initially pure radioactive substance (parent) that decays into another radioactive substance or a chain of radioactive substances, one may observe the growth of the daughter radiations. Eventually, if the parent is long-lived, an equilibrium is established between the ratios of each substance. This is observed in the mixtures of radioactive isotopes found on the earth which may have relatively short half lives all stemming from a very long-lived parent, which continues to replenish them.

During constant production of radioisotopes in a reactor or accelerator, the loss by decay increases until eventually the loss rate is equal to the production rate. For example, after one half life of irradiation, one half of all the radioactive nuclei that will ever be available are present. The next half life of irradiation increases the amount present to only 75% of the maximum, while the third half-life exposure builds it to only 87.5%. For equal additional irradiation times a seriously diminishing return occurs. Extremely short activities can be irradiated in a short time to maximum activity, but must be removed very quickly to be counted. One technique is to transfer the material from target to counter by a "rabbit," which works like a pneumatic letter carrier in a department store. Another method is to have counters located so that the radioactive substance mounted on the periphery of a wheel is passed by them quickly after passing an irradiating beam. Electronic switching methods are also applied to this problem. Measurements of half lives as small as 10^{-3} sec are not difficult with these last two methods.

Two units of radiation are of interest. One, the curie, measures the rate of radioactive disintegration but says nothing about the energy, type of radiation, or hazard. One curie of a substance is that amount which has 3.7×10^{10} disintegrations per

second. This is approximately the number of disintegrations that occurs in one gram of radium per second. For any sample, the strength in curies is obtained by

$$\text{strength} = \frac{1}{3.7 \times 10^{10}} \lambda N, \quad (19)$$

where λ is the decay constant and M is the number of radioactive nuclei in the sample. N is computed by

$$N = \frac{N_0 \rho}{A} V, \quad (20)$$

where N_0 is Avogadro's number ($N_0 = 6.023 \times 10^{23}$ atoms per mole), A is the mass number, ρ is the density in grams per cm^3 , and V is the volume of the sample in cm^3 . A large decay constant λ means a high decay rate and a correspondingly short half life.

The roentgen measures the effect of radiation by indicating the amount of energy released by the radiation. The roentgen is defined as that quantity of x- or γ -rays which produces 1 esu of charge (of either sign) in 1 cm^3 of dry air at standard conditions of temperature or pressure. The rate of exposure is given in roentgens per hour.

Types of Radiation

The radiations from radioactive substances consist of three characteristic types. They were initially named α , β , and γ and were later identified as helium nuclei, fast electrons, and photons, respectively. They differ greatly in their properties, the way they interact with matter, and the processes that govern their emission.

Alpha particles

The longest-lived radioactive substances found in nature are the very heavy elements; these are alpha emitters. Alpha emission is defined by the transition $(A, Z) \rightarrow (A-4, Z-2) + \alpha$. Shorter-lived nuclei that emit alpha particles have been produced artificially. The range of half lives is from 3×10^{-7} to 10^{10} years, or a factor of about 10^{17} , but it is surprising that the energies of the emitted alpha particles vary only by a factor of about two, from 4 Mev to about 9 Mev. The higher the energy, the shorter is the half life.

The strong dependence of half life upon alpha energy was explained by Gamow and independently by Condon and Gurney when they applied a quantum-mechanical analysis to the problem. The nucleus is bound by being in a negative potential-energy state while having a positive potential barrier about it (Fig. C-7). For an outside alpha particle to enter the nucleus easily, it must have sufficient energy to surmount this barrier. On the other hand, an alpha particle within the nucleus can be emitted by tunneling through the potential barrier. The radius of a nucleus is about 10^{-12} cm, and the velocity of an alpha particle within the nucleus might be of the order of 10^9 cm/sec. Therefore, the alpha particle encounters the potential barrier about $10^9/10^{-12} = 10^{21}$ times per second. The probability P that this nucleus will decay in one second is

$$1/\tau = P = 10^{21} P_1, \quad (21)$$

where P_1 is the probability of an alpha particle being emitted in a single encounter with the potential

barrier. (See the lecture "The Statistical Analysis of Data.") The value of P_1 is given by

$$P_1 = e^{-\frac{2\pi zZe^2}{\hbar v}} \quad (22)$$

where z is the charge of the alpha particle, Z is the charge on the nucleus, e is the charge of an electron, $\hbar$ is the Planck constant divided by 2π , and v is the velocity of the alpha particle after it has tunneled through a large potential barrier.

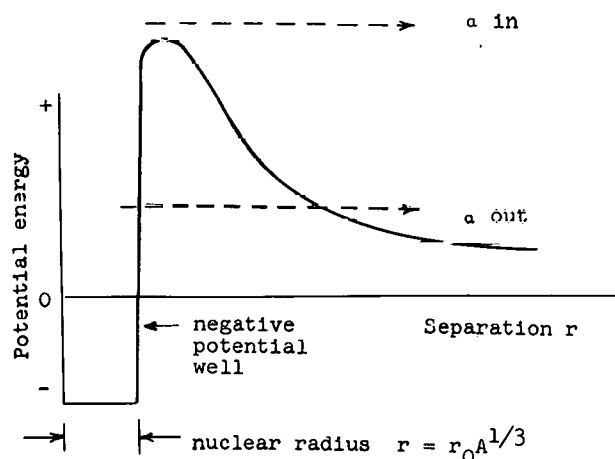


Fig. C-7. Coulomb barrier surrounding the nucleus inhibits entry or exit of charged particles.

Equation (21) gives the lifetime $1/\tau$ of the nucleus. However, there is a more convenient expression for arriving at this same value, called the Geiger-Nuttall rule, an experimentally determined relation:

$$1/\tau = cR^{57.5} \quad (23)$$

where c is a constant and R is the range of an alpha particle in air (expressed in cm). The constant c is different for the three naturally occurring radioactive families. For the uranium family, c is $10^{-42.3}$, for the thorium family c is $10^{-44.2}$, for the actinium family c is $10^{-46.3}$. The range R depends on the energy; for a typical nucleus a change of alpha energy from 6 to 7 Mev reduces the half life by a factor of a million.

Equation (23) thus gives a useful method for evaluating the range R of an alpha emitter when its lifetime can be measured, or conversely, for evaluating the lifetime if its range can be measured.

The mean life of an alpha emitter is very dependent on the radius of the nucleus. Quantum mechanics gives a clear relationship between them. Therefore, by observing the decay rate of an alpha emitter, one obtains a very fine determination of the radius of its nucleus. The nuclear radius is given by

$$r = r_0 A^{1/3} \quad (5)$$

where r_0 is 1.2 to 1.5×10^{-13} cm, depending on which aspect of the nucleus one is viewing.

All alpha particles from one nucleus have almost the same energy, but some are accompanied by gamma rays, which account for the small differ-

ences in α -particle energies in the decay of a single nuclear level. The nuclear change may be symbolized by the diagram shown in Fig. C-8.

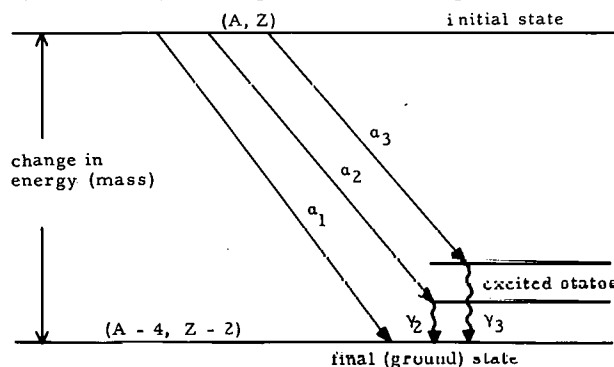


Fig. C-8. Change in nuclear states may proceed in one jump, $(A, Z) \rightarrow (A-4, Z-2)$, or in two jumps via intermediate excited states.

Experimentally, the photon, γ_3 , and α_3 can be shown to be associated. The fraction of all alphas which take a given path is the branching ratio for that path. For instance, if 1% of the decays go to the ground state directly, and 99% go to an excited state first, then the branching ratio for the ground state is $1/100 = 0.01$. The several energy levels to which the alphas decay indicate a condition in the nucleus analogous to that in the atom, in which quantized energy levels exist.

Beta particles

Negative beta particles are identical to those electrons that surround the nucleus in the atom and which are responsible for the chemical properties of an element. The nuclear transitions, which are called beta decay, are of three types. The principal type, called β^- decay, is illustrated by the decay of a neutron (n) into a proton (p) and a negative electron (β^-):

$$n \rightarrow p + \beta^- \quad [\text{see Eq. (24 Rev.) below}]$$

Since mass and energy must be conserved, and since the mass of the neutron and proton are known, one can calculate the energy of the β^- emitted. However, when the experiment is performed and the energies of the β^- particles emitted are measured, an unexpected result appears. Exceedingly few β^- particles actually have the calculated energy. Instead the energy distribution shown in Fig. C-9 is obtained. Where has the missing energy gone? To answer this, as well

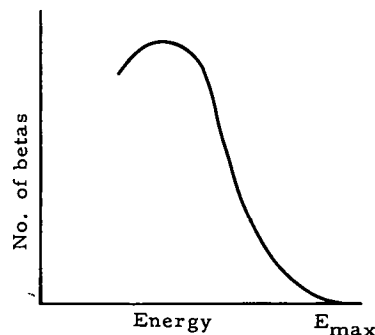


Fig. C-9. Observed counting rate of beta particles in β^- decay. The number calculated from Eq. (25) gives the maximum β energy, $E_{\max}$.

as to satisfy the other conservation laws, one must postulate an additional neutral particle as being involved in the reaction. This particle is called the antineutrino and is designated by $\bar{\nu}$. It carries away the missing energy. Therefore, Eq. (24) should properly be written

$$n \rightarrow p + \beta^- + \bar{\nu}. \quad (24 \text{ Rev.})$$

A general expression for β^- decay is

$$(A, Z) \rightarrow (A, Z + 1), \quad (25)$$

which states that when a beta-active nucleus of mass A emits a β^- particle, the mass of the nucleus remains (essentially) unchanged, because one neutron is converted into a proton. Thus, the final state has $(Z + 1)$ protons, the number of nucleons remaining the same. However, the nuclear charge is increased by one unit.

A second type of beta decay, called positron or β^+ decay, may be described by

$$(A, Z) \rightarrow (A, Z - 1). \quad (26)$$

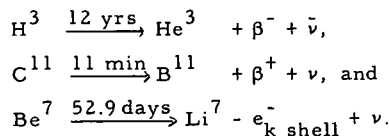
An example of this is the reaction

$$p \rightarrow e^+ + n + \nu, \quad (27)$$

in which a proton yields a positron (positive electron), a neutron, and a neutrino. In this reaction the mass of the nucleus remains (essentially) unchanged, but the nuclear charge is reduced by one unit, since it contains one less proton than formerly. When a positron comes to rest it annihilates with a negative electron, producing a characteristic radiation -- two gammas of 0.51 Mev each, which proceed in opposite directions. The presence of this radiation is a clue that positron emission is occurring.

A third type of beta decay is termed K capture. This occurs when a nucleus is in a state such that its energy can be lowered by the absorption of a negative electron. When one of the inner atomic electrons, e_k shell, gets sufficiently close to the nucleus this process may occur, decreasing the charge of the nucleus by one unit. [This process may also be described by Eq. (26).] After a K capture occurs, there is a hole in the k shell of the atom. Electrons in adjacent shells then cascade inward to the hole, thus emitting gamma radiation, which can be observed.

The three beta processes can be illustrated by the following:



One may expect nuclei rich in neutrons (above the stable ratio) to emit negative electrons in order to move to a more stable ratio, whereas proton-rich nuclei emit positrons. The fission of uranium produces neutron-rich fission products, which are negative-electron emitters.

Gamma rays

Gamma rays are emitted from a nucleus when it reorganizes itself by making a transition from one state to another of lower energy. The numbers and kinds of nucleons do not change, and the reaction

can be symbolized as

$$(A, Z)^* \rightarrow (A, Z) + \gamma, \quad (28)$$

where the asterisk indicates an excited nucleus. The energy of the gamma ray that is emitted is equal to k . This energy k may be exactly the amount that the excited nucleus possesses in excess of its ground state. On the other hand, the excited nucleus may go through several intermediate excited states before reaching the ground state. Such a cascade reaction may be represented by

$$\begin{aligned} (A, Z)^* &\longrightarrow (A, Z)^* + \gamma' \\ &\quad \downarrow \\ &\quad (A, Z)^{*''} + \gamma'' \\ &\quad \quad \downarrow \\ &\quad \quad (A, Z) + \gamma. \end{aligned} \quad (29)$$

In such a case, the original excited nucleus probably cannot proceed directly to the ground state, owing to certain selection rules.

The amount of reorganization that must take place in the nucleus determines the lifetime of the nucleus. The lifetimes of gamma emitters are in general very short, some as short as 10^{-14} sec. However, some nuclei exist in an excited state long enough for their lifetimes to be measured. These are called *isomers*. Isomers may live for seconds, hours, or even years. What prevents some nuclei from jumping immediately to the ground state that is their angular momentum must undergo a large change in the transition. For example, a nucleus of ground-state spin 0 might be in an excited state with a spin of $9/2$. This excited state must possess only a small excitation energy so that its transition will likely be directly to the ground state. It must lose all its spin in one highly forbidden transition to the ground state. Since it can lose spin easily only in increments of $1/2\hbar$, and further since there is only a small excitation energy to drive the reaction, its lifetime is prolonged.

The reorganization of the nucleus changes the distribution of charge, which leads to the emission of electromagnetic energy in the form of a photon. This process is parallel to the emission of electromagnetic energy from an accelerated electric charge.

A process that de-excites the nucleus but results in the emission of a beta instead of a gamma is called *internal conversion*. Internal conversion occurs when an electron of an inner shell of the atom perturbs the nucleus by being in the region of the nucleus. If energy can be released it may be transferred to the electron rather than coming out as a

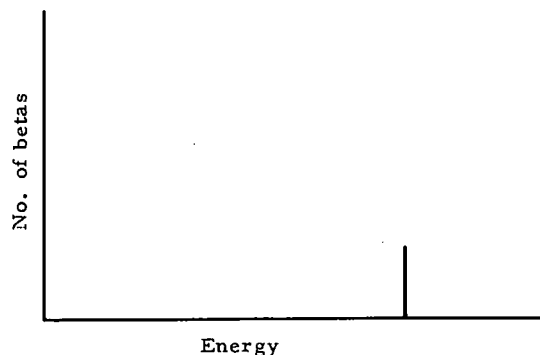


Fig. C-10. Energy spectrum of an internal-conversion transition.

photon. The energy of the beta, E_β , is related to the gamma energy, k , according to

$$E_\beta = k - W, \quad (30)$$

where W is the energy required to remove the electron from its atomic shell. W is called the work function. In contrast to the betas from a beta emission, the internally converted betas have a discrete energy. The graph obtained when the spectrum of beta particles is plotted vs energy, a sharp spike is obtained for the internal-conversion electrons. (Fig. C-10.)

Sometimes a spectrum such as that shown in Fig. C-11 is obtained. The broad spectrum indicates that a beta decay to an excited state occurred, and the sharp spike shows that it was followed by an internal-conversion transition to the ground state or to a lower energy excited state. Thus, by knowing the details of the states in the final daughter nucleus, and by knowing the characteristics of the radiation involved in the transitions, one can learn the properties of the parent nucleus.

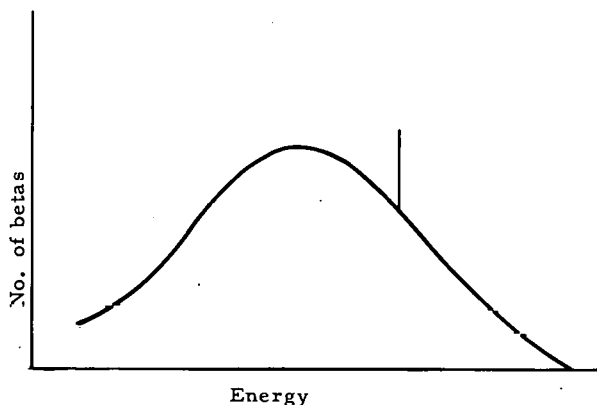


Fig. C-11. Energy spectrum of a beta decay followed by an internal-conversion transition.

In internal conversion a hole is left in the electron shell. Characteristic x-ray cascades are then produced when electrons drop into the k and l shells, which are filled up successively.

In addition to nuclei, certain other basic particles are radioactive. We have already seen that the neutron (a nucleon) decays into a proton, an electron, and an antineutrino [Eq. (24 Rev.)]. In fact, most of the "elementary" particles are unstable, including all the mesons and hyperons. The study of various processes among the mesons and hyperons forms a major portion of the work at this Laboratory in the high-energy field. It is discussed elsewhere in this volume in detail.

Interaction of Radiation With Matter

Each type of radiation behaves differently as it passes through matter. Energy is usually transferred to the matter, which then becomes ionized. This ionization permits detectors to note the presence of the radiation and is the basis for most counters.

Heavy Charged Particles Traversing Matter

Heavy charged particles include all particles heavier than electrons. Their interaction with matter is characterized by ionization along their path.

These heavy particles brush the very light electrons off the atoms in their way and the heavy particles themselves scatter very little in the process. Infrequently, the particles suffer a deflecting collision, which is caused by a nuclear encounter.

The intensity of ionization differs along the path as the particle velocity changes. Figure C-12 shows a Bragg curve for alpha particles. The intensity of ionization in the matter increases as the particles lose energy until a maximum is reached near the end of the path. The ionization per unit length then drops rapidly as the particles come to a stop. The initial low rate of ionization is caused by the short interaction time as the fast particles pass the atoms of the material. As the particle slows down, its interaction time with these atoms becomes longer, at first, but finally its energy becomes too small to cause ionization and the energy loss then drops to zero.

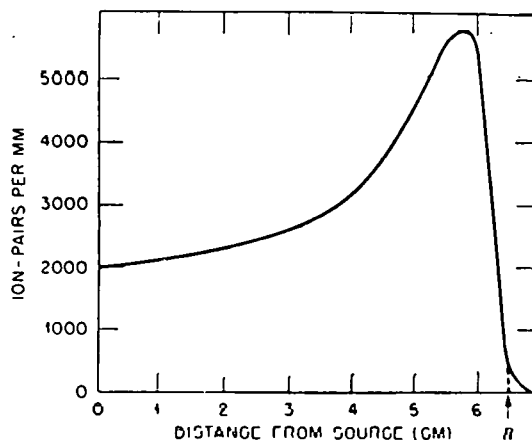


Fig. C-12. Bragg curve -- the number of ion pairs per unit length vs the path length in matter.

A beam of heavy particles of a given energy will have a more or less sharply defined range in any given material, since about the same number of ionizing interactions is required to stop each particle in the beam. This is seen in Fig. C-13. One may define the range as the average path length of a given particle of a given energy.

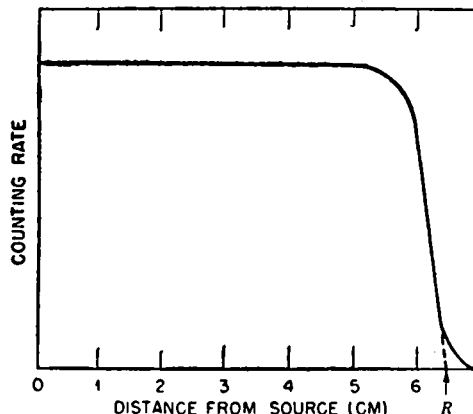


Fig. C-13. Number of heavy particles detected vs the path length in matter.

The rate of energy loss by a particle per increment of path length is $-dE/dx$. When $-dE/dx$ is plotted against the energy E of an incoming parti-

cle, the curve represented in Fig. C-14 is obtained. This contains the same information as a Bragg curve, but is presented in a different way and is more complete. Region A is of little interest to high-energy physicists. In region B the curve is proportional to $1/v^2$, where v is the velocity of the particle. The point of minimum ionization occurs at an energy equal to approximately the rest-mass energy m_0c^2 . For a proton this is about 1 Bev; for an electron it is about 0.5 Mev. For energies higher than this, relativistic effects enter, which result in increased ionization. In region D the curve reaches a plateau, owing to an effect called screening.

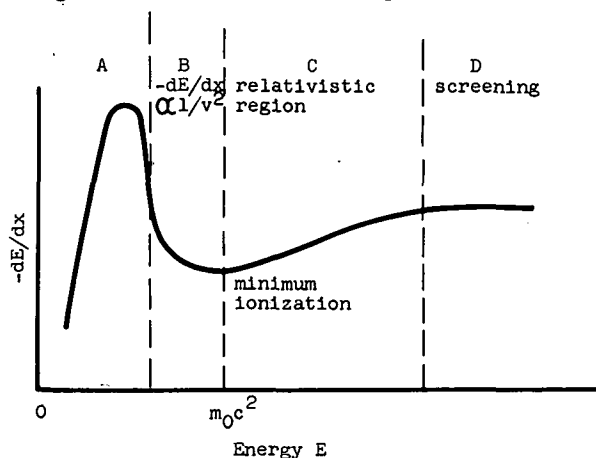


Fig. C-14. Rate of energy loss per increment of path length vs energy for a charged particle traversing matter.

The stopping-power curve in regions B, C, D, and E (Fig. C-13) is given by

$$-\frac{dE}{dx} = \frac{4\pi Z^2 e^4 n}{mv^2} \left(\ln \frac{2v^2 m}{I} \gamma^2 - \beta^2 \right) \quad (31)$$

for a particle of atomic number Z and velocity v traveling through an absorber containing n electrons per cm^2 , where e is the charge of an electron, m is the mass of an electron, γ is the relativistic factor $\gamma = 1/\sqrt{1 - v^2/c^2}$, β equals v/c , and I is the ionization potential of the absorber. The ionization potential is the number of electron volts (ev) required, on the average, to tear an electron off an atom in the absorber. The ionization potential is roughly 30 ev for each ion pair produced, but differs for each element. It should be noted that Eq. (31) is independent of the mass of the incoming particle, as long as the particle is much heavier than an electron.

For convenience, Eq. (31) has been evaluated for various types of particles traversing different materials with specified energies. These values have been tabulated in UCRL-2301 and UCRL-2426 Rev., Vol. II,* which one frequently uses for solving range-energy problems.

For instance, in designing an experiment, one may want to know the energy E' of an emergent particle after it passes through a specified absorber of thickness t , where the energy E of the incident

particle is known. The procedure is, first, to consult the range-energy table and find the range R of the particle in the specified material. The energy E' of the emergent particle is related to the range R' by

$$R'(E') = R(E) - t. \quad (32)$$

Suppose that a 300-Mev proton is incident on a copper plate of thickness $t = 30 \text{ g/cm}^2$. Consulting the table, we find that the range R is 76.44 g/cm^2 . Therefore, the range R' is

$$R'(E') = 76.44 - 30$$

$$R'(E') = 46.44 \text{ g/cm}^2.$$

Again consulting the table, we find that the energy E' corresponding to a particle of range $R'(E') = 46.44 \text{ g/cm}^2$ is about 230 Mev.

Electrons Traversing Matter

The ionization loss by electrons is important, but it is not characterized by the intense ionization and the straight paths of the heavy charged particles. Electrons are more lightly ionizing than heavy particles of the same energy and are scattered about as they encounter atomic electrons (of equal mass) in their path. Thus, their range is greater than that of heavy particles of the same energy, but they tend to lose their initial direction rapidly. When the beam emerges from an absorber it is no longer collimated, but rather fans out in every direction. However, the angular distribution is normal about the direction of the incident beam (Fig. C-15). The number of electrons emerging at any angle θ , with respect to the forward direction, is

$$I(\theta) = \frac{I(0)}{\langle \theta \rangle} \sqrt{\frac{2}{\pi}} e^{-\frac{\theta^2}{2\langle \theta \rangle^2}}, \quad (33)$$

where $I(\theta)$ is the number of electrons incident at $\theta = 0^\circ$, e is 2.718, and $\langle \theta \rangle$ is given by

$$\langle \theta \rangle = \frac{21}{E} \sqrt{t}, \quad (34)$$

where E is the energy of the incident beam in Mev and t is the thickness of the absorber in radiation lengths.

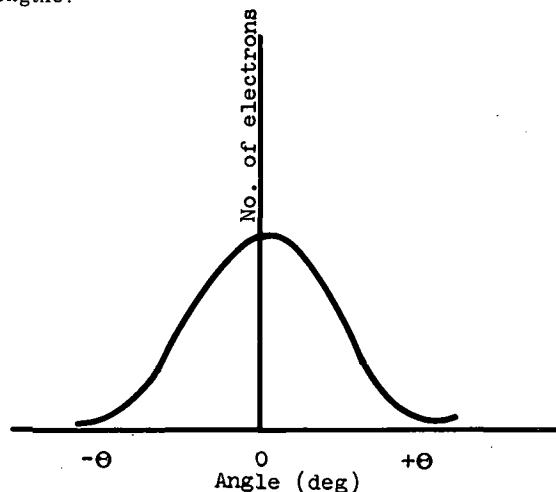


Fig. C-15. Angular distribution of electrons emerging from an absorber.

* M. Rich and R. Madey, Range-Energy Tables, UCRL-2301, March 1954; J. H. Atkinson, Jr., and B. H. Willis, High-Energy Particle Data, UCRL-2426 Rev., Vol. II, June 1957.

As electrons are deflected by their encounters with matter, a deceleration (change in velocity) occurs. Since they are charged particles, the deceleration causes them to lose energy by electromagnetic radiation. This is the source of x-rays, and for high-energy electrons this is known as *Bremsstrahlung*. The *Bremsstrahlung* loss exceeds the ionization loss when the electrons are very relativistic, which occurs at energies of a few Mev in some materials. The ratio of *Bremsstrahlung* loss per cm to ionization loss per cm is

$$\frac{\text{Bremsstrahlung loss/cm}}{\text{ionization loss/cm}} = \frac{ZE \text{ Mev}}{800} \quad (35)$$

where Z is the atomic number of the absorber and E is the energy of the incident electron beam in Mev.

The critical energy E_c is defined as that energy at which the *Bremsstrahlung* loss equals the ionization loss, i. e., for $ZE/800 = 1$. At energies near to and above the critical energy, the phenomenon of cascade showers is observed (Fig. C-16). A high-energy electron traversing matter radiates part of its energy as an x-ray (*Bremsstrahlung*). The x-ray, after traveling a certain distance, converts all its energy into an electron-positron pair. The electron in turn radiates part of its energy as another x-ray; the positron does likewise. In the meantime the original electron again radiates part of its remaining energy, producing another x-ray, which in turn converts its energy, producing another x-ray, which in turn converts its energy into an electron-positron pair, and so on. In this fashion, a single incident electron produces a cascade shower of x-rays and electron-positron pairs.

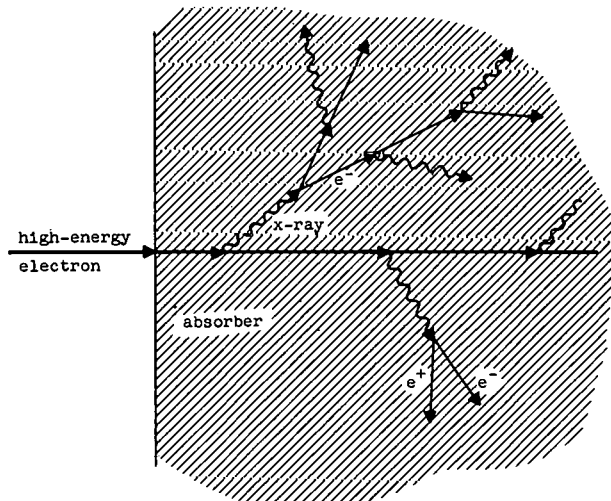


Fig. C-16. Cascade shower of x-rays and electron-positron pairs produced by a single high-energy electron incident on matter.

Gamma Rays Traversing Matter

The interaction with matter of gamma rays or photons occurs by way of three principal processes at low and moderate energies: the photoelectric effect, the Compton effect, and pair production. A single gamma ray traverses matter without interacting until it suffers a single catastrophic collision in which it is totally absorbed. For this reason the attenuation of gamma rays follows an exponential law instead of having a well defined range, as do heavy particles (which lose their energy approximate-

ly continuously in a large number of interactions with atomic electrons, only a small fraction of the particle's energy being transferred in a single interaction):

$$I = I_0 e^{-\mu x}, \quad (36)$$

where I is the intensity after passing through a thickness x of the material, I_0 is the initial intensity, μ is the absorption coefficient, and e is 2.718. (See Fig. C-17.)

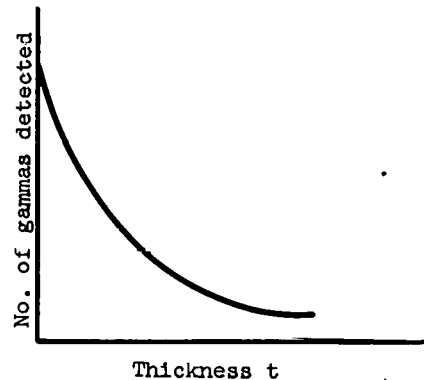


Fig. C-17. Number of gamma rays surviving after traversing a specified thickness of matter.

The photoelectric effect occurs if a gamma ray interacts with an atom and removes a bound electron from an atomic shell (Fig. C-18). The entire energy k of the gamma is imparted to the electron, which leaves with an energy E_{elec} equal to k minus its binding energy in the atom. This process is most important for low-energy gammas (below 1 Mev).

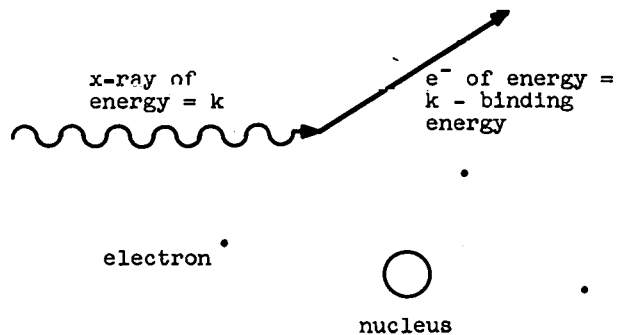


Fig. C-18. The photoelectric effect.

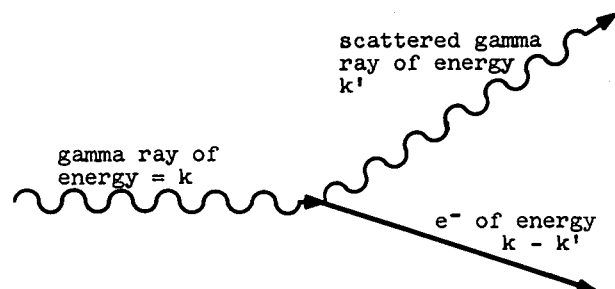


Fig. C-19. Compton scattering.

The Compton effect is the interaction of a gamma ray of energy k with an electron which is essentially free, that is, an electron whose binding energy $\ll k$. (See Fig. C-19.) The application of the rules of conservation of energy and momentum permits this problem to be solved to show that the gamma ray may lose only a part of its energy to the recoiling electron. This effect is important up to gamma energies of several Mev (see Fig. C-20).

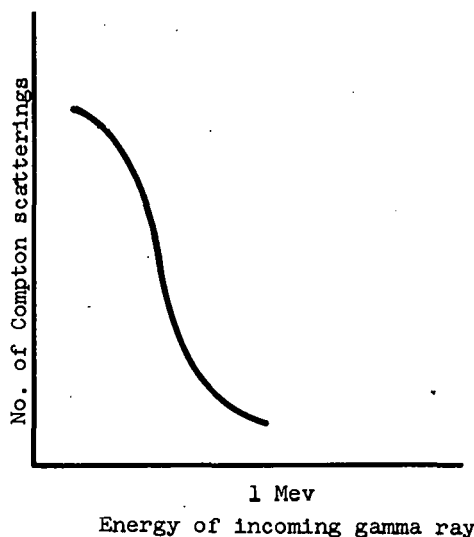


Fig. C-20. Probability curve for the occurrence of the Compton effect.

Electron-positron pair production becomes possible when the gamma energy equals or exceeds twice the rest mass of the electron, i.e., gamma energies in excess of 1.02 Mev. (See Fig. C-21.) A gamma ray may be converted into an electron-positron pair in the presence of a heavy nucleus.

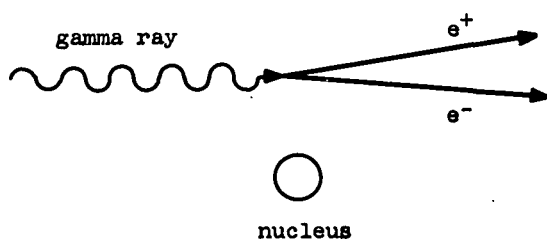


Fig. C-21. Electron-positron pair production.

(See Fig. C-22.) This is the interaction of greatest importance for high-energy gammas, i.e., those above several Mev. The gamma energy in excess of the two rest masses is shared (not necessarily equally) by the electron-positron pair and becomes their kinetic energy.

The probability for any one of these three effects occurring as a photon traverses matter is simply the sum of the three probability curves (i.e., Figs. C-15, C-18, and C-19). This total probability or total cross section is denoted by the symbol σ_T .

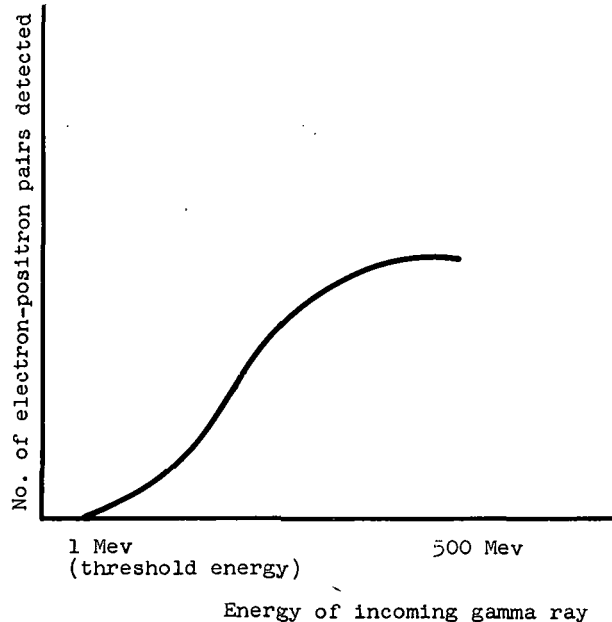


Fig. C-22. Probability curve for electron-positron pair production.

The cross section σ_T for photons interacting with light elements is plotted in Fig. C-23. At low energies the photoelectric effect and Compton effect occur with about equal probability. At high energies, pair production is predominant. In Fig. C-23, the cross section for photons interacting with heavy elements, it is seen that the Compton effect is not manifest.

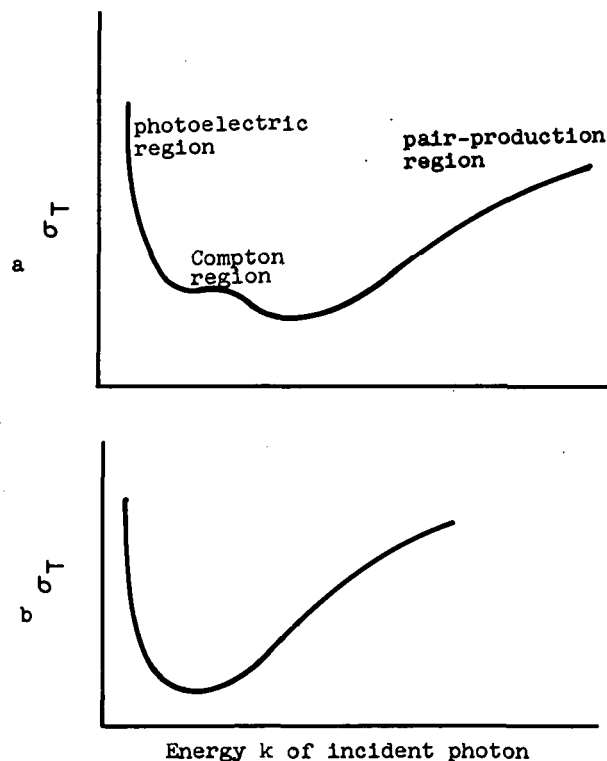


Fig. C-23. Cross section σ_T for photons interacting with (a) light elements or (b) heavy elements.

Cross Section

In all nuclear reactions one is interested in being able to say how probable the reaction is, i. e., under given conditions how frequently will the specific reaction occur? The measure of this probability is called the cross section and it is expressed in cm^2 per nucleus or per particle, whichever is applicable.

The cross section may refer to all processes, when it is called total cross section; or it may refer to a particular process, and it is named to indicate that. For example, one refers to the cross section for fission, for scattering, etc. Specific cross sections are designated as $\sigma(n, f)$ for neutron-induced fission, $\sigma(\gamma, n)$ for photon production of neutrons, etc. Cross sections, total or specific, vary with the process, the energy of the bombarding particle, and the target nucleus.

The convenient unit of cross section for a nucleus is the barn. One barn equals 10^{-24} cm^2 . The area of a nucleus as computed from the radius is about this size. Cross sections for nuclear processes usually have values in the neighborhood of a barn, with a range from a millibarn to 1000 barns. Some nuclear cross sections lie outside these values, and some high-energy cross sections are measured in microbarns (10^{-30} cm^2). At the threshold of any reaction, the cross section is zero, by definition, so that a cross section may take values between zero and these foregoing small-sounding numbers, depending upon the circumstances involved.

The cross section applying to the nucleus alone is called the microscopic cross section and is usually designated by σ . The microscopic cross section when multiplied by the number of nuclei per cm^3 of the substance gives the macroscopic cross section, sometimes designated by Σ :

$$\Sigma = \frac{N_0}{A} \rho \sigma. \quad (37)$$

The reciprocal of the macroscopic cross section is the mean free path in the particular matter

under the conditions implied by ρ , its density. N_0 is Avogadro's number and A the mass number.

Differential cross section is the term applied to the probability that a nuclear reaction will result in the particle's going out at an angle to the incident beam. Differential cross section is designated by $d\sigma/d\Omega$, in which Ω is the solid angle into which the particle is directed. The integral of the differential cross section over all solid angles is the microscopic cross section for the process under study. The differential cross section gives added information about the nuclear structure and nuclear forces.

Bibliography

1. H. A. Bethe, What Holds the Nucleus Together?, Scientific American, 58 ff. (Sept. 1953).
2. D. Halliday, Introductory Nuclear Physics, Second Edition (Wiley, New York, 1955).
3. S. Glasstone, Sourcebook on Atomic Energy, Second Edition (Van Nostrand, Princeton, 1958), Ch. XII.
4. Jones, Rotblat, and Whitrow, Atoms and the Universe (Scribner, New York, 1956).
5. M. G. Mayer, The Structure of the Nucleus, Scientific American, 22 ff. (Mar. 1951).
6. R. E. Peierls, The Atomic Nucleus, Scientific American, 75-82 (Jan. 1959).
7. R. F. Post, Fusion Power, Scientific American, 73 ff. (Dec. 1957).
8. V. F. Weisskopf and E. P. Rosenbaum, A Model of the Nucleus, Scientific American, 84 ff. (Dec. 1955).

D. SOLID-STATE PHYSICS - INTRODUCTION

Roger W. Wallace

Solid-state physics is largely a development of the past twenty years. Today it is one of the most active areas of research, owing largely to the development of the transistor and its very important role in the electronic industry.

Transistors are formed of semiconducting materials, i. e., those having a resistivity roughly midway between that of good conductors and good insulators. The semiconducting materials most often used in transistors are silicon or germanium crystals. In this lecture we shall discuss the properties of crystals and consider how, under certain conditions, they can be used to rectify or amplify electrical currents.

It is easy to understand why a diatomic crystal like sodium chloride can be formed, since there is a simple electrostatic attraction between the positive sodium ions and the negative chlorine ions. However, it is not obvious why sodium metal alone or any other monatomic substance can form a crystal. In fact, crystal formation by an element cannot be explained on a classical basis. Instead, one must consider a quantum-mechanical picture.

Figure D-1 plots the potential energy between two atoms as a function of the separation of their centers. As the two atoms approach one another, the potential and kinetic energy of their electrons change. At an infinite separation the potential energy may be regarded as zero. As the atoms come together the energy of their outermost electrons decreases, producing an attractive force between the atoms. This continues, until at a separation of about one angstrom the energy increases quite rapidly, producing a repulsive force between the two atoms. If this did not occur, the electrons in their outer shells would approach so closely that they would have identical quantum numbers. This is prohibited, as explained by the Pauli exclusion principle. When a separation dis-

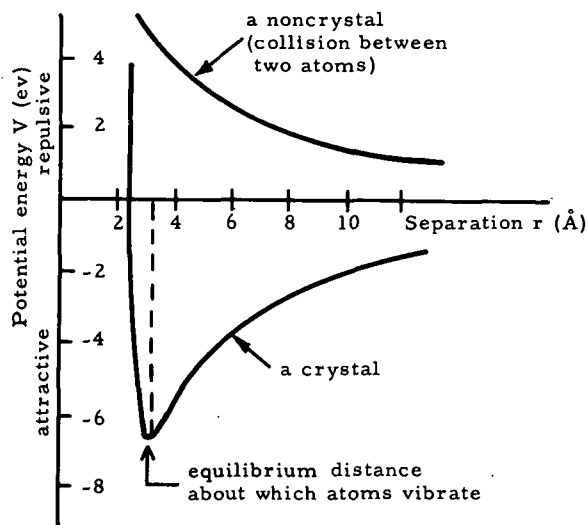


Fig. D-1. Energy between two atoms as a function of the separation of their centers.

tance is reached at which the repulsive force balances the attractive force, the atoms no longer come together but merely vibrate about this equilibrium position, thus forming a crystal.

On the other hand, if the attractive force between two atoms is insufficient to cause binding, no crystal is formed. This case is represented by the upper curve in Fig. D-1.

A more sophisticated explanation of crystal formation may be obtained by considering the wave functions associated with the outer valence electrons of two atoms, say of sodium. These two outer valence electrons may be thought of as sitting in separate boxes or potential wells (Fig. D-2a), which of course

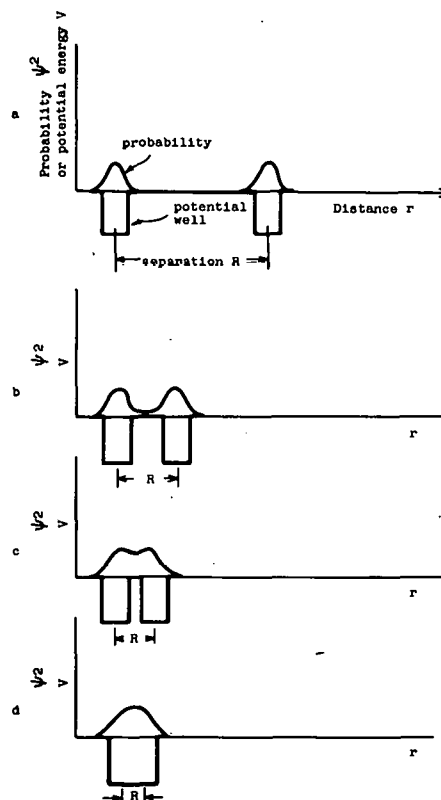


Fig. D-2. Case A: probability curves for two atoms (having wave functions of the same sign) at different separations.

represent their potential energy. At a certain separation, the two electrons do not influence each other. The probability is greatest that each electron will be in the center of its box, and less that it will be at the edge of the box. This is represented by the probability curves in Fig. D-2. [It will be recalled that the probability P is equal to the wave function squared, i. e., to ψ^2 . (See Section B "Atomic Physics.")]

As the two sodium atoms are brought together, the picture changes as shown in Fig. D-2b, c, and d. Eventually the probability curves merge. They must be joined by a smooth curve, since the combined probability curve must remain well-behaved. This means that the junction point does not go all the way down to zero. Finally, the two potential wells come together, and there is no longer any barrier between them. The combined probability curve now appears as one big hump, the depression in the center having disappeared.

The foregoing picture is valid for atoms whose wave function ψ never changes sign. The probability is always positive regardless of the sign of the wave function, since ψ^2 is always positive whether ψ is positive or negative.

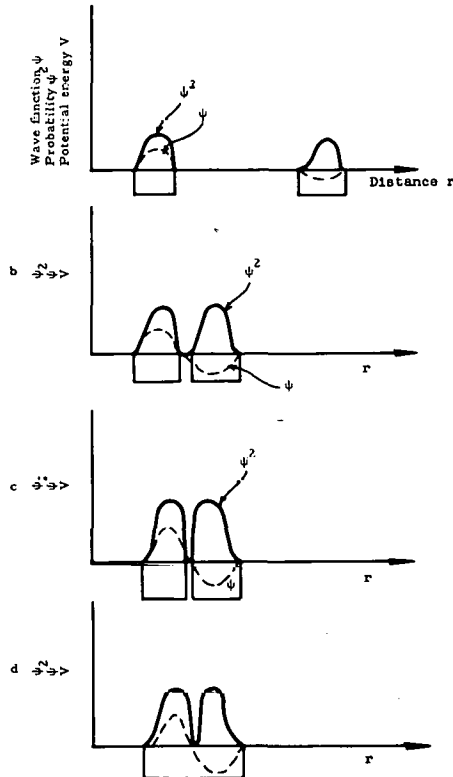


Fig. D-3. Case B: probability curves for two atoms (having wave functions of opposite signs) at different separations.

On the other hand, when the wave function ψ changes sign, a different picture results (Fig. D-3). If by chance, as will be the case in half of all the pairs of atoms considered, one wave function is negative, then as the two atoms are brought together in order for the probability curves to join, the combined curve must always go through zero. Therefore, a double-humped probability curve results. This means that the electrons have different physical positions, i. e., there are two points of greatest probability for the electron positions.

The first case (A), in which ψ never changes sign, is called the symmetrical case. It is a lower-energy state than the second case (B), called the antisymmetrical case, in which ψ changes sign. This situation is analogous to a vibrating piano string. The fundamental (one hump) has the lowest energy associated with it, whereas the first harmonic (two humps) has a higher energy.

In a crystal containing, say, 10^{20} atoms one finds 2×10^{20} energy levels associated with a given state, e. g. the 1s state. (There are 10^{20} levels, each allowing one spin-up and one spin-down electron, so there are 2×10^{20} effective levels.) The spacing between levels is so fine (10^{-20} eV) that the separate curves are indistinguishable and one refers to the group as an allowed band (see Fig. D-4). The width of this band does not grow with the number of electrons in the crystal; the number merely determines the closeness of individual curves within a band. The energy region between two neighboring allowed bands is referred to as a forbidden band.

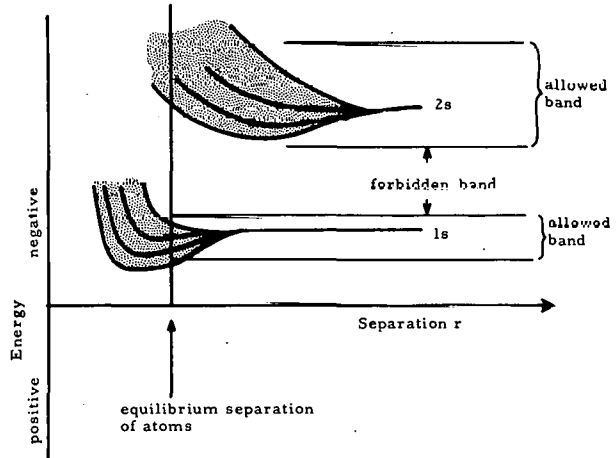


Fig. D-4. Energy levels as a function of inter-nuclear separation for a crystal containing a large number of atoms. An allowed energy band, consisting of innumerable curves, is associated with each state.

It is evident from Fig. D-4 that the average energy of the electrons in the various allowed bands is negative, indicating that the crystal is bound. This effect is known as metallic binding and accounts for the attraction of noncharged atoms.

Consider a crystal such as sodium that has a single valence electron. An electron on one atom can move to a neighboring atom in which the corresponding electron on that atom has the same quantum numbers, if it flips its spin so that the two spins are not the same. Thus, an electron can move about from atom to atom until it encounters an exclusion situation, i. e., one in which the spin quantum numbers are identical. Therefore, the outer electrons are bound to the crystal as a whole but not to a particular atom. This ability of an electron to drift throughout the crystal is what makes the material a conductor.

On the other hand, a material like sulfur that has its outer shell filled cannot become a conductor. An electron cannot move to a neighboring atom without getting into an exclusion situation since there are already two electrons there, one of each spin orientation. Therefore, such materials are nonconductors.

There is another class of materials known as semiconductors, in which some electrons can migrate but others cannot. The four important semiconductors (all of which have a diamond-like crystal structure) are silicon, diamond, germanium, and gray tin. They should all be nonconductors, since their energy bands are full. However, if a high enough potential is applied to a semiconductor, it begins to conduct. This occurs when an electron gains sufficient energy from the ex-

ternal electrical field, between collisions with the crystal lattice, to pass into a higher-energy band. The voltage required depends on the separation of the bands.

Figure D-5 shows the energy bands for silicon. The 3s band is completely filled, and the 3p is partially filled. Silicon becomes conducting when a 3s electron gains 1.9 ev, enough to allow it to jump into the 3p band. For germanium, only ~ 0.6 ev is required for this jump, so that it conducts more easily.

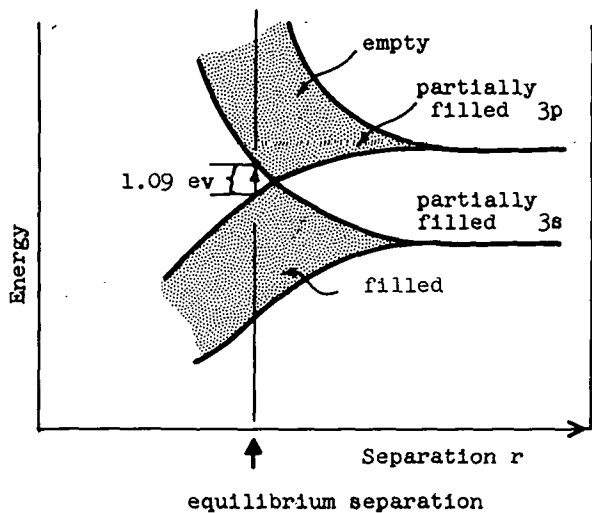


Fig. D-5. Energy bands in silicon as a function of internuclear separation.

Impurities are added to semiconducting crystals to change their characteristics. For example, a material like phosphorous that has five valence electrons may be added in small amounts to germanium, which has four valence electrons (see Fig. D-6). The outermost electron from the phosphorous is so far from its parent atom that it can move about with considerable ease. A crystal of this sort is called "N type," since it has an excess of negative charges.

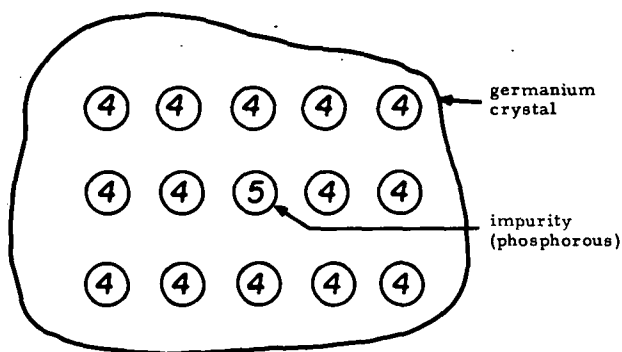


Fig. D-6. N-type crystal, in which an impurity having an excess of valence electrons is added.

On the other hand, the impurity may have only three electrons in its outer shell. This variety of crystal is called a P type, since it has a deficiency of negative charges. Instead of the extra electron being free to move about, we speak of a "hole" as moving through the crystal. Conductivity by means of a hole is not quite as effective as by an excess electron.

In the manufacture of transistors, the crystals are grown artificially. What one does is to allow the crystal to grow for a while, adding an impurity like phosphorous with a valence of five, then shifting to an impurity like arsenic with a valence of three. The resulting single crystal, part with an electron excess and part with an electron deficit is called a PN junction diode (Fig. D-7).

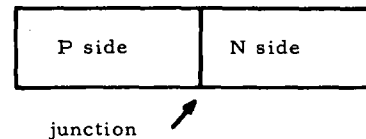


Fig. D-7. PN grown-junction diode.

The energy-band diagram of a PN junction diode is represented in Fig. D-8. In order for the diode to conduct, an electron must jump from a band that is full to a higher-energy or conduction band, which is not full. Thermal vibration of the crystal provides a certain average energy (called the Fermi energy) to an electron, but this is not sufficient to allow the electron to make the jump. However, if one applies a voltage across the diode in such a way that the N-side is negative and the P-side positive, the applied voltage adds to the Fermi energy and the electrons are caused to flow across the junction of the diode.

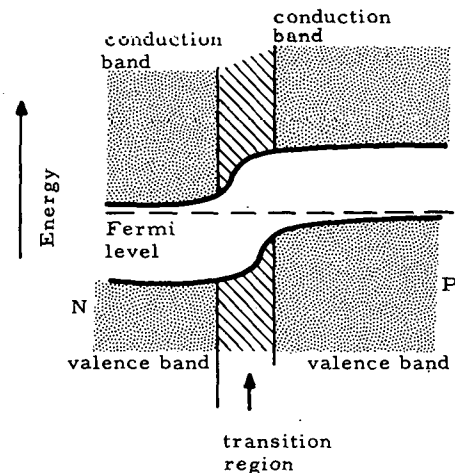


Fig. D-8. Energy-band diagram of a PN junction diode in equilibrium.

It should be emphasized that the applied voltage does not push current through the diode. Rather it acts as a valve, allowing electrons to jump to a higher band by virtue of their Fermi energy. The thickness of the transition layer between P and N type is important to good operation of the diode as well as several other characteristics.

If the voltage is applied in the reverse direction, the Fermi energy is opposed and the electrons cannot make the jump. Thus, no current flows in the reverse direction through a PN diode. This property of diodes makes them useful as rectifiers.

It is also possible to produce triodes, for example an NPN junction triode (Fig. D-9). In this type of triode one can control the flow of electrons by varying the potential across the P-section, which

is called the base. Thus, the base in an NPN junction triode is analogous to the grid in a vacuum-tube triode. By controlling the base current, a transistor may be used as a current amplifier.

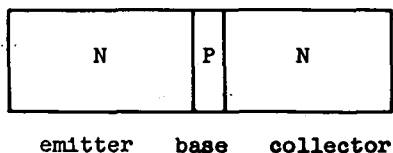


Fig. D-9. NPN grown-junction triode.

Bibliography

1. A. J. Decker, Solid State Physics (Prentice-Hall, New York, 1957).
2. Charles Kittel, Introduction to Solid State Physics (J. Wiley, New York, 1956).
3. F. H. Rockett, The Transistor, Scientific American, 52ff. (Sept. 1948).
4. M. J. Sinnott, The Solid State for Engineers (J. Wiley, New York, 1958).
5. M. Sparks, The Junction Transistor, Scientific American, 28ff. (July 1952).
6. R. L. Sproull, Modern Physics; A Textbook for Engineers (J. Wiley, New York, 1956), Ch. VIII and IX.

E. ELEMENTARY PARTICLES

Robert L. Thornton

What are the elementary particles? Physicists today know of some thirty particles and antiparticles of which atoms are made. Some of these are probably more elemental than others in terms of explaining physical phenomena. Unfortunately, we do not know which ones to choose for this distinction. We must be content to list them all and see what properties they have.

An analogous situation existed during the last century in regard to the chemical elements when the Russian scientist Mendeleef published his periodic table. He systematized the elements, arranging them in rows and columns that indicate similar chemical properties. Although Mendeleef did not understand the underlying principles, he recognized the usefulness of emphasizing the regularities that were observed. With the discovery of the nuclear atom and quantum mechanics in this century, one can now interpret and explain the periodic table.

At present we have succeeded in systematizing the particles but have no general theory which adequately explains them. In this lecture we shall discuss principally the systematics of the particles.

Much is known about the more common particles -- the proton, electron, and neutron -- and how they interact. From these three particles the atoms of the periodic table are made. Why then was it necessary to seek beyond this simple array?

The reason is that in many experiments there are facts which cannot be explained in terms of these three alone. One example of this is in beta decay -- the emission of electrons from radioactive nuclei, as in the case of the decay of a neutron into a proton and an electron. In this reaction momentum and energy cannot be conserved unless one hypothesizes the emission of an additional particle, the neutrino. The neutrino is a neutral particle with no mass. A particle that had to be postulated in order to explain, for example, the photoelectric effect is the photon. This particle, which has no mass and always travels at the velocity of light, is the carrier of electromagnetic energy. The meson is another type of particle which was hypothesized in order to explain an observed phenomenon, in this case the very strong forces that hold the proton and neutron together within the nucleus.

Thirty particles or antiparticles have been observed or postulated. Some of these, such as the hyperons, were discovered more or less accidentally. Others were first predicted to satisfy an otherwise incomplete theory and then found after a deliberate investigation. A few antiparticles that have been postulated have not as yet been observed.

One of the most useful ways to classify these particles is according to how strongly they interact, i.e., according to the intensity of the force between two particles that exists when they are in intimate contact. Three types of interactions are well recognized. These are the strong, weak, and electromagnetic. The strong interactions are those associated with the nuclear forces. The weak interactions are involved in the spontaneous decay of electron-emitting

radioactive materials or unstable particles. The electromagnetic interactions occur between charged particles by virtue of their electric charges or magnetic moments. These interactions differ enormously in strength. If we arbitrarily place the strong interactions on a scale of one, the electromagnetic interactions then have a value of $1/137$, and the weak interactions a value of only 10^{-14} .

The strong interactions occur in an extremely short time. For example, consider a proton approaching a stationary nucleus with a velocity nearly equal that of light. Since a nuclear diameter is 10^{-12} cm, the proton passes by the nucleus in 10^{-22} sec. If otherwise possible, a strong interaction may well result in a nuclear reaction in this brief period. The electromagnetic interactions are an order of magnitude slower, i.e., they occur in about 10^{-21} sec. The weak interactions, on the other hand, are inordinately slow by comparison, taking about 10^{-8} sec at the fastest.

Table E-I shows how particles are classified on the basis of strong or weak interactions.

Table E-I

Classification of particles as to how strongly they interact			
	Interaction type		Spin (angular momentum) ^a
	Strong	Weak	($\hbar$)
fermions	baryons	leptons	1/2
bosons	mesons	photons	0 (or 1)

^aThis is not to be confused with isotopic spin, which is discussed below. Angular momentum is measured in terms of $\hbar$, which is Planck's constant h divided by 2π .

All the leptons that have so far been discovered are light particles -- the electron, the neutrino, the muon, as well as their respective antiparticles (see Table E-II). However, the basis for calling a particle a lepton is not its mass, but rather the fact that it undergoes a weak interaction. Baryons are heavy particles that interact strongly. These include the nucleons (the proton and neutron, and their antiparticles) and the hyperons (lambda, sigma, and xi particles, and their antiparticles). The mesons are strongly interacting particles which are heavier than leptons and lighter than baryons. The photon, although it has no mass, is not considered a lepton. It is in a class by itself and only undergoes electromagnetic interactions.

The terms "fermion" and "boson" have to do with properties such as the angular momentum (spin) of a particle. Fermions all have half-integral angular momenta; in fact, all those presently known have a spin of $1/2\hbar$. Bosons have a spin of zero, except

the photon, whose spin is $\hbar$. Another property of fermions is that they obey the Pauli exclusion principle, whereas bosons do not. It is not necessary to go into this further for our present purpose.

Table E-II
Classification of particles

	Particle	Antiparticle ^a	Mass ^b (Mev)	Mean life (sec)
Leptons	neutrino ν	antineutrino $\bar{\nu}$	0	stable
	electron e^-	positron e^+	0.51	stable
	muon ^c μ^-	antimuon μ^+	106	2.2×10^{-6}
Mesons	pion π^+	π^-	140	2.6×10^{-8}
	π^0	(its own antiparticle) π^0	135	10^{-16} to 10^{-15}
	K meson K^+	K^-	494	1.2×10^{-8}
	K_1^0	$\bar{K}_1^0$	494	10^{-10}
	K_2^0	$\bar{K}_2^0$	494	3×10^{-8} to 10^{-6}
Baryons	proton p	antiproton $\bar{p}$	938	stable
	neutron n	antineutron $\bar{n}$	940	stable
	lambda Λ^0	$\bar{\Lambda}^0$	1115	2.7×10^{-10}
	sigma Σ^0	$\bar{\Sigma}^0$	1190	
	Σ^+	$\bar{\Sigma}^-$	1184	0.7×10^{-10}
Hyperons	Σ^-	$\bar{\Sigma}^+$	1197	1.5×10^{-10}
	xi (cascade) Ξ^0	$\bar{\Xi}^0$	1320	
	hyperons Ξ^-	$\bar{\Xi}^+$	1320	10^{-10} to 10^{-9}

a For every particle of ordinary matter there is a corresponding antiparticle, with an opposite electric charge (in the case of charged particles) and an opposite magnetic moment, but with the same mass and half life. A few of the listed antiparticles have not yet been observed.

b The difference in mass between a charged and a neutral particle of a given type presumably arises from the electromagnetic interaction.

c Formerly, the term "mu meson" was used for this particle. The muon is now recognized as a lepton because it does not interact strongly. "Meson" is reserved for strongly interacting particles.

Before proceeding further with a discussion of the systematics of elementary particles, we must review the various conservation laws of physics (see Table E-III). The first five listed in Table E-III are always conserved in any nuclear interaction. For instance, when two particles collide and form several new particles, the mass of the system remains unchanged, providing one considers the mass that is converted into energy. The last three listed in Table E-III are conserved in strong interactions, but not always in weak ones.

Table E-III

Conservation laws	
1. energy and mass	} always conserved
2. electric charge	
3. momentum (angular and linear)	
4. baryon number	
5. lepton number	} usually conserved, and always in strong interactions
6. parity	
7. strangeness	
8. isotopic spin	

The first three conservation laws have been explained in Section C, "Nuclear Physics." The others are perhaps unfamiliar. "Baryon number"

and "lepton number" do not refer to the number of baryons or leptons present; rather, they are quantum numbers. The lepton number is always +1 for leptons, -1 for antileptons, and 0 for particles that are not leptons. An example of the lepton number being conserved is in a reaction where a gamma ray produces an electron-positron pair:

$$\gamma \longrightarrow e^- + e^+$$

The lepton number for the gamma ray is zero, since it is not a lepton. It is +1 for the electron, and -1 for the positron, because the positron is an antiparticle. The value on the left side of the equation cancels the sum on the right.

In a similar way, the baryon number is +1 for a baryon, -1 for an antibaryon, and 0 for nonbaryons. Since the baryon number must be preserved in any interaction, the proton cannot decay. This is one way of saying that our universe exists. For instance, if a proton were to decay into π mesons, the baryon number would be +1 on the left side of the equation and 0 on the right. Therefore, this reaction cannot occur.

There is no "meson number" that has to be conserved. This implies, as is true, that a single meson can be produced in an interaction, whereas leptons and baryons occur in pairs.

"Parity" refers to the fact that "right" and "left" are not fundamental properties of nature. [For a discussion of this conservation law and an account of its overthrow a few years ago, see Philip Morrison, "The Overthrow of Parity," Scientific American, 45-53 (April 1957).]

"Strangeness" is also a quantum number. This concept is discussed later. "Isotopic spin", or I-spin as it is often called, must be conserved in a manner analogous to that of the lepton or baryon number. The concept of isotopic spin originated in connection with neutrons and protons. Except for the fact that the proton carries a charge and the neutron does not, these particles are very similar. They have almost the same mass and interact with about the same strength. The question naturally arises as to whether or not they are merely different aspects of the same particle, which we now call the nucleon.

It must be emphasized that isotopic spin is totally unrelated to electron spin, which is a true angular momentum and is denoted in units of $\hbar$. However, it is instructive to consider electron spin because there are certain analogies that can be made. In a magnetic field, electrons can spin with one of two possible orientations, either parallel or antiparallel to the field. If parallel, we say that the electron spin is +1/2. If antiparallel, we say the spin is -1/2. The direction of the magnetic field is often designated the z axis of a coordinate system.

In an analogous way, we think of an isotopic spin (I-spin) associated with nucleons as follows: This spin has two components in a preferred direction in space (not ordinary space!) called the z direction. Nucleons are assigned an I-spin of 1/2. The z component (I_z) for a proton is said to be +1/2, and for a neutron -1/2.

It is then possible to write an expression for particles that relates their charge Q, in units of the

electronic charge; z component of isotopic spin I_z ; and baryon number N . We have

$$Q = I_z + N/2. \quad (1)$$

Consider a proton, which has the values $I_z = +1/2$ and $N = +1$. We write

$$Q = 1/2 + 1/2$$

$$Q = +1,$$

which we know is the charge of the proton. For the neutron, we have $I_z = -1/2$ and $N = +1$:

$$Q = -1/2 + 1/2$$

$$Q = 0,$$

indicating that the neutron is a neutral particle. This type of reasoning, which of course is a formalism, is nevertheless useful for such things as predicting particles presently unknown.

We use the term "singlet" to describe a type of particle whose I -spin can have only one possible orientation. An example is the lambda (Λ^0). (See Table E-IV.) Similarly, the terms "doublet" and "triplet" describe particles capable of two and three possible orientations, respectively. For a singlet, the I -spin is taken as zero, in which case I_z must also be zero. For a doublet, the I -spin is $1/2$; I_z may then be $+1/2$ or $-1/2$. For a triplet, the I -spin is 1; I_z may be $+1$, 0, or -1 .

Table E-IV

Strangeness number and isotopic spin for baryons, antibaryons, and mesons			
	Baryon no. N	Strangeness no. S	Isotopic spin
			Singlet $I = 0$ $I_z = 0$
Baryons	1	-2	$\Xi^0 \Xi^-$
	1	-1	Λ^0
	1	0	$p \quad n$
Anti-baryons	-1	0	$\bar{p} \quad \bar{n}$
	-1	1	$\bar{\Lambda}^0$ (not observed)
	-1	2	$\bar{\Xi}^+ \bar{\Xi}^0$
Mesons	0	-1	$\bar{K}_0 K^-$
	0	0	π^0 (not observed)
	0	+1	$K^+ K^0$

The number of particles of a given type which can exist is the same as the number of possible orientations of the I -spin. Thus, a doublet like the nucleon can exist in two states; hence the proton and neutron. A singlet such as the lambda can have only one state, and that is neutral.

These systematics were generalized to include the strange particles (K mesons and hyperons) by Pais and Gell-Mann. This involved introducing a new quantum number, the "strangeness" number S (see Table E-IV). The strangeness number is related to isotopic spin I and charge Q by the expression

$$S = 2(Q - I) - \text{baryon number}.$$

As an example, consider a Σ^- particle. From Table E-IV we see that its isotopic spin is -1 . Its charge is -1 and its baryon number is 1. Therefore, we have

$$S = 2[-1 - (-1)] - 1$$

$$S = -1.$$

In order to accommodate the strange particles, Eq. (1) for electric charge must be generalized as follows:

$$Q = I_z + N/2 + S/2. \quad (2)$$

Where S is zero, as in the example given with Eq. (1), Eq. (2) reduces to Eq. (1). Consider a particle where S is not zero. An example is the xi zero (Ξ^0), the neutral cascade hyperon discovered recently at this Laboratory. For the Ξ^0 we have $I_z = +1/2$, $N=1$, and $S = -2$ (from Table E-IV). Therefore, we write

$$Q = 1/2 + 1/2 + (-2)/2$$

$$Q = 0.$$

It should be noted that certain blocks are excluded in Table E-IV, i.e., they contain no particles. The reason for this is that for a particle to fit into one of the excluded blocks, it would have to have one-half an electronic charge. Since this is impossible, we believe that this table is complete and that no other particles can exist. Some particles are listed which have not yet been found; the singlet π^0 meson and certain antiparticles. Experiments are now underway which seek to discover these particles. It is the general belief that they can exist and will be found, with the possible exception of the singlet π^0 .

Table E-IV shows very simply and concisely the relationship between various properties of the particles. Why the particles can be arranged in this way is not known. Perhaps some day we will be able to explain it, just as we can now explain the periodic table of the chemical elements.

Bibliography

1. J. S. Blair and G. F. Chew, Subnuclear Particles, in *Annual Review of Nuclear Science*, Vol. 2 (Annual Reviews, Stanford, California, 1953), pp. 163-186.
2. S. De Benedetti, Mesonic Atoms, *Scientific American*, 93 ff. (Oct. 1956).
3. W. F. Fry, Hyperfragments, in *Annual Review of Nuclear Science*, E. Segrè et al., Eds., Vol. 8 (Annual Reviews, Stanford, California, 1958), pp. 127-162.
4. M. Gell-Mann and E. P. Rosenbaum, Elementary Particles, *Scientific American*, 72-88 (July 1957).
5. M. Gell-Mann and A. H. Rosenfeld, Hyperons and Heavy Mesons, in *Annual Review of Nuclear Science*, J. G. Beckerley et al., Eds., Vol. 7 (Annual Reviews, Stanford, California, 1957), pp. 407-478.
6. S. Glasstone, *Sourcebook on Atomic Energy*, Second Edition (Van Nostrand, Princeton, 1958), Ch. II and XVIII.

7. R. E. Marshak, Meson Physics, in Annual Review of Nuclear Science, Vol. 1 (Annual Reviews, Stanford, California, 1952), pp. 1-42.
8. R. E. Marshak, Pions, Scientific American 84 ff. (Jan. 1957).
9. P. Morrison, The Neutrino, Scientific American, 58 ff. (Jan. 1956).
10. P. Morrison and E. Morrison, The Neutron, Scientific American, 44 ff. (Oct. 1951).
11. P. Morrison, The Overthrow of Parity, Scientific American, 45-53 (Apr. 1957).
12. P. Morrison, A Survey of Nuclear Reactions, in Experimental Nuclear Physics, E. Segrè, Ed., Vol. II (Wiley, New York, 1953), Part VI.
13. J. Rainwater, Mu-Meson Physics, in Annual Review of Nuclear Science, J. G. Beckerely et al., Eds., Vol. 7 (Annual Reviews, Stanford, California, 1957), pp. 1-30.
14. S. B. Treiman, The Weak Interactions, Scientific American, 72-84 (Mar. 1959).

F. ANTIPARTICLES

Owen Chamberlain

The subject of antiparticles must begin with electrons and their antiparticles, positrons. The electron was discovered in 1895, and much physics has gone in to the study of its behavior since that time. Another theme which must be drawn on heavily in this lecture is the theory of relativity, propounded in 1905. Of primary importance here is the assertion that energy and mass are equivalent.

There appeared in 1926 the theory of quantum mechanics, and two years later the theory of Dirac. This last theory is quite complicated. It is well known that the momentum P of a particle can be expressed by the equation

$$P = \frac{M_0 v}{\sqrt{1-v^2/c^2}}, \quad (1)$$

where M_0 is the rest mass of the particle; v is its velocity, and c is the velocity of light. Also, the total energy E of a particle can be expressed by

$$E = \frac{M_0 c^2}{\sqrt{1-v^2/c^2}}. \quad (2)$$

Physicists usually combine these relations into a fairly simple form:

$$E^2 = (Pc)^2 + (M_0 c^2)^2, \quad (3)$$

or

$$E = mc^2, \quad (4)$$

where m is the total mass of the particle. Dirac was trying to follow this reasoning to its natural conclusion. One of the strange properties of Eq. (3) is that when it is solved, the energy E can be either positive or negative:

$$E = \pm \sqrt{(Pc)^2 + (M_0 c^2)^2}. \quad (5)$$

Notice that in Eq. (5) the minimum value for the energy is $M_0 c^2$, which would be for a particle that has zero momentum. As the momentum is increased, the value for energy can go higher but not lower. In terms either negative or positive, the energy can only move above the minimum value. The picture derived from Dirac's theory is consistent with this equation.

Figure F-1 may be used to illustrate Dirac's ideas. At first Dirac did not know what to make of the negative energy state. It does not make sense in our everyday way of thinking to talk about negative energy. That is, what could take place according to this theory? Imagine a single electron and a positive-energy state. This might correspond to a vacuum box that was empty except for one electron. The electron could give off a gamma ray and be reduced to one of these lower energy states. The electron could then give off another gamma ray and drop to a still lower energy state. This presents other difficulties, however. To get around this, Dirac said that these states were the possible states

of an electron, but there is an $E = 0$ line, in Fig. F-1. Above this there are allowed states which are usually empty. Below this line there are allowed states of an electron but they are usually full. It will perhaps be remembered that electrons in atoms are restricted to only one electron to each quantum state. There are only two electrons in the innermost k shell; the reason that there are no more than two is connected with the Pauli exclusion principle that two electrons cannot fit identically in the same state.

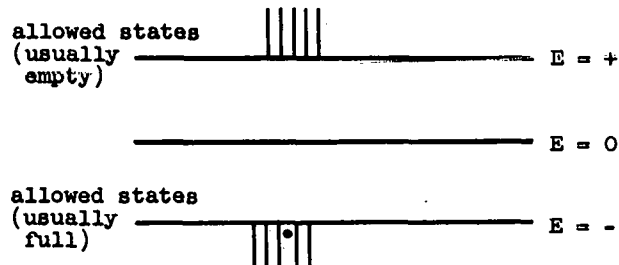


Fig. F-1. Energy diagram suggested by Dirac.

One electron has the spin pointing up and the other the spin pointing down. Thus, there are actually two states and two electrons - there cannot be more. The next shell has room for only eight electrons. The same thing is illustrated in Fig. F-1. Dirac said that there is only one electron to each possible state. Moreover, in his theory this peculiar situation is the way a vacuum is represented. In other words, the vacuum box in this theory is an infinite sea of filled electron states of negative energy.

Curiously enough, if all these electron states are completely filled, none is missing. It is true then that most of the things one would seek experimentally must look just like a vacuum. For instance, there is no sign of an electric current flowing anywhere, because for every state of electrons moving in one way there is also a state of electrons moving in the opposite direction at the same speed. They must cancel each other perfectly. The same is true for all other possible directions. Actually, this is a peculiar way to represent a vacuum, but this is the best that Dirac could do by way of illustrating his argument.

Then Dirac asked another question. What would be the behavior of one electron found in one of the positive energy states? His answer was that it would correspond to the typical behavior of an ordinary electron. What then, he asked, happens when one electron is missing from this infinite sea? The situation might be viewed in this way. One starts with a filled band but removes just one electron. The dot in Fig. F-1 indicates that one electron is missing. Dirac found that in this situation, for all experimental purposes, the empty space would look like a positive charge. This makes sense because from a complete closed system one negative charge has been removed and the result is one positive charge. The hole acts like a positive charge and it moves like a positively charged part of what was removed. Dirac studied this in some detail and what he found

was that this particle would move much as though it were the opposite charge of the electron. So the logical results was to think of an electron whose charge is negative, and a hole which behaves just like a particle with a positive charge. He examined the angular momentum or spin that such a particle would have. Many of these particles are like tiny tops constantly spinning. They have an angular momentum built into their nature as particles. The spin of the electron was known to be $1/2$, where $\hbar$ or Planck's constant divided by 2π is the unit. He also found that the spin for this hole would be $1/2$ unit. Similarly, he identified the mass of this hole. He was a little in doubt about the mass, as there were various possibilities; however, he found that if we call the mass of the electron m_e , it could be strongly suspected that the mass of this hole in his infinite sea would have the same mass as the electron.

In 1932 this particle was discovered by C. D. Anderson of the California Institute of Technology. He called it the positron. Figure F-2 is a diagrammatic representation of Anderson's original picture, which was a cloud chamber photo of the first identified antiparticle. A lead plate was mounted through the center of the chamber. Because of the plate he could be quite certain of the direction in which the particle was moving. It had lost momentum by its encounter with the lead plate. The cloud chamber was in a magnetic field. Therefore, the curvature of the track in the cloud chamber indicates the curvature of the particle that made the track, and this may be used to give a direct measure of the particle's momentum. It is evident that the curvature was greater at the lower side of the plate.

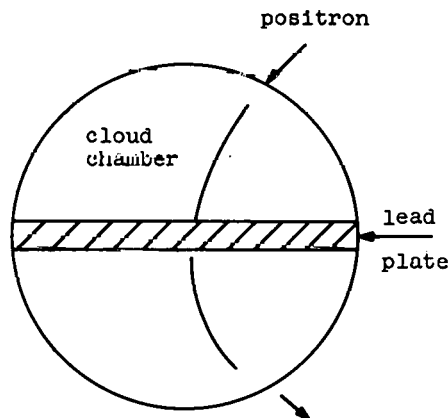


Fig. F-2. Drawing of the cloud chamber photo obtained by Anderson. Note the greater curvature on the lower side of the lead plate.

The direction of travel for the particle was very important because this would determine whether it was a positive or negative electron. The presence of the plate allowed him to tell the direction the particle followed. He could then say that the curvature was such that the particle was positively charged. If it were an ordinary negative electron it would curve in the opposite direction. This was proof that the positively charged particle existed.

Figure F-3 is another drawing of a cloud chamber picture, again with a lead plate. Occasionally an electron pair is seen originating in the plate. The large spiral track shows where an electron has made many turns in the magnetic field of the cloud chamber. Unseen gamma rays have come in from

the top and made a pair at the point of origin. One of the pair was of rather low energy and circled for a time in the chamber. In many cases where such a pair is created the particle does not remain very long in the cloud chamber. Usually it spirals up or down and quickly leaves the chamber.

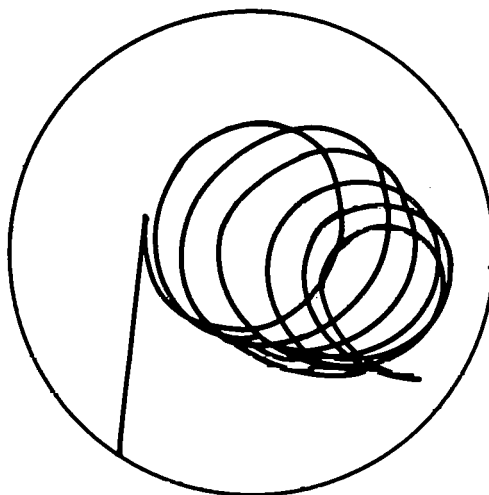


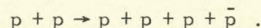
Fig. F-3. Drawing of a cloud chamber photograph.

Pair production involving gamma rays is a familiar subject. A gamma ray may change into an ordinary electron and a positron. This process does not occur in vacuum because if one applies simultaneously the conservation of energy and the conservation of momentum, it turns out that a gamma ray cannot turn into an electron and a positron without the presence of matter. A nearby nucleus which has a large electric charge can absorb a little bit of the momentum of the gamma. Lead plates are ideal because of the large charge of lead nuclei.

Considering the success of Anderson and others in identifying the positron, it was only natural to look for other particles. What might be expected of them? The proton is well known. Its charge is positive, its spin is also $1/2$ unit, and its mass is known. The question was asked, "Can one say the same things about an oppositely charged particles which might be called antiprotons?" If there is to be an antiproton, what is believed to be true about it from the Dirac theory indicates that it must have a charge opposite that of the proton. If the proton is plus, its antiparticle must be minus. It should have the same spin and presumably the same mass. There are a number of other properties to be predicted by this theory, such as the magnetic moment, but they can perhaps be omitted for the present.

The main problem not many years ago was to decide under what circumstances one might be able to make an antiproton. It is known, for example, that a proton is a stable particle. By this is meant that it does not suffer a radioactive decay and change into something else spontaneously. According to Dirac's ideas, if the proton is stable the antiproton should also be stable. This means that an antiproton in a vacuum should last forever. The fact that the proton is completely stable has made us suspect for a long time that there is a very rigid rule of physics that the only way to destroy a proton is to arrange an annihilation process with an antiproton. The reason the proton is stable is the fact that there are no antiprotons around to destroy it.

This rule that protons are never lost suggested very strongly that if an antiproton were ever made, it would have to be made as a member of a pair with another proton. Thus, the following experiment was suggested: Cause two protons to dash into each other at high energy. In the process they might form not only the original two protons but an extra pair, a proton and antiproton. This reaction may be written



(The antiproton is indicated by $\bar{p}$, as is customary.) A pair would be created if the reaction can be made to proceed. How much energy must be put into the colliding protons in order to make these extra particles? This refers back to Einstein's relation between mass and energy, as noted above. If we are to create in the final products some extra mass then we must supply the equivalent amount of energy. This can be computed simply from $E = mc^2$.

The mass of a proton times c^2 , in energy equivalent units, is 938 million electron volts (Mev). Compare this with the electron. The mass of an electron times c^2 is just about a half million volts (0.5 Mev). This is quite a difference. The proton is roughly 2,000 times heavier than the electron and must have 2,000 times as much rest energy. The energy that must be supplied in order to make this creation appear is just twice this amount, because it is assumed that the antiproton and the proton have the same mass. More correctly, it should be asserted that $2M_p c^2$ would be 1878 Mev, where M_p is the mass of a proton. If it were possible to cause one proton to collide with another while they were moving toward each other, then this would be the minimum amount of energy required. One would put 938 Mev into one proton and the same on the other; when they collide there would be just enough energy available to make a pair possible.

Up to the present time we are restricted to having but one proton in motion and causing it to collide with a proton that is virtually stationary. The only part that counts is the effective energy of these two colliding. The actual minimum energy that must be supplied, called the threshold energy, turns out to be six times the mass of the proton times c^2 , which is about 5,600 Mev. When Professors Lawrence and McMillan were making plans for the Bevatron, they had this figure in mind so that antiproton production might be possible. Indeed the machine was planned to deliver a fraction over 6,000 Mev. It was worthwhile therefore, to seek this reaction.

Dirac had predicted the existence of this antiproton many years ago. However, when Professor Anderson scanned his cloud chamber pictures showing cosmic ray tracks he looked very carefully for any evidence of the antiproton as well as the positron. However, he did not observe the antiproton. There was a growing suspicion that the antiproton did not exist. When the Bevatron became a usable tool for research the opportunity was first presented for finding it by machine methods. The apparatus shown in Fig. F-4 was set up to detect any possible antiprotons resulting from the reaction already noted. At the upper left of Fig. F-4 we see the circulating proton beam in the Bevatron. This beam strikes a target T consisting of stationary protons. Any antiprotons that might be formed would go predominantly in the forward direction. Those particles having just the right momentum would follow the line going through D and Q1. These

particles would be focused into a beam by magnet Q1, which serves as a magnetic lens. This beam is then passed through a bending magnet M1, is again focused, is passed through a second bending magnet M2, and is again focused (by Q2). In our first experiment the apparatus was limited to what is shown in Fig. F-4. Later, additional features were added; Figure F-5 shows the apparatus schematically. Again at the top of the picture is

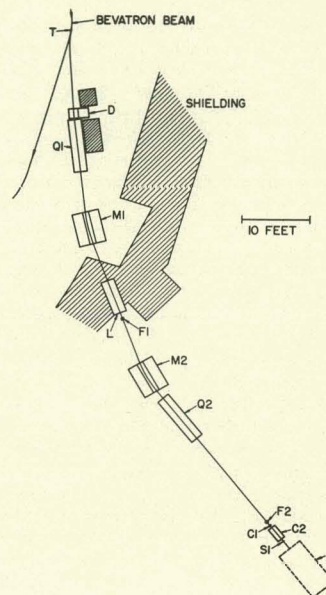


Fig. F-4. The antiproton experiment (preliminary arrangement).

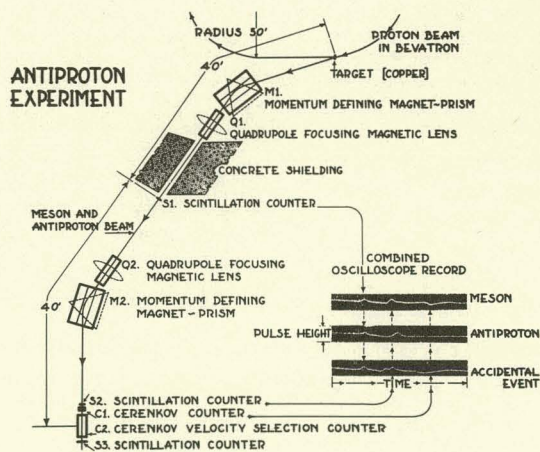


Fig. F-5. The antiproton experiment (final arrangement).

the Bevatron beam. A copper target was the source of target protons used in antiproton production. The first bending magnet is indicated by a prism because of the very close analogy between the way a magnet bends a charged particle and the way a prism will bend the path of light. Think of light of different colors being bent by the prism. In the same way particles of different momentum are bent differently by the analyzing magnet. If you wish, you might think of this as a very simple spectroscope. The essential minimum is one prism and one lens. Their function is to separate out a sort of image of

the target. If the color of the light is correct you will see an image at the point S-1 in Fig. F-5, where a focus is indicated. The diagram shows a second spectroscope working in the same way. S and S₂ are actually scintillation counters. These were placed at the two focus points. Counts being recorded at the time indicated that one particle had gone through one counter, through this magnetic system, and on to the other counter.

The problem was complicated by the fact that many particles might have the right momentum to move through this system successfully. Many other particles were produced from the same target. In fact there were many π^- mesons, which are likewise negatively charged. We had hoped that there might be some antiprotons and also expected that there would be quite a number of electrons coming from the copper target. It was necessary to inquire into the velocity of these particles as they came through the magnetic system. The magnet selected one value of momentum. Knowing the momentum of these particles, we could calculate their velocity for any assumed mass. It was determined that the pi mesons would be going 99% of the velocity of light. Antiprotons are heavier and need not move as fast to have the same momentum. They would be going 76% of the velocity of light, while the electrons with slight mass would be moving essentially at the velocity of light. On this basis we hoped to separate antiprotons. Some effort went into measuring this velocity precisely. This was done by measuring the time it took the particle to go from counter S1 to S2. Using refined techniques, we were able to measure this time quite accurately. Furthermore, we used counters which specifically measure particle velocity. These are Cherenkov counters (see Section K: "Scintillation and Cherenkov Counters"). They proved especially valuable in this experiment. At the lower right of Fig. F-5 is seen the oscilloscope traces of the electrical signals from the various counters. Particles in rapid flight show pips or deflections close together. The slower moving antiprotons give pips that are more widely separated.

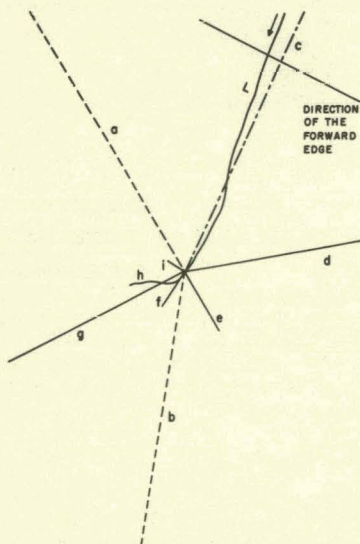


Fig. F-6. An antiproton stopping in emulsion.

Evidence of antiproton production was gathered by this sort of apparatus. Also immediate attempts were made by other procedures to collect further confirming evidence. One procedure was to stop these particles in nuclear emulsions. Figure F-6

is one of the first records made of a stopping antiproton. The track labeled L comes in from the top as a heavy ionization and represents an antiproton coming to rest. When at rest, it combined with a proton from a nucleus in the nuclear emulsion resulting in their annihilation. This might be described as an antiproton and proton essentially eating each other up and regurgitating five mesons or so.

In a short time another picture of a similar sort was found which was very important to us (Fig. F-7). This was the first picture to show that the energy of this annihilation was so great that it could be nothing but the $\bar{p}$ -p annihilation. Let us see how this could be proved to be an annihilation process. The mass of that particle coming in could be measured and compared with the known mass of the proton. The measuring of the mass, in effect, came from knowing the particle momentum, by careful control of the magnetic field, and measuring its range in the emulsion stack. The resulting figure definitely agreed with the mass of the proton; also there was justifiable reason for saying that it brought with it energy of 938 Mev. In the annihilation process, however, there are two things being destroyed: the incident antiproton and a proton. The process should contain twice the energy that the antiproton brought in, because two proton masses--one proton and one

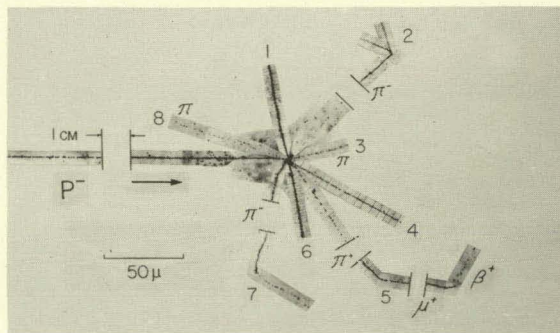


Fig. F-7. A proton-antiproton annihilation.

antiproton--are involved. It is in this sense that the energy release was definitely close to the amount expected, proving that it could not have been merely a particle decaying by itself.

Logically, the next question to ask was: "Now that the antiparticle of the proton has been found, what others might exist?" It is not reasonable to expect that the neutron would also have its antiparticle? The antineutron was also found rather recently. The neutron has no charge, and likewise the antineutron has no charge, but each has a spin. Therefore, the only difference that can exist between the neutron and antineutron is the direction of spin. This antineutron was found first by Wenzel, Piccioni, and Lambertson at this Laboratory.

Figure F-8 shows the antineutron process diagrammatically. The original photograph was taken with the 30-in. propane bubble chamber. The particle entering from the right can be identified as an antiproton, but one which suddenly disappears. What happened is that it suddenly traded one meson with a particle in the chamber and became an antineutron. This antineutron coasted along for a short distance, leaving no track in the chamber (as it was without charge). Then it ends in an annihilation star. Just as the antiproton can be eaten up in an annihilation process, so also can the antineutron.

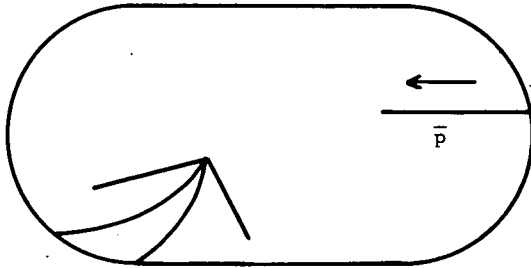


Fig. F-8. Diagram of a photograph taken with the 30-in. propane bubble chamber.

Because of the presence of ordinary matter, antimatter cannot live for long as such. If there is any place in our universe where antimatter exists independently and in isolation, we do not know about it. There is no way known to tell whether the distant

stars and galaxies are composed of antimatter or ordinary matter. As far as is known, matter or antimatter would look exactly alike. There is a faint hope that someday a process can be seen in which ordinary matter and antimatter might collide out in space and result in great showers of pi mesons. So far, nothing like this has been identified with certainty.

Bibliography

1. E. Segrè, Antinucleons, in Annual Review of Nuclear Science, E. Segrè et al., Eds., Vol. 8 (Annual Reviews, Stanford, California, 1958), pp. 127-162.
2. E. Segrè and C. Wiegand, The Antiproton, Scientific American, 37 ff. (June 1956).

G. DESIGN OF EXPERIMENTS

Robert W. Kenney

Proficiency in designing experiments comes only with experience, just as skill in playing the piano is attained only through practice. Each physics experiment presents its own unique problems of design. No simple set of rules exists for solving all of these. However, there are some general principles that, if followed, may save the experimenter time and effort and enable him to obtain more meaningful results.

To investigate a phenomenon intelligently requires a thorough understanding of what is already known about it. No one wants to repeat work that someone else has done, but more than this, it is essential to comprehend the theoretical implications of existing knowledge. Also, one should be familiar with the experimental background. Knowing what has failed and what has succeeded may allow the experimenter to avoid the same mistakes and capitalize upon the good points.

After one has a complete grasp of the relevant physical principles and of the pertinent techniques and materials, he can proceed to design the "nuts and bolts," i. e., the apparatus that will be used. In their younger and more foolish days, some people go about this in the reverse order. First they design the nuts and bolts, then they decide what it is they are hoping to accomplish, and finally they examine the existing theory.

In general, we can divide experiments into two classes: those that explore a new field, and those that gather precise data. In approaching a new area of investigation, one usually does an exploratory experiment first. This is to give a broad view of the situation and reveal what phenomena exist. The apparatus used in this type of experiment must not restrict one's over-all view.

In performing exploratory experiments one should notice and remember rare events, because new effects sometimes appear in this way. For example, two Australians, Rochester and Butler, were scanning a large number of cloud chamber photographs of cosmic rays in 1947. They noticed two tracks that were unlike anything they had previously seen. One track resembled a hook and one a fork. Because they were careful experimenters, Rochester and Butler tried to understand everything they saw. It would have been easy to ignore these unusual tracks, ascribing them to some malfunction of the equipment. Instead they treated these tracks as being due to unknown particles. As it turned out, these were tracks of previously unknown "elementary" particles. Today the major effort of the physics groups at this Laboratory is involved in the study of these particles.

Once the over-all field has been assayed, one then designs an experiment to give precise answers to specific questions. Previously, resolution was sacrificed in order to take in the largest possible area of information. But now one is interested in detail. Therefore, the broad view must be sacrificed in favor of resolution. For instance, a beam of nuclear particles coming out of a cyclotron invariably has some energy spread. In an exploratory experiment the physicist might be interested in looking at

all these energies; the beam intensity would therefore be great. However, in a precision-type experiment, he would exclude all particles except those of a very narrow energy interval. That is, he would sacrifice beam intensity for resolution. Resolution and intensity are thus complementary; if one is small, the other is large, and conversely.

Suppose that a new type of nuclear particle has been found in an exploratory experiment. The physicist then wants to know its physical properties, i. e., its size, mass, charge, energy, mode of decay, etc. To do this, he must be familiar with the kinds of instruments that are suitable for taking these measurements.

Of fundamental importance in this type of experiment is the question, "Just how precise must the measurement be in order to be meaningful?" Suppose, for example, that you wanted to know the time of day so that you would not miss a train. A calendar would be useless, because its smallest time increment is 24 hours. Likewise it would be uneconomical for this purpose to build and maintain an elaborate electronic device that would measure time intervals to, say, 10^{-9} sec. All that you need is a reasonably accurate watch. Thus, the answer to our question is that a measurement must be accurate enough to give useful results. That is, for a given purpose, how much inaccuracy can be tolerated? In many cases, an answer to this will involve a statistical analysis. The lecture entitled "The Statistical Analysis of Data" discusses this topic in more detail.

It may turn out that it is impossible to measure a given quantity with the required accuracy using known techniques. At present the practical limits of accuracy are indicated in Table G-I.

Table G-I

The best precision now possible in measuring certain quantities

Quantity	Accuracy	Method
mass	1 part in 10^5	calibrated weights
length	1 part in 10^7	interferometer methods
time	1 part in 10^7 to 10^8	electronic methods
voltage	1 part in 10^4	standard cell
magnetic field strength	1 part in 10^7 to 10^8	precession frequency of proton magnetic moment

To measure mass as accurately as 1 part in 10^5 requires using carefully calibrated weights. This might be improved to 1 part in 10^6 , but only with great effort. Length is measured much more easily, by means of an interferometer. This is an optical instrument which measures length or thickness in terms of known wave lengths of light. Time

can be known with an accuracy of 1 part in 10^7 or 10^8 by listening to the time signals broadcast by WWV, the Bureau of Standards radio station. Voltage can be measured to 1 part in 10^4 by means of a standard cell. Currents cannot be determined as accurately. Magnetic fields, on the other hand, can be determined to 1 part of 10^7 or 10^8 . This is done by measuring the precession frequency of a proton in a magnetic field. The magnetic moment is a well known parameter of the proton. When a proton is in a magnetic field it spins with its axis oriented in a certain direction relative to the field. If a radio-frequency is then applied, this axis then precesses, or flips. By measuring the frequency at which this occurs, one has a very precise indication of the magnetic field strength.

Thus, for results that are the most precise, one should design an experiment so as to measure mass, length, time, or magnetic field strength, avoiding measurements of voltages or currents.

A very important class of measurements might be called "differential-type measurements." An example of this is determining the difference in length between two long metal tapes. One might be a tape whose length is well known, say 1/4 mile. The other tape is supposedly 1/4 mile long also, but in fact is slightly less than that. To determine the error in the second tape one does not take a third tape, measure the first two with it, and then compare the difference. Rather, he puts the two tapes side by side with, say, the left-hand ends together. Then he observes the right-hand ends and notes the slight difference in length as indicated by the tape whose length is well known.

Another case in point is in regard to measuring the mass of a given element, say, helium 4. In nuclear physics, the mass of oxygen is taken as 16,000... atomic mass units, in order to have some standard. We would then want to compare the mass of He^4 to the mass of O^{16} . A mass spectrograph would be used for this (Fig. G-1). It utilizes the principle that a

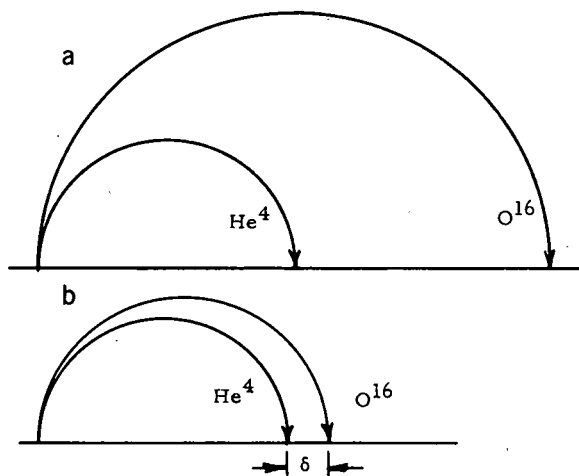


Fig. G-1. Method for determining the mass of an element with a mass spectrograph: (a) incorrect; (b) correct.

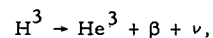
charged particle is deflected in a magnetic field. The radius of curvature ρ is proportional directly to the ratio of mass to charge, and inversely to the magnetic field:

$$\rho \propto \frac{m}{e} \frac{1}{B}$$

Thus, for singly ionized particles, oxygen will have a radius of curvature about four times as great as helium 4 (see Fig. G-1a). To compare the mass of He^4 to that of O^{16} by this method we would have to measure the two radii of curvature extremely accurately. Furthermore, we would have to measure the magnetic field strength quite precisely at each point. The mass of He^4 could be obtained, but not very accurately.

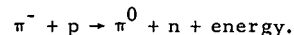
A more precise method would be to use O^{16} that has been ionized four times and He^4 that has been ionized only once. Their radii of curvature in the mass spectrograph would then be almost the same. By measuring the small difference δ (Fig. 1b) with only moderate precision, one could then calculate the mass of He^4 very accurately, to about 1 part in 10^5 or 10^6 .

It is also possible to determine the mass of particles by another type of differential measurement. This method depends upon knowing very accurately the mass of one particle involved in a nuclear reaction. Consider, for example, the reaction



in which tritium decays to helium 3 by emitting a beta particle and a neutrino. The mass of He^3 is known very precisely. The energy given off in the form of the beta and the neutrino can be measured. By adding this to the mass of the helium-3 nucleus, one can then find the mass of the tritium nucleus quite closely, because energy is conserved in the reaction.

Perhaps a better example of a differential measurement is the experiment performed about 10 years ago by Professor Panofsky in which he determined the mass difference between the π^0 and π^- meson. He did this by observing the reaction



The mass of the proton and neutron were known with good precision. Therefore, by measuring the energy released, he could calculate the difference in mass between the π^- and π^0 . The accuracy of his result would depend directly upon how closely he could measure this energy. To improve the precision of this measurement, Panofsky designed an experiment in which this energy would first be amplified about six times. This meant that the result would be about six times as accurate for a given precision of the measurement itself.

Panofsky knew that the energy that is released serves to kick the two reaction products, the π^0 and n , away from each other with equal and opposite momenta. This energy would then reside only in the kinetic energy of motion of the two particles. Panofsky then measured this kinetic energy by observing the velocity of the π^0 . He could have done this by observing the neutron velocity, but this was not nearly so conducive to careful measurement.

The principle upon which he measured the π^0 velocity is explained as follows: The π^0 decays into two γ -rays (photons) almost immediately, before it has a chance to lose any energy. Because the π^0 was moving before it decayed, the laboratory energy of the photons will be shifted from the energies measured as seen from an observer moving with the π^0 meson. This is analogous to the toot of a moving train; as the train approaches us, the toot is higher in pitch than that heard by the engineer and is still higher

than when the train is going away. This Doppler effect also applies to photons. If the photon is moving toward us, the frequency of the light that we observe is increased. But if it is moving away, the frequency is decreased. Panofsky measured the frequency of the light produced in many of these reactions and related it to the photon energy. His results are given in Fig. G-2. He found that the energy ranged between 83 and 55 Mev, a spread of 28 Mev.

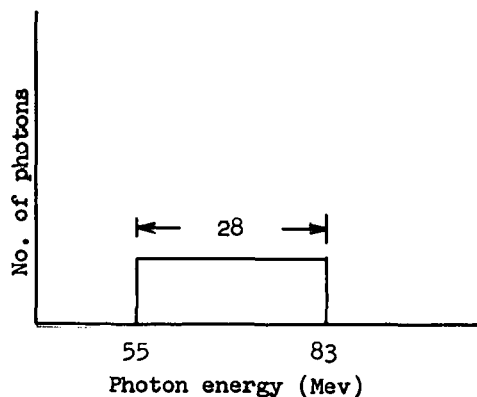


Fig. G-2. Spread in photon energy due to the Doppler effect.

If Panofsky had measured directly the energy released in the reaction that he was studying, he would have found it to be only 5 Mev. Thus, by designing the experiment so that this was amplified to 28 Mev, he obtained an amplification factor of $28/5 = 5.6$. This is a differential-type measurement, because one does not measure the energy directly, but rather an energy difference.

A precaution that has to be observed in any experiment, whether it is a precision experiment or an exploratory one, is to beware of background. By background is meant all the extraneous effects that enter into a measurement. This applies not only to experiments in physics, but to those of every field as well. In the case of nuclear physics, the term background has a rather specific meaning, i. e., radiations other than those being studied that in some way enter a detector. Background radiations come from cosmic rays, from particles penetrating the shielding of a cyclotron, or from other sources. For example, particles scattering from a material other than the target can sometimes enter the counter. If this happens, a false counting rate is observed. Of course, steps should be taken to reduce the background to as low a value as feasible, but some background is unavoidable. In fact, in some experiments where low levels are being studied, there may be equal contributions from the background and the effect. No matter how low the background is, its value should be determined and taken into account in evaluating one's results.

We will conclude this lecture with a brief discussion of the design of the "nuts and bolts" of an experiment. After the purpose, the method of approach, and the required accuracy have been determined, one then proceeds to design the nuts and bolts. To do this well requires a great deal of practical knowledge and insight. One must know what types of materials and apparatus can be purchased commercially, as well as those that can be fabricated here at the Laboratory. Also, there is the matter of cost. One does not use platinum or osmium when brass will do. On the other hand, one should not practice false economy.

The cost of labor must be considered in addition to that of the material. Brass is more expensive than steel, but it is so easily machined that the total cost is often less. Also, brass does not rust. Stainless steel does not either, but it is one of the more difficult materials to work. For example, we often make cryogenic containers of Mylar (rather than glass) since it does not crack when filled with liquid hydrogen or liquid helium. Mylar is the same material as the Orlon used in textiles except that it is called by a different name when it comes in blocks or sheets. Mylar is as strong as steel on the basis of weight, it is tough, and it is inexpensive. Like other plastics, Mylar can be easily formed into various shapes. Also, it comes in sheets as thin as 0.001 inch.

Very useful cements for use with Mylar or many other substances are the epoxy resins, which maintain their adhesive properties over a temperature range from absolute zero up to approximately 100°C.

In designing the nuts and bolts one must know the limitations of the materials that he is going to use, such as their elasticity, hardness, etc. These constants should preferably be memorized for the more common materials so that it is not necessary to consult the appropriate tables at every step.

Another essential in designing the nuts and bolts is to be familiar with how Nature itself operates in the range being studied. An example of this is in regard to designing an electromagnet. Steel saturates at perhaps 15 to 20 thousand gauss. At higher field strengths steel is no more effective than air, so it would be senseless use steel in a 100-kilogauss magnet.

A design that looks good on paper may be overly optimistic unless all the relevant effects have been considered. Suppose that you are designing an electromagnet to give a field of one hundred kilogauss (a very unusual magnet!). You might use some standard formula for computing the number of ampere-turns required. But your answer would be considerably in error unless you took into account the heat loss in the coils, which becomes exceedingly great at the large currents required to produce a field of this magnitude. Once you become aware of this difficulty you might then decide to utilize the phenomenon of superconductivity to circumvent the heat losses. This involves cooling the coils down to a temperature very close to absolute zero in order to reduce the resistance almost to zero. Unfortunately, in a field larger than a few thousand gauss, a conductor loses all its superconductivity. Therefore, this approach would also be fruitless. Very recently a suggestion came from the Livermore Laboratory for obtaining very strong magnetic fields by cooling the coils down to about 80°K, in the case of copper coils. Although superconductivity does not exist at this temperature, the resistance of the coils is greatly decreased. This reduces the energy loss in the coils. With this technique one can get by with less copper and steel for a given field strength. But the gain is largely offset from a financial point of view because a large refrigerator is now required to cool the coils. However, in those applications where a small magnet is needed, this method may be very useful. For example, sometimes only a limited space is available near an accelerator for a magnet but the refrigeration equipment can be located remotely. Another advantage of this might be that the bulk of the heat loss occurs not in the experimental area, but at some remote point where it does not interfere with any measurements.

Bibliography

1. R. Furth, The Limits of Measurement, Scientific American, 48 ff. (July 1950).
2. B. T. Matthias, Superconductivity, Scientific American, 92 ff. (Nov. 1957).
3. J. Strong, Procedures in Experimental Physics (Prentice-Hall, New York, 1938).

H. HIGH-ENERGY PARTICLE ACCELERATORS

David L. Judd

The first accelerated particles used in physics research were supplied by natural radioactivity. Beta rays and alpha particles of a few Mev were available from these sources. Not until the end of the nineteenth century were particles first accelerated artificially, by J. J. Thompson and others, using direct-voltage machines. The principle of machines of this type is illustrated in Fig. H-1.

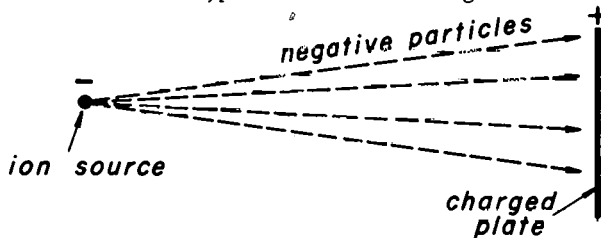


Fig. H-1. A direct-voltage accelerator for charged particles.

Charged particles emitted from an ion source, such as a heated filament, are attracted toward an oppositely-charged metal plate. The apparatus is enclosed in a tank which is evacuated so as to keep the particles from scattering as a result of collisions with air molecules. To accelerate negative electrons, for instance, the plate is positive. The amount of acceleration is limited by the voltage that can be applied to the plate.

Voltages much higher than those previously available were obtained in the 1920's with the Cockroft-Walton voltage multiplier and the Van de Graaff generator. In fact, protons with sufficient energy to cause nuclear transformations were produced by Cockroft and Walton in 1932. These machines are restricted to a few Mev because they operate on the "one-push" principle. That is, particles experience but one push in a single electric field before they strike a target.

Modern high-energy machines make use of repetitive acceleration, in which a particle is made to speed up by giving it a series of pushes. The first successful repetitive accelerator was the Sloan-Lawrence linear accelerator, built in 1928 here at Berkeley. (See Fig. H-2.)

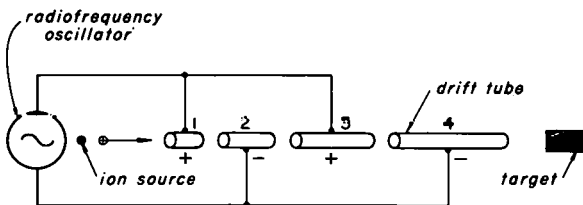


Fig. H-2. A linear accelerator.

Particles emitted from an ion source are attracted toward an oppositely-charged cylindrical tube. Once inside the tube, the particles are shielded from the electric field and merely drift through. In the meantime the sign of the potential on the drift tubes has been changed. This is done by connecting the drift

tubes to a radiofrequency oscillator, a source of high-frequency alternating potential. When the particles emerge from the first drift tube, they are repelled by it, since it now has the same sign as the particles. However, the particles are attracted toward the second drift tube because it has the opposite sign. The particles enter the second drift tube, and again the sign of the potential on the drift tubes is changed. As this process is repeated, the particles gain energy each time they are accelerated across a gap. Owing to their increased velocity, the particles travel progressively farther in a given time interval. Therefore, the drift tubes must be made successively longer. If this were not done, the particles would get out of step with the alternating voltage and would arrive at a gap too soon to be accelerated.

With a linear accelerator one can obtain high-energy particles even though a relatively low voltage is applied across each gap. For example, if 100 gaps were used one could obtain 100-Mev particles by applying only about one million volts between each pair of drift tubes.

The early linear accelerators did not accelerate protons, because the oscillators then available could not operate at sufficiently high frequencies. However, since World War II many proton linear accelerators have been constructed. The first of these was the 32-Mev machine built under the direction of Dr. L. W. Alvarez at this Laboratory. A few years ago it was dismantled and sent to USC at Los Angeles. The largest proton linear accelerator, a 60-Mev machine, is at the University of Minnesota.

The heavy-ion linear accelerator (hilac) at this Laboratory also operates on the principle described above. It can accelerate ions as heavy as argon to an energy of 10 Mev/nucleon. This means that the energy of an oxygen ion, for example, would be $16 \times 10 = 160$ Mev.

A linear accelerator is essentially inefficient, as each accelerating gap is used only once for a given particle. Higher energies can be obtained only by using a very large number of gaps. This difficulty was overcome by Professor E. O. Lawrence in 1930 when he invented the cyclotron (Fig. H-3).

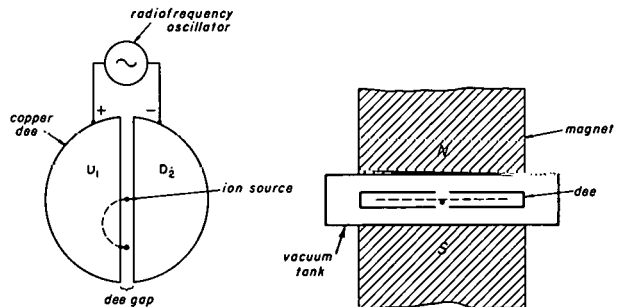


Fig. H-3. A cyclotron.

In a cyclotron a single accelerating gap is used over and over again to accelerate positive ions. This is accomplished by using a magnetic field to bend the ions into circular orbits.

A positive ion is emitted from the ion source into the region between the circular electrodes, or "dees." At that instant one of the dees, say D_1 , is charged negatively. The ion is attracted toward D_1 . Once inside the dee, the ion is shielded from the electric field. It would travel in a straight line were it not for the magnetic field. The magnetic field bends the ion so that it circles around toward the dee gap.

Before the ion reaches the dee gap, the sign of the electric field is reversed by the oscillator, so that now D_2 is negative and D_1 is positive. The ion is consequently attracted toward the negative dee, D_2 . Again the ion is bent by the magnetic field, and again it enters the gap. As it gains speed and energy it spirals outward until it strikes a target inserted in its path or is extracted for use as an external beam.

The sign of the electric field must be alternated at precisely the right frequency, so that when the ion arrives at the dee gap it will be accelerated. The angular frequency is constant for a given type of ion. In other words, it takes an ion the same amount of time to complete a small orbit as a large orbit, because its velocity increases directly as the distance traveled. Because this angular frequency is constant the radiofrequency oscillator can operate at a fixed frequency.

To understand why the above is true, we must consider the forces exerted on the particle. The force F_m due to the magnetic field is given by

$$F_m = \frac{e}{c} vB, \quad (1)$$

where e is the charge of the particle in esu, c is the velocity of light in cm/sec (a conversion factor to put the unit of charge into the electromagnetic system), v is the particle velocity in cm/sec, and B is the magnetic field strength in gauss. The centrifugal force F_c is given by

$$F_c = \frac{mv^2}{r}, \quad (2)$$

where m is the particle mass in grams and r is the radius of the particle orbit. Because F_c must equal F_m we may equate Eqs. (1) and (2):

$$\frac{e}{c} vB = \frac{mv^2}{r}. \quad (3)$$

Canceling and rearranging terms, we obtain an expression for the angular frequency ω , which by definition is v/r :

$$\omega = \frac{v}{r} = \frac{eB}{mc}. \quad (4)$$

It is important to note that ω does not depend on the particle energy. For a given type of particle, ω depends solely on B , the magnetic field strength. Since the magnetic field strength B is held constant in a cyclotron, the angular frequency ω is constant.

The cyclotron is an efficient machine for accelerating particles to energies up to about 20 Mev. By expending large amounts of power, perhaps 100 Mev could be obtained, but this is not very practical.

Two conflicting demands prevent the cyclotron from obtaining higher energies. One of these is that when a particle approaches the velocity of light its mass begins to increase, as explained by the theory of relativity. The total mass m of the particle is given by

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}}, \quad (5)$$

where m_0 is its rest mass, v is its velocity, and c is the velocity of light. From Eq. (5) we see that when v is very small, the total mass m is virtually the same as m_0 . However, as the particle approaches the velocity of light, the ratio v^2/c^2 approaches 1. The denominator thus becomes zero and the particle mass approaches infinity. Figure H-4 illustrates this relativistic mass increase.

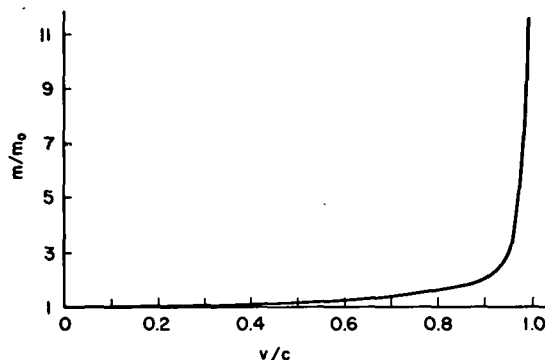


Fig. H-4. Increase in mass of a particle as it approaches the velocity of light c : m is its total mass, m_0 its rest mass, and v its velocity.

From Fig. H-4 it is seen that to accelerate an infinitely heavy object would require an infinite amount of energy. However, there is a problem with a cyclotron if the particle mass increases even by as little as a few percent. Referring back to Eq. 4, we see that if the mass m begins to increase instead of remaining constant, then the angular frequency ω must decrease. As the particles slow down, they arrive at the dee gap late and are decelerated.

The easiest way to keep the particles in phase might seem to be to allow the magnetic field strength B to increase, in order to keep the ratio B/m in Eq. (4) constant. This could be done by changing the shape of the magnet poles to appear as shown in Fig. H-5. This pole shape would cause the field to increase radially.

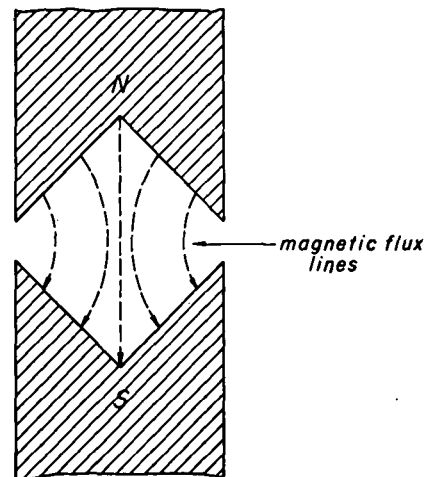


Fig. H-5. Shape of magnet pole faces required to give a radially increasing field (exaggerated).

Unfortunately, this cannot be done, owing to the second conflicting demand. That is the necessity for focusing the particle beam. As a particle spirals outward it occasionally collides with molecules of the residual gas inside the vacuum tank. This causes it to wander from the proper path. The particle gyrates into the dees unless restoring forces exist.

We may analyze the required properties of a restoring force by referring to Fig. H-6. The displacement of the particle from its proper orbit is plotted as a function of time t . We would like the displacement x always to be zero. However, the best we can hope for is that when the particle begins to drift from the proper orbit, a force is created which directs the particle back toward the correct orbit. If this occurs, then the particle oscillates about the correct orbit in the manner shown in Fig. H-6. Such a restoring force F_x must be proportional to the displacement x and be directed so as to minimize the displacement:

$$F_x = -kx \quad (6)$$

(k is a proportionality constant). In a radially increasing field, like that illustrated in Fig. H-5, the force F_x is directed the wrong way. That is, when the particle drifts off the correct orbit a force is created which pushes it farther away, in an exponential manner. The particle quickly strikes a dee.

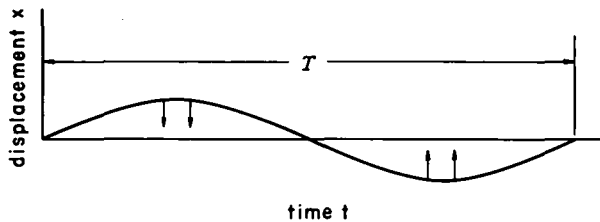


Fig. H-6. Displacement of a particle from its proper orbit, as a function of time.

On the other hand, a practical way to provide the required restoring force is to have a radially decreasing magnetic field, as shown in Fig. H-7. In this type of field, whenever a particle gets off center it is immediately pushed back, since the restoring force is directed toward the midplane.

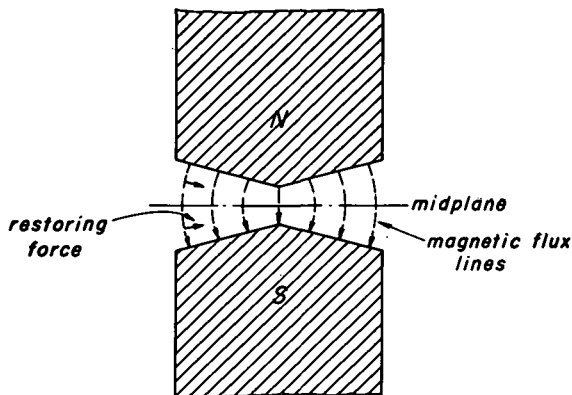


Fig. H-7. Magnet pole faces required to give a radially decreasing field.

However, if we use a decreasing magnetic field we have only complicated the problem of keeping the particles in phase with the alternating potential. A number of suggestions were offered for avoiding this

dilemma. The earliest was by L. H. Thomas in 1938. His ideas were not very well understood, and physicists were so busy building conventional cyclotrons that they took little notice of them. Not until 1949 was a Thomas cyclotron constructed.

In 1945 a different solution was proposed, by Veksler in Russia and independently by McMillan, present Director of the Lawrence Laboratory. Their principle of phase stability was incorporated into the design of the 184-inch cyclotron, then under construction here. As a result, the 184-inch machine was able to produce particles of twice the energy previously planned and yet use much lower dee voltages. The 184-inch cyclotron has since been rebuilt so that now it gives even higher energies (735 Mev, in the case of protons).

The basic idea of phase stability is that it is unnecessary to try to keep the particle exactly in step with the alternating voltage. It is much better to allow the particle to oscillate back and forth slightly in phase with respect to the applied voltage. The principle of phase stability is illustrated in Fig. H-8. The change in energy ΔE that a particle would receive as it passes through the dee gap of a static system (constant field, constant rf frequency) is plotted as a function of time. The points where the curve crosses the horizontal axis are equilibrium positions. What are the conditions to be met in order that these be stable?

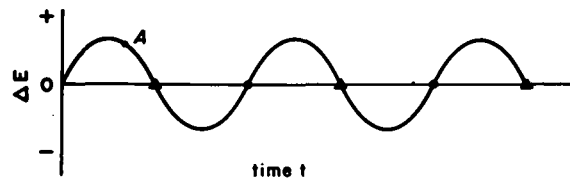


Fig. H-8. Change in energy ΔE that a particle receives as it passes through the dee gap, as a function of time (for a static system).

Consider a particle at Point A. It has arrived early at the dee gap and receives a positive increment of energy, causing it to travel in too large an orbit. Therefore, its angular frequency decreases so that it arrives at the dee gap later on its next encounter. It may arrive on time, but if it arrives too late its angular frequency increases and it arrives early the following time. In this way, the particle oscillates about its correct energy at a given orbit. The positions marked with a triangle are stable for a cyclotron or other simple circular machine. Those marked with a dot are stable for linear accelerator.

Once these equilibrium positions are stable, one can then carefully decrease the frequency of the potential applied to the dees to keep it in step with the decrease in angular frequency of the particles caused by their relativistic mass increase and by the change in magnetic field required for focusing. As the radiofrequency is decreased, the particles are accelerated to higher energies.

A machine which uses this principle is called a synchrocyclotron or a frequency-modulated cyclotron. This type of accelerator emits a pulsed beam. One batch of particles is injected from the ion source, and the frequency of the dee potential is then decreased through a certain range to accelerate the particles to full energy. This process is then repeated for another batch (See Fig. H-9). The average beam current of a synchrocyclotron is much

less than for a conventional cyclotron, which emits a steady beam.

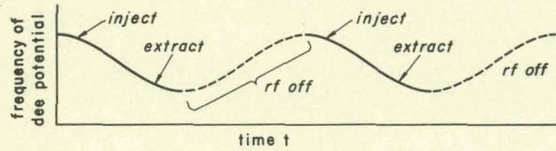


Fig. H-9. Variation of frequency of the dee potential, as a function of time (for a synchrocyclotron).

The principle of phase stability has also made possible the multi-billion-volt proton synchrotrons. Four of these are now operating — the Cosmotron at Brookhaven, the Bevatron at Berkeley, the Synchrophasotron in Russia, and the Saturne at Saclay, France.

In a proton synchrotron the particles are accelerated in a circular orbit of constant radius. This is done by increasing the magnetic field strength at the same time the frequency of the potential on the accelerating electrodes is increased. (Fig. H-10.) When the particles are injected, the field is very low. As the particles gain energy from the accelerating electrodes, the magnetic field must increase in order to keep the particles in a circular orbit.

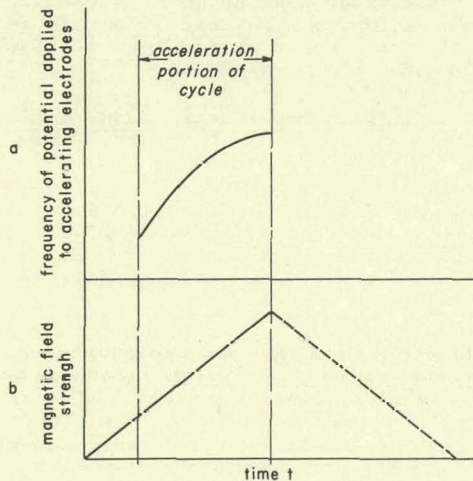


Fig. H-10. (a) Frequency of the potential applied to the accelerating electrodes, as a function of time; (b) magnetic-field strength, as a function of time. (These curves are for a proton synchrotron.)

Figure H-11 shows the 6.2-Bev Bevatron. The radius of its accelerating "donut" is 50 ft. Although its magnet contains 10,000 tons of steel, a synchrocyclotron of the same energy would require a magnet many times heavier. This is because a proton synchrotron uses a ring-shaped magnet, rather than a solid one.

A proton synchrotron requires an injector. In the case of the Bevatron, the injector must accelerate the protons to an energy of 10 Mev before they are traveling fast enough to be accelerated by the Bevatron itself. The injector consists of a Cockcroft-Walton machine, which produces about 1/2 Mev, and a linear accelerator which next takes over and boosts their energy to 10 Mev. The protons are then introduced into the Bevatron via an electrostatic inflector, which bends the beam of particles into a direction tangent to the accelerating chamber.

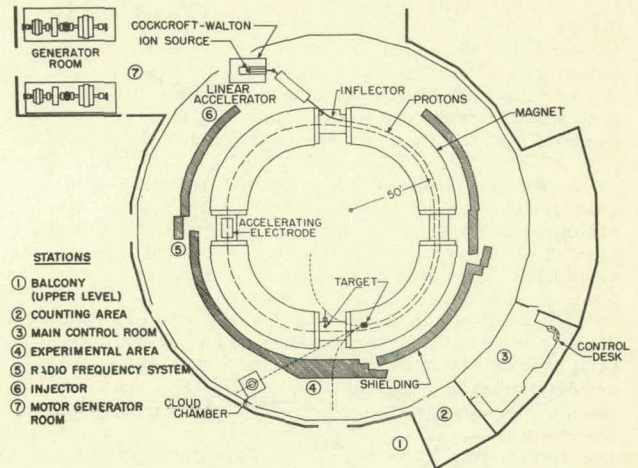


Fig. H-11. The Bevatron, the 6.2-Bev proton synchrotron at Berkeley.

Figure H-12 is a particle-eye view of the Bevatron. The small tubes seen on the top and bottom are coil windings that are used to make minor adjustments in the magnetic field strength. The flags sticking up in the accelerating chamber are targets. They lie flat while the particles are being accelerated (about 1.75 sec, during which the protons make 4,100,000 revolutions), and then are flipped up into the full-energy beam.

Several giant proton synchrotrons now under construction throughout the world will use the new principle of alternating-gradient focusing. For example, the 30-Bev machine at Brookhaven has a diameter of 840 feet. Yet the beam must be guided through a tube only a few inches in diameter. Very strong focusing is required to do this.

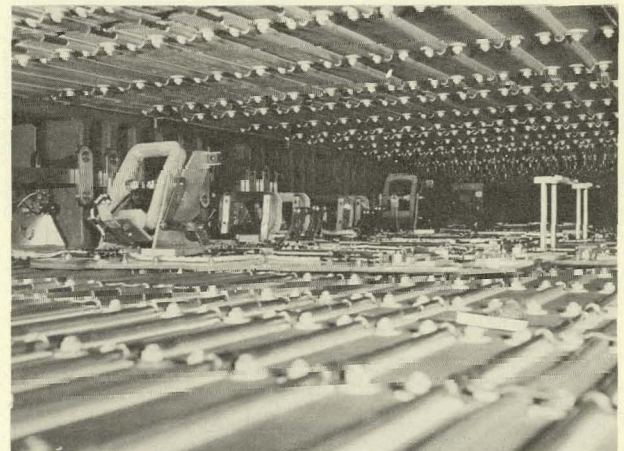


Fig. H-12. Interior of the accelerating chamber of the Bevatron.

The idea of alternating-gradient focusing is illustrated in Fig. H-13. The "ideal" orbit is represented by the dotted circle. With conventional focusing, when the particle begins to drift off this idealized orbit, a restoring force is set up which pushes it back where it should be. In alternating-gradient focusing just the opposite is done. First the particle is thrown violently one direction, and then violently in the other by properly shaping the magnetic field. The net effect, surprisingly, is to

produce much sharper focusing than is possible by the conventional method.

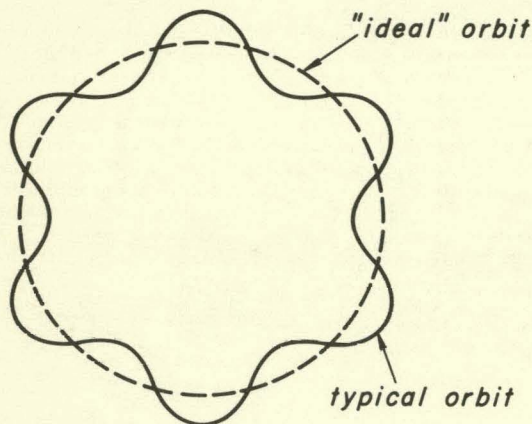


Fig. H-13. Diagram illustrating alternating-gradient focusing.

The principle of alternating-gradient focusing can be easily demonstrated using an upside-down pendulum (Fig. H-14). Everyone knows that this is unstable — as soon as the pendulum moves off center it continues to fall over. If its support is given a hard jerk upward when the pendulum starts to fall over, it tends to fall even faster. On the other hand, if its support is jerked downward, the pendulum returns to its upright position. Now if this process of jerking up and down is repeated at a fast enough rate, the pendulum becomes stable and oscillates slightly about an upright position. If you should deflect the pendulum it immediately returns to the upright position.

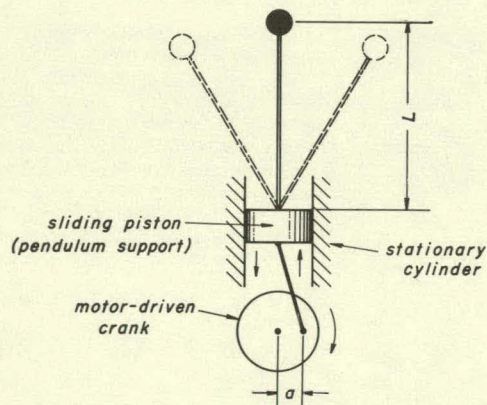


Fig. H-14. An upside-down pendulum illustrates the principle of alternating-gradient focusing.

It can be shown that the frequency f at which the pendulum support must be driven up and down in order for the pendulum to be stable must be greater than a certain value:

$$f > \frac{\sqrt{2 L g}}{2 \pi a}, \quad (7)$$

where L is the length of the pendulum arm, g is the acceleration of gravity (32 ft/sec^2), and a is the radius of the motor-driven crank (Fig. H-14) which drives the pendulum support up and down.

Bibliography

1. J. P. Blewett, Recent Developments in Proton Synchrotrons, in *Annual Review of Nuclear Science*, J. G. Beckerley et al., Ed., Vol. 4 (Annual Reviews, Stanford, California, 1954), pp. 1-12.
2. E. L. Chu and L. I. Schiff, Recent Progress in Accelerators, in *Annual Review of Nuclear Science*, Vol. 2 (Annual Reviews, Stanford, California, 1953), pp. 79-92.
3. B. L. Cohen, Cyclotrons and Synchrocyclotrons, in *Handbuch der Physik*, S. Flugge, Ed., Vol. 44 (Springer-Verlag, Berlin, 1959), pp. 105-169.
4. S. Glasstone, *Sourcebook on Atomic Energy*, Second Edition (Van Nostrand, Princeton, 1958), Ch. IX.
5. G. K. Green, The Proton Synchrotron, in *Handbuch der Physik*, S. Flugge, Ed., Vol. 44 (Springer-Verlag, Berlin, 1959), pp. 218-340.
6. D. L. Judd, Conceptual Advances in Accelerators, in *Annual Review of Nuclear Science*, E. Segre et al., Eds., Vol. 8 (Annual Reviews, Stanford, California, 1958), pp. 181-216.
7. M. S. Livingston, *High-Energy Accelerators* (Interscience Publishers, New York, 1954).
8. M. S. Livingston, *High-Energy Accelerators (Proton Synchrotron)*, in *Annual Review of Nuclear Science*, Vol. 1 (Annual Reviews, Stanford, California, 1952), pp. 169-174; (Standard Cyclotron), pp. 157-162; (Synchrocyclotron), pp. 163-168.
9. E. M. McMillan, *Particle Accelerators*, in *Experimental Nuclear Physics*, E. Segré, Ed., Vol. III (Wiley, New York, 1959), Part XII.
10. W. Panofsky, *The Linear Accelerator*, *Scientific American*, 40 ff. (Oct. 1954).
11. J. C. Slater, *High-Energy Accelerators (Linear Accelerators)*, in *Annual Review of Nuclear Science*, Vol. 1 (Annual Reviews, Stanford, California, 1952), pp. 199-206.
12. L. Smith, *The Bevatron*, *Scientific American*, 20 ff. (Feb. 1950).
13. L. Smith, *Linear Accelerators*, in *Handbuch der Physik*, S. Flugge, Ed., Vol. 44 (Springer-Verlag, Berlin, 1959), pp. 341-389.
14. Thomas, Kraushaar, and Halpern, *High-Energy Accelerators (Synchrotrons)*, in *Annual Review of Nuclear Science*, Vol. 1 (Annual Reviews, Stanford, California, 1952), pp. 175-198.
15. R. R. Wilson, *Electron Synchrotrons*, in *Handbuch der Physik*, S. Flugge, Ed., Vol. 44 (Springer-Verlag, Berlin, 1959), pp. 170-192.
16. R. R. Wilson, *Particle Accelerators*, *Scientific American*, 64-76 (Mar. 1958).

I. SURVEY OF PARTICLE DETECTORS

Roger W. Wallace

All radiation detectors utilize in some way the phenomenon of ionization, which is the formation of positive and negative charges when an atom is struck by an incoming charged particle or photon. Radiation detectors may be classified according to the method utilized in detecting this ionization. These categories are:

- a. the Geiger counter, proportional counter, and ionization chamber;
- b. the cloud chamber and bubble chamber;
- c. the photographic emulsion;
- d. the scintillation and Cherenkov counters.

Detectors in category (a) are seldom used in high-energy physics research, but find their widest application in industry, in radiological monitoring, and in low-energy research. The cloud chamber and photographic emulsion are frequently used in cosmic ray research. The bubble chamber, the scintillation counter, and the Cherenkov counter, as well as the photographic emulsion, are the principal detectors employed in present-day high-energy physics.

Ionization Instruments

The detectors in category (a) are usually referred to collectively as "ionization instruments," although it must be borne in mind that all radiation detectors utilize ionization in some manner. The reason for this designation is that the detectors in this category produce an electrical signal as a direct consequence of the production of ionization, whereas in the other types of detectors the ionization serves to initiate or trigger a release of energy or brings about a transformation of energy by some other mechanism. This secondary release of energy is then recorded.

An ionization instrument consists basically of an enclosed vessel containing a gas (Fig. I-1). The walls of the vessel are metalized, forming an outer electrode. A central electrode is connected to an amplifier and counting circuits. When a charged particle enters the vessel through one of its walls the gas inside is ionized producing an electron and a positive ion (called a primary ion pair). The electron quickly travels toward the central electrode, while the positive ion migrates slowly to the outer electrode. By using an electronic detecting circuit

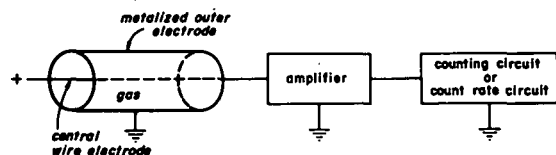


Fig. I-1. Basic arrangement of an ionization instrument.

with a short time constant (i. e., a fast circuit), one can record the pulse due to the electron and then ignore the ion pulse. (However, this is not always done.)

Curve A in Fig. I-2 represents a case in which a larger number of primary ion pairs is produced than for curve B. It is evident that these curves can be divided into six regions. In region I the pulse height increases with an increase in the applied voltage. In region II the pulse height is unaffected by an increase in the applied voltage. It is in this region of about 100 to 300 volts that an ionization chamber operates. If should be noted that the geometry shown in Fig. I-1 is more common to the Geiger and proportional counters; ion chambers frequently have parallel plates instead.

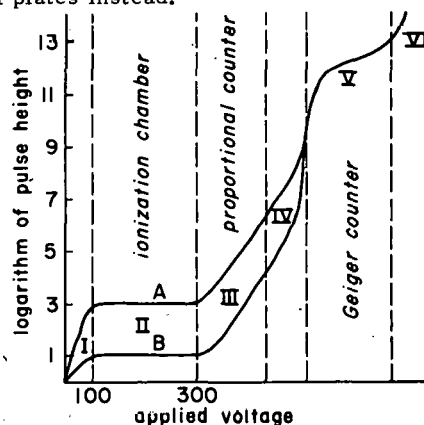


Fig. I-2. Dependence of pulse height on applied voltage in an ionization instrument. Curve A represents a case where more primary ion pairs are formed than for Curve B.

In region III of Fig. I-2 the pulse height again increases with an increase in the applied voltage, owing to an effect at the central wire called multiplication. (See Fig. I-3) The field gradient near the wire is sufficiently high so that an electron can pick up enough energy between collisions with the gas to ionize many molecules. This results in the release of additional electrons, which in turn cause the release of even more electrons. For a given voltage, the multiplication is constant. This is evidenced by the fact that the vertical separation between curves A and B of Fig. I-2 is constant in region III. Thus, for a given voltage the pulse height is proportional to the number of primary ion pairs formed. The proportional counter operates in this region.

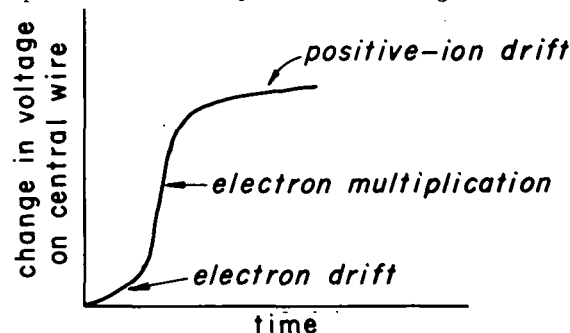


Fig. I-3. Buildup of a single pulse in a proportional counter.

The Cloud Chamber

In region IV of Fig. I-2 the curves again rise with an increased voltage, but at different rates. No radiation detector operates in this zone.

In region V of Fig. I-2, curves A and B coincide, indicating that for a given voltage the pulse height is independent of the number of primary ion pairs formed. The Geiger counter operates in this region. What is occurring is that the multiplication process has become so energetic at these higher voltages that photons of light are released, which ionize other gas atoms near the wire as they proceed down the vessel. This causes a cascade all along the wire, a phenomenon known as a corona discharge. The discharge is so great that the pulse height is no longer proportional to the applied voltage, but is much larger. Thus all information is lost as to the size of the original pulse. One incoming particle forms as large a Geiger pulse as ten million particles do.

However, with an ionization chamber or proportional counter, one can readily tell whether one particle enters or two, and whether it is a heavily ionizing particle or a lightly ionizing particle. Therefore, the ionization chamber and proportional counters are very useful as research instruments. An important technique involving these two instruments is that of pulse-height analysis. By means of it one may (under certain conditions) sort the radiation being observed as to the number of alpha, beta, or gamma particles it contains.

The great advantage of the Geiger counter is its simplicity. For example, one does not have to control its voltage very closely. Its chief limitation is that its dead time is of the order of 100 microseconds, owing to the time required for the discharge to be quenched. Any particle entering during that time is not recorded. If one particle entered every microsecond, the Geiger counter would record only 1% of the incoming particles, since it can produce only one count every 100 μ sec. For this reason, one should be very suspicious of the readings of Geiger counters, especially if they are used in an area of high instantaneous flux such as near an accelerator.

On the other hand, a proportional counter has a dead time of only 2 to 3 μ sec. While this is a considerable improvement, there is still a problem of lost counts when the counting rate is high. Two particles entering at once cause a pulse twice as high as that produced by a single particle. A third particle entering during this period causes the pulse to pile up even higher. These pileups become confusing when one is attempting to sort out gamma and alpha radiation, because a pileup of gammas may look like the pulse of an alpha. Due to this limitation the proportional counter is also of little value in high-energy experiments. It may be very useful elsewhere, however.

The ionization chamber also is beset by this problem of dead time; and in addition it is very sensitive to vibration, which may cause it to give a spurious count. This difficulty is partially overcome by using amplifiers that are insensitive in the audiofrequency region. One of the most useful types of ionization chamber is the integrating type, which does not detect individual pulses, but instead adds up the current. This type of instrument is often used for detecting background radiation.

The cloud chamber and bubble chamber are track-imaging devices. In the former, an incoming particle produces a track of droplets in a super-saturated gas, whereas in the latter the track consists of tiny bubbles in a superheated liquid. Since a separate lecture is devoted to bubble chambers, we shall confine our attention mainly to the cloud chamber.

Figure I-4 is a pressure-volume plot of representative isothermal curves for a real gas. The top curve ($T > T_k$) is for a gas that is above its critical temperature (the temperature at which a liquid and its saturated vapor coexist in equilibrium). In the case of a perfect gas, as the volume increases the pressure drops according to the $PV = NRT$ rule, where P is the pressure, V is the volume, N is the number of moles, R is the perfect gas constant, and T is the absolute temperature. The perfect gas approximation is fairly good in the vapor region to the right of the two-phase region.

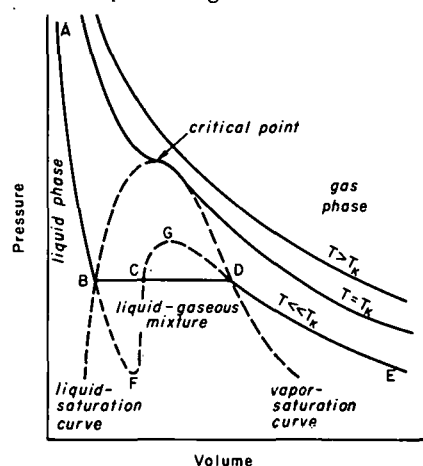


Fig. I-4. A pressure-volume plot of isothermal curves for a real substance.

However, for a substance that is at or below its critical temperature this rule no longer applies; instead the curve is more complicated (see the lower curve in Fig. I-4). The solid curve ABCDE represents the stable condition. In region AB the substance is in the liquid state, in region BCD it is in the liquid-gaseous or two-phase state, and in region DE it is in the gaseous state. If the substance is in the condition represented by CD, and the pressure is suddenly reduced, the curve jumps to CGD, an unstable state. That is, the gas becomes super-saturated.

A charged particle passing through this super-saturated gas ionizes the atoms, leaving ion pairs. The gas then condenses on these ions, producing a track of tiny droplets. Thus, the charged particle triggers an unstable condition, causing it to return to a stable state. The energy thus released to restore this stable state produces an amplification mechanism capable of rendering visible the tracks of all particles, including those producing minimum ionization. (Minimum ionization occurs for all charged particles at an energy approximately equal to the rest-mass energy of the particle; at this energy the particle produces the fewest ion pairs per centimeter of track.)

The cloud chamber has proven to be a very useful instrument. Its beauty lies in the fact that one can see what is happening to individual particles. If the chamber is placed inside a magnetic field, charged particles are caused to curve, the positive ones in one direction and the negative particles in the other. Their radius of curvature r is given by

$$r = \frac{mv}{Be}$$

where m is the mass of the particle, v is its velocity, e is its charge, and B is the magnetic field strength. Thus, by knowing three of these variables, the fourth can readily be calculated.

Unfortunately, cloud chambers suffer from several disadvantages. Since a gas is not very dense, relatively few nuclear collisions occur. A great deal of time and effort are used waiting for an interesting reaction. This can be partially overcome by placing lead plates inside the chamber. However, since the reactions occur within the metal, one sees the tracks of the reaction products only after they leave the lead.

Another problem is that large cloud chambers suffer from turbulence in the gas and nonuniformity in the expansion process, the latter effect causing the chamber to be insensitive in certain regions.

One of the most serious limitations of the cloud chamber is that it can be cycled only once every minute or two, although a beam pulse from an accelerator is usually available every few seconds or less. In cosmic ray research this is not usually a problem, owing to the comparatively long intervals between the entrance of particles. To record the passage of a cosmic ray, one places a Geiger counter above and below the chamber. These are connected electronically in such a way that a particle traversing both the upper and lower set of counters triggers the expansion of the chamber, flashes the illuminating tubes, and exposes a photograph in the proper time sequence. This technique was developed by Carl Anderson of the California Institute of Technology, who discovered the positron with such an arrangement.

Various methods have been developed for improving the cloud chamber. One is to make it continuously sensitive by the diffusion technique. In the diffusion chamber a volatile liquid such as methyl alcohol is vaporized in a warm region at the top. The vapor diffuses toward the bottom, which is cold. In an intermediate layer the gas is in a supersaturated condition and is therefore sensitive to ionizing radiation. The gas is usually put under a pressure of 20 to 35 atmospheres to increase its density and thus allow more nuclear collisions. Unfortunately, the sensitive layer is very shallow. Those reactions occurring above or below it are not observed.

One great advantage to both types of cloud chamber is that the droplet density along a track is related to the ionization density of the initiating particle.

Scintillation Counters

Scintillation counters are widely used in high-energy physics, owing to their extremely short time resolution of 10^{-9} sec. A scintillation counter consists of a transparent material called a scintillator that emits a flash of light when it is struck by a charged particle, and a photomultiplier tube for observing this flash (Fig. I-5).

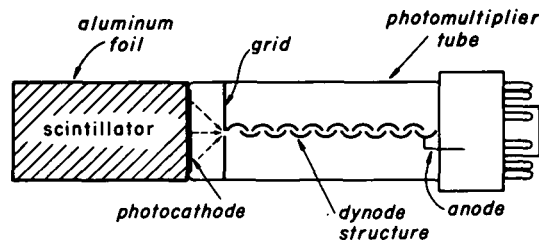


Fig. I-5. A scintillation counter.

Scintillators

The various types of scintillators and their characteristics are summarized in the appendix. The factors that must be considered in the choice of a scintillator are:

- its speed (i. e., how rapidly the light pulse forms and dies away);
- its light yield (i. e., the size of the light pulse for a particle of a given energy);
- its sensitivity to different types of radiation;
- the linearity of its response to different types of radiation (this is the most important consideration);
- its density (which affects the number of nuclear events occurring in its volume); and
- its ease of manufacture.

The organic solids like anthracene and stilbene are difficult to cut without cracking. Also, they are temperature-sensitive. Their chief advantage is that they have an excellent light yield. Moreover, they are very fast (10^{-9} sec). Their sensitivity to gamma radiation is low. This effect depends on the atomic number of the material; for hydrocarbons the gamma sensitivity is low, whereas for heavy elements it is high. The neutron sensitivity, on the other hand, is very good for the organic solids, because neutron sensitivity is dependent on the hydrogen content. The higher the hydrogen content, the more proton recoils one obtains from neutrons. In the case of organic solids the linearity of the response to electrons is good, but for protons it is poor. Linearity is dependent on the spacing of the primary ion pairs produced by the incident ionizing radiation. Most scintillators are quite linear to electrons, because electrons leave the ion pairs spaced rather far apart, and jamming of the light-production process does not occur. However, this is not true for other types of radiation. From Fig. I-6 we see that separate curves

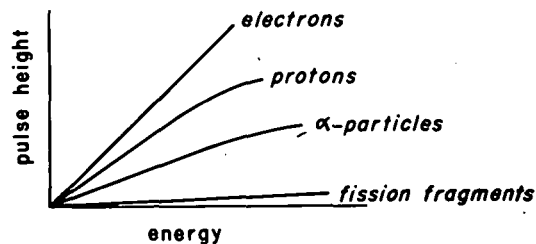


Fig. I-6. Pulse-height--energy relationships for particles with different linear ionization rates in a typical scintillator.

are obtained for protons, alphas, and fission fragments. We would prefer to obtain one curve, a straight line, for all of these so that the energy of the various particles could be measured directly.

The most commonly used inorganic solids are sodium iodide (with a thallium activator), potassium iodide, and lithium iodide. Their chief drawback is their slow speed of 10^{-5} to 10^{-6} sec, as compared with the 10^{-9} -sec speed of the organic solids. An advantage is their high density.

The organic liquids like toluene can be used where large volumes are required. For instance, at the 184-inch cyclotron there is a tank of toluene the size of a desk. A drawback is the fact that the toluene deteriorates the gaskets and fittings and is highly inflammable. However, the organic liquids are very fast (10^{-9} sec) and have a good light yield.

The plastic scintillators are solid versions of the organic liquids. They are produced by heating styrene and adding a small amount of an inorganic phosphor. Since plastic can be easily machined into complex shapes, this type of scintillator is the most popular and versatile. Their light yield is good and they are fast (10^{-9} sec).

The noble gases are seldom used, although they have the best linearity. One straight line is obtained for all types of radiation. For example, the pulse produced by a 100-Mev fission fragment is 100 times as large as the pulse from a 1-Mev gamma. With other types of scintillators the pulses from fission fragments are quite small compared to gamma-ray pulses.

Photomultipliers

The photomultiplier tube represented in Fig. 1-5 is of the end-window type. Inside the glass envelope at the end of the tube is a material that emits photoelectrons when it is struck by light. Electrons from this photocathode are accelerated by a potential of ~ 300 volts between the grid and the photocathode. After passing through the grid, the electrons strike a series of dynodes, which are also coated with a photoelectron-emitting material. Therefore, for each electron striking a dynode, four or five are given off. Since there are perhaps 10 to 14 dynodes, one electron striking the photocathode may cause a million to ten million electrons to leave the final dynode. Each dynode is held at a higher potential than the preceding one, in order to obtain an acceleration down the tube. Thus, in a scintillation counter a usable electric pulse is produced by the passage of a single particle through the scintillator. The electronic circuits used in connection with scintillation counters are discussed in Section K "Counters Used in High-Energy Physics."

Bibliography

1. S. Glasstone, Sourcebook on Atomic Energy, Second Edition. (Van Nostrand, Princeton 1958), Ch. VI.
2. Miner, Shackleton, and Watson, Teaching with Radioisotopes (U. S. Atomic Energy Commission, Washington, 1959). (Available from the Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C.; price 40 cents.)
3. R. L. Platzman, What is Ionizing Radiation?, *Scientific American*, 74-83 (Sept. 1959).
4. W. J. Price, Nuclear Radiation Detection (McGraw Hill, New York, 1958).

Appendix

Radiation Detectors Giving an Electrical Output

I. Ionization instruments

- A. These instruments utilize an electric field to collect ions formed by radiation in a gas.

1. For x-rays and γ -rays, ionization is due to:
 - a. Photoelectrons (cross section $\propto Z^2$);
 - b. Compton recoils (cross section $\propto Z/E$ in the range from well above the k-shell energy to > 1 Mev); or
 - c. Positron-electron pair production (cross section $\propto Z^2$ for energies > 1.02 Mev).
2. Charged particles produce primary ion pairs; the energy relationship is

$$\frac{dE}{dx} \propto \frac{z^2 NZ}{mv^2},$$

where v is the velocity of the charged particle, z is the charge of the particle, Z is the charge of the medium, N is the number of atoms of the medium per cm^3 , and m is the mass of one electron.

3. Neutrons do not themselves ionize the gas; but when they enter a detector they give rise to charged particles, owing to recoils or nuclear reactions. These charged particles then ionize the gas. In these indirect ways neutrons can be detected.
- B. The three types of ionization instruments are the ion chamber, the proportional counter, and the Geiger counter.
 1. Ion chamber
 - a. The two kinds of ion chamber are: the electrometer type and the ion-pulse counter.
 - (1) The electrometer (integrating type) gives an indication proportional to the total charge collected in a given time.

Three kinds of electrometers are:

- (a) Electron tube; this type may be portable and has a sensitivity of 10^{-12} to 10^{-15} amp.
- (b) Electrostatic, of which the Lindeman quartz fiber is an example; its sensitivity is 10^{-15} to 10^{-16} amp. This type of ion chamber, as well as type (c) below, is used in laboratory installations.

- (c) Vibrating-reed capacitor (sometimes called a dynamic condenser or a synchronous condenser); this type uses ac amplification in order to achieve a high sensitivity (10^{-16} amp).
- (2) Ion-pulse counter (detects individual pulses)
- (a) Electron collection (fast): a fast amplifier having a low gain in the audiofrequency band must be used in order to detect the electrons but ignore the ions. One of these gases is used: A, CH_4 , CO_2 , BF_3 , or N_2 . For fast electron collection one may use 95% A and 5% CO_2 . The electron-collection speed is 1000 (cm/sec) (cm/volt).
- (b) Ion collection (slow): air can be used, in which case electrons attach themselves to oxygen, losing mobility. Mechanical vibrations can cause a false response (microphonic trouble). The ion collection speed is 1 to 2 (cm/sec) (cm/volt).
- b. Remarks
- (1) Factors affecting dosimetry with ion chambers:
- If:--the counter walls have approximately the same Z as the Z of the gas; and
- the counter is large relative to the range of secondaries in the gas (walls may be neglected; ions collected are proportional to the volume of the gas enclosed); and
- the counter is small (with walls that are thick relative to the range of secondaries in the walls);
- Then: With low-Z walls and using air as the gas, one can calibrate the counter in roentgens. One roentgen/sec gives $1/3 \times 10^{-9}$ amp/cm³ in STP air for x-rays and γ -rays.
- (2) Background
- (a) Cosmic rays give 3 ion pairs/cm³-sec in STP air.*
- (b) Gammas and alphas from walls and earth give 2 to 10 times this.
- (c) Total background is $\sim 10 \mu\text{r/hr}$. Fluctuations in the background limit low-level measurements.
- (3) Recombination of electrons is caused by high gas pressure or high dE/dx . Recombination reduces the pulse height and/or the number of counts.
- (4) Oxygen, water, or chlorine contaminations cause negative ions to form, reducing the response and producing errors.
- (5) For an ion-pulse counter:
- (a) A preamplifier must be used which matches the line, in order to get above the line noise.
- (b) A 10- μamp 1-millivolt output from the ion chamber is necessary in most practical cases.
- (c) A gain of 10^4 to 10^6 is needed.
- (6) Cautions
- (a) Good insulators must be used (10^{17} to 10^{18} $\Omega\text{-cm}$) between the central anode wire and the surrounding cathode.
- (b) High-voltage guard rings should be used to reduce electrical leakage across the insulator and the resulting noise.
- (c) The following factors must be considered when evaluating a measured counting rate:
1. Microphonics
 2. Insulator-leakage noise
 3. Power-line transients
 4. Inductive and radiative pickup
 5. Pileup of small pulses
 6. Multiple grounding paths
 7. The pulse length relative to the time constants of the electronic circuits
 8. The position where the ions are formed changes the pulse height.
- (d) If the desired pulses are small and are present in a background of large pulses, a non-overloading amplifier may be required.
2. Proportional counter
- a. Advantages over the ion chamber
- (1) The pulse is nearly independent of the point of origin, since the volume near the wire in which charge multiplication occurs is a small fraction of the total volume.

- (2) The shape of the counter is not important, only the length of the wire.

b. Advantages over the Geiger counter

- (1) Discrimination against unwanted pulses is possible.
- (2) The proportional counter gives fast pulses (0.1 μ sec).
- (3) It is more reliable.
- (4) It has a longer life.
- (5) The height of the output pulse is proportional to the input ionization, whereas with the Geiger counter all pulses are the same height.

c. Disadvantages of the proportional counter relative to the Geiger counter

- (1) The proportional counter requires better voltage regulation.
- (2) It needs more amplifier gain.
- (3) It requires more linearity of amplification.

d. Specifications of a representative proportional counter

- If: --the radius of the central wire is 1 mil, and
- the radius of the counter chamber is 5 cm, and
- the gas is 5% CO_2 and 95% A at a pressure of 20 cm Hg, and
- the applied voltage is 2000 volts,

Then: the gas multiplication will be 1000.

3. Geiger counter

a. Features

- (1) The pulse spreads along the wire, causing all pulses to be the same size. This is a disadvantage if one wants to separate the various types of radiation.
- (2) The pulse height increases with voltage.
- (3) A pulse is quenched when the field around the wire is canceled by the positive-ion-sheath space charge.

- (4) After a pulse is quenched by the process described in (3), the counter is dead until the positive ions drift outward and the electric field returns to the wire. This dead time is $\sim 100 \mu$ sec. This often makes the Geiger counter useless as a detector in high-energy physics research.

- (5) To improve the resolution of a Geiger counter, the pulse can be deliberately quenched, either externally or internally. External quenching is done by removing the applied high voltage. Internal quenching is effected by adding an appropriate inorganic gas to the gas filling the counter. The latter method is used more often because it is simpler; however, it limits the life of the tube to $\sim 10^8$ or 10^9 pulses owing to the consumption of the quenching gas.

- (6) A nonlinear amplifier with low gain is satisfactory because pulse heights are all the same anyway and there is more internal amplification than with the proportional counter.

- (7) The operating point for the voltage is located on the plateau of the curve of Fig. 1-2, i. e., in region V.

- (8) The Geiger counter counts 100% of the charged particles passing through it, providing a particle does not enter while a discharge is occurring as a result of a previous particle.

- (9) The γ -ray and x-ray efficiency depends on the wall material.

- (10) The Geiger counter is inexpensive, sensitive, and readily portable.

b. Caution: the dose measured with a Geiger survey meter may be in considerable error if:

- (1) counts are more frequent than one each 100 μ sec (since this is its dead time), or
- (2) the character of the radiation is very different from that used in calibrating the instrument, since all pulses are the same size.

II. Scintillation and Cherenkov counters

A. Comparison (see Table I-1)

B. Remarks

1. For electrons >100 kev, one photon is obtained for every 70 ev lost. As a rule of thumb, one photocathode electron is released for a loss of 2.5 kev by an electron in a plastic scintillator. This includes some solid-angle loss.
2. Light pipes using internal reflection in Lucite are widely used to conduct light from the scintillator to the photomultiplier photocathode.
3. Photomultipliers
 - a. Ten to 14 stages of dynodes are used.
 - b. There is a voltage increase across each stage of 100 to 150 volts.
 - c. The gain doubles for each 10-volt rise per stage.
 - d. A gain of 4 to 5 per stage is obtained.
 - e. The total gain may be 10^6 to 10^7 .
 - f. For each μ watt of light input (blue), 20,000 to

80,000 microamperes of output current are obtained.

- g. Ten to 20 incident photons cause one electron to emerge from the photocathode.
- h. Phototubes are very sensitive to magnetic fields, even those under 0.1 gauss. Therefore, the phototube must be shielded with an inner layer of Mu metal and one or more outer layers of soft iron.
4. The pulse-height resolution is proportional to $\sqrt{N}/N$, where N is the number of counts received during a given time interval for a nondecaying source. This is because the arrival times of the particles are random; this is consequently a statistical process.
5. In a Cherenkov counter the number of photons released per cm is

$$500 \left(1 - \frac{1}{n^2 \beta^2} \right).$$

When one of these photons hits a photocathode there is about a 10% chance of an electron being released.

Table I-1

Type	Material	Dens	Decay time (sec)	Rel. β or γ light	$\alpha/\beta\%$ light	Neutron sens.	Ease of fabrication	Linearity with $\int (dE/dx) dx$	Size
Organic solids	anthracene trans-stilbene	1.25 1.15	2.7×10^{-8} 3.7×10^{-9}	100 60	9 9	good good	hard hard	no	few cm
Inorganic solids	NaI (Tl) LiI (Eu) Zn S (Ag)	3.67 4.10	2.5×10^{-7} 10^{-5}	210 75 200	44 95 100	poor thermals poor	fair fair thin	no	to 9"
Organic liquids	xylene and terphenyl and diphenylhexatriene polyvinyltoluene and terphenyl and diphenyl stilbene		$\sim 10^{-8}$	48 48	9 9	good good	any size some problems	no	any
Plastics with organic phosphors	polystyrene and terphenyl	1.1	5×10^{-9}	15		good	easy	no	few feet
Noble gases	He, A, Kr, Xe Gas, Liq, Solid		$\sim 10^{-9}$	small	100	He good Xe poor	need purification	yes	depends on pressure and temp.
Cherenkov media	glass (sometimes with high lead content) liquids, solids gases	2.5-6 1.2-2 0-	$\sim 10^{-9}$ depends on size	small		depends on H content	easy	depends on velocity	20"

J. EMULSION AS A PARTICLE DETECTOR

Harry H. Heckman

Introduction

The importance of nuclear emulsion in particle detection is such that at least one discussion in this series should be devoted to it. This lecture includes information about the history of its use, something about its advantages and shortcomings as a detecting medium, and a brief description of techniques contributing to its effectiveness.

Nuclear Emulsions

Ideally, a nuclear emulsion is one that is sensitive to all particles of any mass, any velocity, and any charge. This implies that it should be sensitive to a single charged particle of any velocity. A typical curve describing energy loss of a particle going through matter is represented in Fig. J-1. The differential dT/dR is its rate of energy loss, and E/M is the ratio of its total energy to its rest mass. At the point of lower velocities (0.001) the rate of energy loss is greatest. As velocity increases, the rate of energy loss decreases until a point is reached where it is at a minimum. From this point, energy loss increases and flattens off. The ideal emulsion is one that ionizes at the minimum point. In the year 1947, emulsion technique permitted the recording of particles with an E/M of approximately 0.05. In 1949 this improved to slightly better than 0.1, and in 1950 the ideal was achieved.

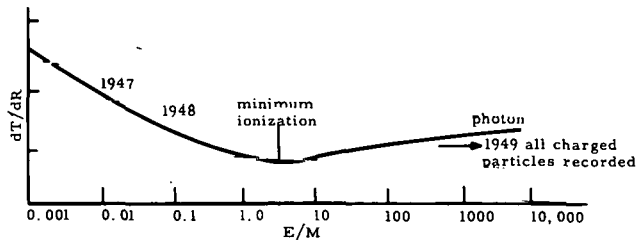


Fig. J-1. Energy loss per unit track length vs E/M , where E is the total energy of a particle and M is its rest mass.

Emulsions of many types have been developed throughout the years of research. As a group, they are capable of noting particles of any charge and any velocity, although individual sensitivities differ. A particular emulsion, for example, may be better than others for recording protons of less than 40 Mev. Another type is used when proton energies are between 50 and 60 Mev. Perhaps one desired an electron-sensitive emulsion, or an emulsion adopted for noting fission tracks. Emulsions have been developed for these widely varied uses.

The emulsions commonly used in high-energy physics are made almost exclusively by the Ilford Company of London, England. This firm produces electron-sensitive emulsions (types known as G.5 and, more recently, K.5). These will record tracks of particles having the very highest velocity. These emulsions are used at the Bevatron, for example, when 6-Bev protons, K

mesons, and antiprotons are studied. Sensitivity of an emulsion is primarily dependent upon the diameters of the silver bromide crystals (the larger the crystals, the greater the sensitivity). The crystals in G.5 emulsions are typically about 0.3μ in diameter. The various emulsions are essentially identical in other respects. They look alike, have the same color, and vary little in chemical composition. As is true in photographic emulsion, the sensitive crystals or grains are silver bromide in a mixture of gelatin. By weight, an emulsion is 80% silver bromide, by volume it is about 50% silver bromide and 50% gelatin. The density of emulsions is high, being about 3.815.

Nuclear emulsions are available to the physicist in many forms and thicknesses, depending upon the requirements of the experiment. Often, the emulsion is simply poured on glass plates, dried, and cut to any suitable size, say 1×3 in. After being poured and dried, the emulsion is peeled from the mounting to produce what are called "pellicles" - sheets of pure photosensitive material which can be stacked together to form an emulsion stack. A typical stack might be a number of these stripped emulsion layers, each 0.6 mm thick, placed together like a stack of cards and secured.

A useful stack for Bevatron experiments is one that is fabricated from some 240 sheets of 0.6-mm-thick emulsion each measuring 6×9 in. The height of the stack would then be about 6 in. This gives the experimenter essentially a solid body of emulsion which is sensitive to particle movements throughout its volume, there being no glass or paper separators. The block is then grooved down the back side so that after the pellicles are separated, their relative positions in the stack during exposure can be determined. After exposure to a beam of π mesons, K mesons, etc., the stack is taken apart. Some kind of a grid is then printed on the surface of each sheet to facilitate the finding of identical positions from sheet to sheet.

What the experimenter ends up with is a library devoted to some set of information. However, before this information is useful each pellicle must be scanned using a compound microscope. This involves measuring 3-dimensional coordinates at representative points along the particle track. The scanner might take the leading pellicle and search for an incident particle track or an event. The stack is aligned in the beam with the pellicles parallel to the beam, so that incident particles tend to move in a single pellicle. However, since interactions are three-dimensional, the tracks of scattered particles may move through the stack in any direction. The scanner must therefore follow a track through the first pellicle to the point where it leaves the local surface. In the second sheet, he locates (by means of the grids) the point where the particle entered and follows it on through succeeding pellicles until the event is reconstructed datawise.

An advantage of emulsions over those detectors having a certain "dead" time is their continual sensitivity. This is not an unmixed advantage, how-

ever, because from the time they are manufactured until developed, any incident particle produces ions in the volume of the emulsion. It is desirable, therefore, to make up the stack as close to the time of exposure as practicable in order to minimize the inevitable background, largely from cosmic radiation. A stack of emulsion standing around three or four months could be sufficiently contaminated by background radiation that one might consider not using it at all. The emulsion has been, however, an indispensable tool of cosmic ray work. Since they are comparatively simple and lightweight, blocks of emulsions can be attached to balloons. Whatever particles they "see" at high altitudes are recorded.

Still another advantage of emulsions is their high spatial and angular resolution. When tracks are observed through the microscope, a very small volume is observed. Distances are resolved, not to meters or centimeters, but to microns (1 micron = 0.001 mm).

The outstanding disadvantage is the uncertainty of interaction products. Since emulsions are heterogeneous substances, it is not possible to say from what nucleus, for example, a scattered proton may have come. It may be from the light elements, C, N, or O (that constitute the gelatin), or from the heavier elements, principally silver and bromine. Track-following, too, is difficult and does not lend itself readily to automatic devices. A track may be 10 centimeters long, or longer, and move up or down through the emulsions, and the scanner can see only a few microns at a time.

Emulsions proved their worth, however, and came fully into their own in the years 1947 and 1948, especially through the efforts of Professor Powell and his co-workers at Bristol, England. By use of emulsions they were able to locate and identify the π meson (which earlier had been predicted by Yukawa) as the particle responsible for holding the nucleus together. From that time, emulsions have been widely used in the study of interactions, masses, decay modes, and numerous other characteristics of mesons.

About this same time machine physics began to dominate the field, and techniques developed in cosmic ray work were now applied to this new area. Figure J-2 shows the first published photograph of an artificially produced π meson recorded in emulsion. The π^- entered at the left and came to rest at the point of the star at the extreme right of the track. Notice that the tracks are composed of

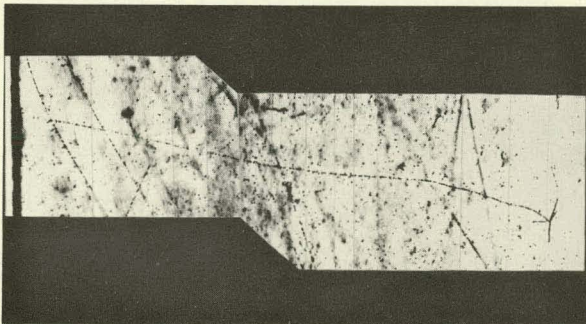


Fig. J-2. First published photograph of an artificially produced π meson recorded in emulsion (Gardner and Lattes, University of California Radiation Laboratory, Berkeley).

small bunches of silver grains. Observe, too, the intrinsic scatter of a particle in emulsion and how the grain darkens as the particle slows down. Being negative, the π was captured by a nucleus in the emulsion. Its rest mass was transformed into energy, and the resultant kinetic energy can be observed as a star. Here a four-pronged star was initiated by the π explosion. A particle thus leaves its own signature in the emulsion. An ability to read these signatures comes from experience and careful observation. Tracks of π^- and μ^- mesons look much alike. By covering up the star, the track perhaps could not be correctly identified. The presence of the star, however, identifies it specifically as a π^- meson.

In Fig. J-3, two particles of different masses are shown coming to rest in an emulsion. Note again the π meson with its characteristic scattering and star. Again, the star is four-pronged. This is not the most probable number; actually, two or three prongs are more apt to be observed. Below is a proton coming to rest. Note that it scatters less than the π meson (owing to its greater mass) and that there is no interaction at the point of rest. The proton is positively charged, and at the point where its kinetic energy is lost it simply acquires an electron and becomes a typical hydrogen atom. Ionization, however, increases toward the end.

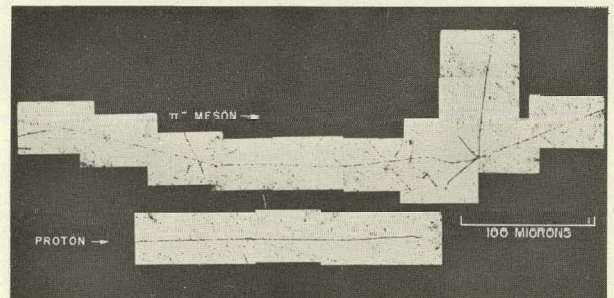


Fig. J-3. Proton and pion coming to rest in emulsion.

Figure J-4 shows the signature of a π^+ meson. Since this meson carries the same charge sign as a nucleus, it will be repelled by nuclei. When it comes to rest, the π^+ undergoes spontaneous decay into a μ meson and a neutrino. It is a typical two-bodied decay, so that the neutrino moves out in a direction diametrically opposite that of the μ . In order that energy be conserved, this decay always involves a μ meson with a unique energy. Here then is a chain of events. The π^+ comes to rest in the emulsion and decays into a μ meson. The track of the latter is similar to that of the former. The μ must carry the charge of the π^+ ; hence it too will come to rest without being absorbed into a nucleus when it decays (the product will be a positive electron, i. e., a positron). The π decay shown here is in an electron-sensitive emulsion; otherwise the positron would not have been detected. Coming off simultaneously with the μ^+ are a neutrino and antineutrino. The energies of positrons resulting from μ^+ decay are therefore not unique but may have a wide range of energies. Pictures like this helped to solve the whole riddle of the π meson. It is now known that π mesons are produced by cosmic rays. They decay in about 1/100 billionth of a second, so that about one half of them are gone in the first 1/100 millionth of a second after production. Therefore, not many π mesons live to reach the earth. Muons, however, interact

less strongly with matter than do pions and also live some hundred times longer. For these reasons, the mesons observed at sea level are all muons.

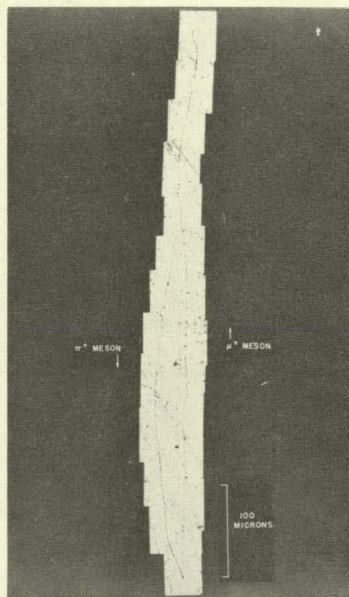


Fig. J-4. Signature of the π^+ meson.

In Fig. J-5 we have yet another particle. Recorded in Ilford G.5, it is called a K meson. With the discovery of the π meson, cosmic ray researchers sought intently for other particles. One such particle looked much like the π but had characteristics which failed to conform to its signature, as already noted. It was neither a π meson nor proton. Subsequent study showed it to have three or four times the mass of the π and about one half the mass of the proton. The K meson in this picture was produced by the Bevatron in an electron-sensitive emulsion. The multiple scattering is less than that for the π . After careful study of these particles it might be detected that the K scatters a little more than the heavier proton. In this instance the particle decays by beta emission (e^+), which is one of its rarer modes.

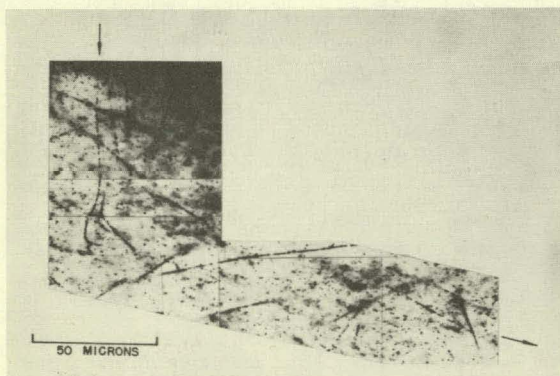


Fig. J-5. The K meson, as recorded in Ilford G.5 emulsion.

Figure J-6 is the history of a K particle. It is that mode of decay of a K particle which has been called the tau (τ) and decays into three pions. The production sequence is at the left. An incident 6.2-Bev proton strikes an emulsion nucleus, probably

silver or bromine. The result is a terrific "blast" and a shower of alphas, protons, and almost every conceivable particle. The tau is produced and moves to decay into three mesons. Were one to observe this decay on edge, it would be seen to lie in one plane (a coplanar decay). Moving out at the upper right is the π^- , resulting in a small star. The π^+ moving upward leaves the emulsion and cannot be followed, but that moving downward comes to rest. It in turn decays into a muon, and the muon then into an electron. The tracks, apart from revealing the signature of the particles producing them, illustrate the history of the degradation of energy: the transformation of mass energy of the unstable particles ultimately to electrons, positrons, and gamma radiation.

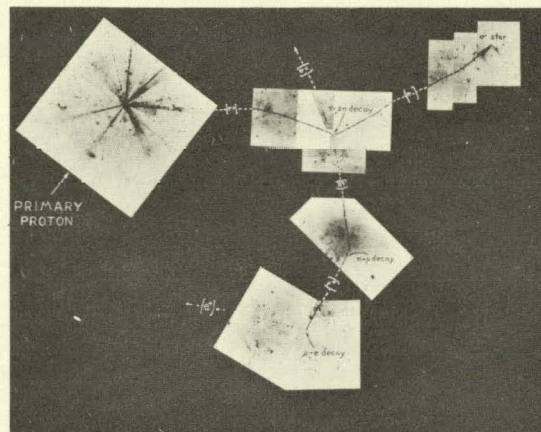


Fig. J-6. The τ decay mode of a K meson.

Figure J-7 is a mosaic of the K meson. Having a negative charge, it can be captured by a nucleus and will interact. Here the K comes in from above and interacts with a heavy nucleus. Several neutral particles may have been emitted from this point, but the fact that they lacked charge means that they could leave no observable tracks in the emulsion. A new particle called the sigma (Σ^-) continues downward through the emulsion, stops, and is captured by a nucleus. Here is produced yet another type of particle called a hyperfragment. It is essentially an excited nucleus resulting from the sigma interaction. A number of

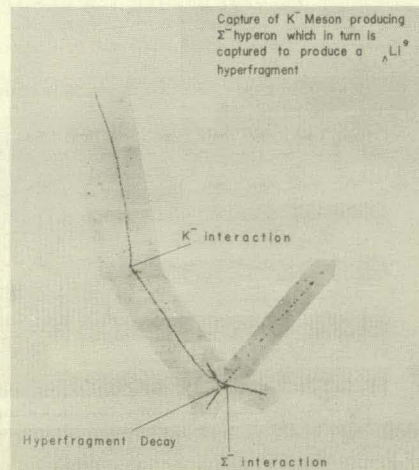


Fig. J-7. Interaction of a K meson with a heavy nucleus.

particles have thus been produced by the initial triggering of the K particle. These multiple events make the K's interesting objects of study. They are sources of numerous particles which we call hyperons and hyperfragments.

Figure J-8 is a diagram of one of the first antiproton interactions recorded in emulsions. The experiment resulting in the discovery of this particle was set up with counters as the detectors. The track left by the antiproton looks exactly like one left by a proton. Note that it bumped into a heavy nucleus and experienced a large-angle scatter, about 90°. At the point of its capture, it annihilated with a proton or neutron, producing several pions in the initial interaction. Secondary reactions of the pions with other nuclei of the nucleus are responsible for the observed charged particles such as alpha particles, protons, and deuterons, etc. The energy represented in the interaction is more than 1.8 Bev, confirming that the reaction was actually an annihilation process involving two particles of protonic mass.

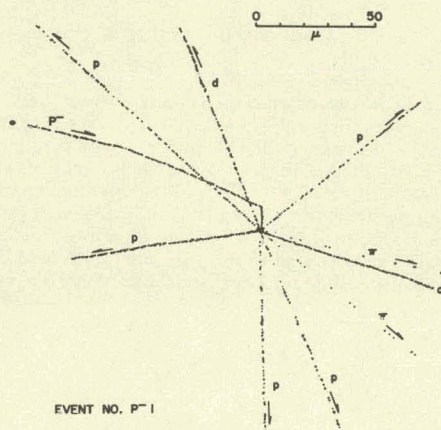


Fig. J-8. Interaction of antiproton at rest.

It has recently been possible to accelerate and record a number of heavy ions ranging in mass from carbon to argon. This is done with the heavy-ion linear accelerator (hilac) at this Laboratory. In Fig. J-9 are seen heavy-ion tracks as they appear at full machine energy (10 Mev per nucleon) in G.5 emulsions and moving from right to left. This emulsion is electron-sensitive, so that one sees the copious production of delta rays as a particle collides with many electrons along its path. Since the

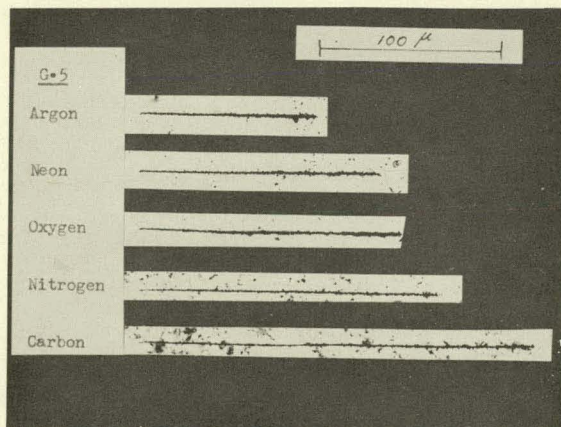


Fig. J-9. Heavy-ion tracks in Ilford G.5 emulsion.

particle is multicharged its profile shows a great number of scattered electrons; the track narrows toward its end.

Emulsion Techniques

The object of this picture-taking of course is to identify particles and analyze events in this microscopic world. When the scanner looks through the microscope, a number of things are done. First of all, the distance a particle moves into the emulsion is measured. Also, angles are measured. These angles may be either the change in direction of a particle after a scatter, or the relative angle between two outgoing particles from a star or interaction. The angles between particles resulting from decays may also be important. Furthermore, individual silver grains can be seen through these microscopes. A count of these grains can give a measure of ionization and energy lost by the moving charged particle. These techniques, which are elemental, as well as others that are composites of them, must be mastered by the emulsion physicist.

Range Measurements

The stage of the scanning microscope is essentially a three-dimensional micrometer. This means that it is capable of making measurements along the X, Y, and Z axes. As an example, the track shown in Fig. J-10 might be measured at the points indicated by dotted vertical lines. The object here is to measure the range of the track between the interaction at $X_0 Y_0 Z_0$ and its terminus at $X_4 Y_4 Z_4$. Let us assume that a π decays into a μ at $X_0 Y_0 Z_0$, so that the measured range will be that of the μ . The total range of the μ is called R, which is given by

$$R = r_1 + r_2 + r_3 + r_4, \quad (1)$$

where r_1 is the range of the particle in the first pellicle, r_2 is the range in the second, etc. Each of the r ranges approximates a straight line. The track, it must be kept in mind, does not lie in the same plane throughout the emulsion stack (in Fig. J-10 we see a 2-dimensional projection of the true track). The angle at which the track dips must therefore be considered in the range measurement or the result will be grossly inaccurate. Each segment r must be measured as an individual and its true range determined by the Pythagorean theorem as follows:

$$r_1^2 = (X_1 - X_0)^2 + (Y_1 - Y_0)^2 + (Z_1 - Z_0)^2,$$

$$r_2^2 = (X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2,$$

etc.

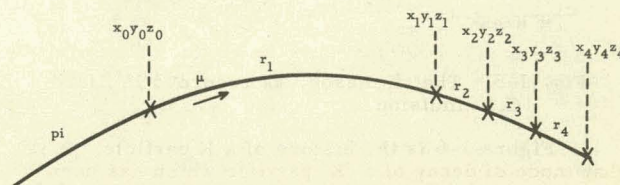


Fig. J-10. Two-dimensional projection of a particle track in emulsion.

Measurements of Relative Masses

When a charged particle travels through matter, such as lead, tin, or emulsion, it continually loses energy and eventually comes to rest. As has already been pointed out, the mechanism by which the particle loses energy is called ionization. This ionization and excitation of atoms in the stopping medium is due to their interaction with the electromagnetic field associated with the moving charged particle. The change in energy ΔE that the incident particle undergoes in passing through a small distance in matter depends upon the density of the stopping material, the velocity of the particle, and the charge. The ionization process does not depend upon the mass of the moving particle. In very general terms, the equation may be written

$$\frac{\Delta E}{\Delta R} = Z^2 F(v), \quad (2)$$

where

$$\frac{\Delta E}{\Delta R} = \frac{\text{the change in energy}}{\text{the change in range}},$$

Z = the charge on the moving particle,

and

$F(v)$ is a function whose value depends only on the velocity of the particle.

Let the particles under consideration be of unit charge (such as protons, electrons, or mesons). Then Eq. (2) becomes

$$\frac{\Delta E}{\Delta R} = F(v).$$

This says that two particles of unit charge at the same velocity have the same rates of energy loss regardless of their difference in mass. From this result a very useful relationship between the masses of two particles, M_A and M_B , and their ranges, R_A and R_B , can be derived when their velocities are equal. The interesting result is that if two particles enter an absorber at equal velocities, their ranges will be proportional to their respective masses:

$$\frac{R_A}{M_A} = \frac{R_B}{M_B}, \quad (3)$$

or

$$\frac{R_A}{R_B} = \frac{M_A}{M_B}. \quad (4)$$

When the scanner observes tracks of a charged particle in a nuclear emulsion, the number of silver grains that are seen in a small distance is a measure of the rate of energy loss. It follows that if two singly charged particles produce the same number of grains in a given segment of track length (i. e., the tracks are of equal grain density) the rates of energy loss are the same. Hence, from the equation $\Delta E/\Delta R = Z^2 F(v)$, the velocities are the same.

One is now led to a method of determining the relative masses of particles that come to rest in an emulsion. Consider two tracks A and B that come to rest along a line which will be called L (Fig. J-11). Measurements are made, or counts

taken, of grain density along each track. It is found that these are identical at points 1 and 2, at ranges R_A and R_B respectively. This means that the velocities are the same for the two particles at points 1 and 2. From the equations already given, it is seen that the ratio of the ranges R_A/R_B is equal to the ratio of the masses M_A/M_B .

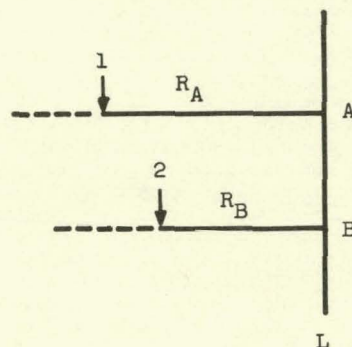


Fig. J-11. Tracks of two particles coming to rest at line L.

An example of this procedure is as follows: A pion decays into a muon according to the reaction $\pi \rightarrow \mu + \gamma$. The range of the muon from this decay is always about 0.6 mm in emulsion. The mass of the pion is known to be 273 electron masses (Me). It is observed by the scanner that the grain density of the muon at the point of decay is the same as the grain density of the pion at a point 0.793 mm from the end of its range. The mass of the muon relative to that of the pion is therefore

$$\frac{M_\mu}{M_\pi} = \frac{0.600}{0.793} = 0.757,$$

and the absolute mass of the muon is

$$M_\mu = 0.757 M_\pi = 0.757 \times 273 = 207 \text{ Me.}$$

Identification by Scattering Measurements

The identification of particles by means of the method outlined above assumes that they come to rest in the emulsion. The velocity, on the other hand, may be such that an incident particle does not stop but passes on through the stack. In this event a different technique must be used. As the particle moves through the emulsion, there will be numerous collisions with electrons along its path. The track therefore does not proceed in a true straight line. The sum total of these small deflections is called multiple scattering. The more massive the particle and the greater its velocity, the less will be its scattering. It should be possible therefore, to use this property as an aid in identifying a particle. A quantitative measure of multiple scattering may be accomplished by first dividing the track into 100- μ units called cells. The inclination of the tangent to the track is measured at each division.

The physical quantity that we actually determine when we measure the amount of multiple scattering a particle undergoes in emulsion is its pv , that is, momentum times velocity. The expression of pv in term of a , the mean scattering angle that is determined by measurements on the track, is given by

$$\alpha^0 = \frac{K}{pv} \left(\frac{S}{100} \right)^{1/2}, \quad (5)$$

where

α^0 = angle in degrees,
 K is a constant (about 25.0),
 p = momentum,
 v = velocity, and
 S = cell length in microns.

In general, we can write $p = m \gamma v$, where

$$\gamma = \frac{1}{\sqrt{1 - (v/c)^2}}$$

A measure of this gives the observer an estimate of

$$m \frac{v^2}{\sqrt{1 - (v/c)^2}},$$

which is the mass times a function of velocity. If we next determine the velocity by measuring the grain density of the particle, we are able to deduce the mass of the particle.

An improved technique for measuring the angles of scatter was suggested by Fowler of the Bristol group. In the method briefly outlined above, angles are measured by a goniometer on the eyepiece of a microscope. Fowler suggested the use of a stage that would give a straight-line reference for the angle measurements. The track is aligned with the X-axis of the stage. As the track is then moved along the X-axis, it appears to move up and down in the "Y" direction owing to its multiple scattering. The observer can then measure the distances of scattering from a straight line. The track might appear as shown in Fig. J-12 when aligned under the scope.

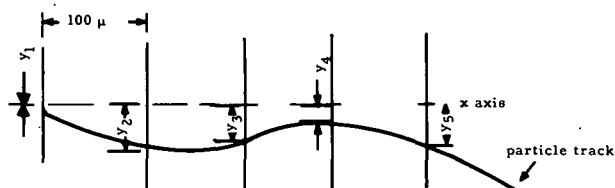


Fig. J-12. Particle track in emulsion, as aligned with the X-axis. The track is divided into 100-μ cells.

The track is divided into 100-μ cells, as described above. The top horizontal line is the reference from which the deflection of the track is measured. Sample measurements of a track made on the above hypothetical track are given in Table J-I.

Table J-I

Sample track measurements			
	Y	ΔY	$\Delta^2 Y = \Delta Y_2 - \Delta Y_1$
Y_1	10	$= Y_{n+1} - Y_n$	
Y_2	12	+2	
Y_3	11	-1	-3
Y_4	13	+2	+3
Y_5	14	+1	-1
			$\frac{7}{3} = 2.33$

The term Y is a measure of the deviation of the track from the reference line; ΔY is the first difference obtained by subtracting Y_2 from Y_1 , Y_3 from Y_2 , etc; ΔY is proportional to the tangent of the line segment between adjacent Y readings; $\Delta^2 Y$ is the second difference, obtained by subtracting the ΔY 's in their regular order. In Fig. J-12, we find $\Delta Y_1 = Y_2 - Y_1$, and

$\Delta Y_2 = Y_3 - Y_2$; $\Delta^2 Y_1 = \Delta Y_2 - \Delta Y_1$. The term $\Delta^2 Y$ then is a measure of the change in the direction of the tangent of the line segment in going from cell to cell. In Table J-I, the Y columns are the Y constants of the track: $\Delta Y = Y_2 - Y_1$, etc., and $\Delta^2 Y = \Delta Y_2 - \Delta Y_1$, etc. Second differences are summed disregarding their plus and minus signs. The average of $\Delta^2 Y$ (when converted to microns) divided by the cell length S (100μ in the example) gives the average change in the tangent (which we call α) of the line segment. Again, we may obtain the particle's pv from Eq. (5):

$$pv = \frac{C}{\alpha} (S/100)^{1/2},$$

where C is a constant, though different from K in Eq. (5).

Summary

Some of the fundamentals of emulsion techniques, therefore, are the counting of grains, the measuring of ranges, and the measuring of angles. Skill comes from a familiarity with particle tracks and from combinations and modifications of these fundamental procedures.

Bibliography

1. A Beiser, *Nuclear Emulsion Techniques*, *Reviews of Modern Physics* **24**, 273-311 (1952).
2. Y. Goldschmidt-Clermont, *Photographic Emulsions*, in *Annual Review of Nuclear Science*, J. G. Beckerley et al., Eds., Vol. 3 (Annual Reviews, Stanford, California, 1953), pp. 141-170.
3. *Nuclear Track Emulsion*, in *Fundamental Mechanisms of Photographic Sensitivity*, *Proceedings of a Symposium Held at the University of Bristol in March 1950* (Butterworths, London, 1951), pp. 259-345.
4. C. F. Powell, *The Study of Elementary Particles by the Photographic Emulsion Method* (Pergamon, New York, 1959).
5. M. M. Shapiro, *Nuclear Emulsions*, in *Handbuch der Physik*, S. Flugge, Ed., Vol. 45 (Springer-Verlag, Berlin, 1958), pp. 342-346.
6. C. Waller, *Nuclear Emulsion Techniques*; *A Review of Progress 1948-1953*, *Journal of Photographic Sciences (British)* **1**, 41-54 (1953).

K. COUNTERS USED IN HIGH-ENERGY PHYSICS

LeRoy T. Kerth

Dr. Wallace gave a general survey of the various types of counters and what they are used for. We will not repeat that material but instead will dwell mostly on the kind of counting experiments that we do in high-energy physics. This will include the philosophy behind the counting experiment, how it is set up, and what it is intended to do.

Before our discussion can proceed very far, we must explain the jargon that has developed in connection with counters. It is very familiar to the physicist, but the words are quite frequently used differently than they are in other professions. A glossary of terms is presented in Table K-I.

Table K-I

Glossary of terms	
Term	Definition
Pulse	A voltage change
Count	When a particle produces a pulse
Coincidence	When pulses arrive at same time
Coincidence circuit	A device that detects coincidences
Anticoincidence circuit	A device that stops coincidences
Delay	Time delay of a pulse
Scaler	A device that counts pulses

We will first explain them and then proceed to discuss the experiments themselves.

The first term is a pulse. Dr. Wallace explained that a counter is a device which produces an electrical pulse when a charged particle goes through its sensitive volume. Now, what exactly is an electrical pulse? It is a change in voltage at some instant. For example, if you quickly switch a light on and off, an electrical pulse occurs in the wires. That is, the voltage went on and then it went off. Suppose now that a pulse is produced at some point along a conductor and that we connect to that conductor what is called a cable or transmission line. This pulse or electrical disturbance then travels down the cable at the velocity that light would travel through the insulation of the cable. Therefore, it takes a definite amount of time for a pulse to travel from one end of a cable to the other. This is an important concept.

As stated above, a counter is a device which, when a charged particle passes through it, somehow transforms this event into an electrical pulse. A scintillation counter, for instance, consists of a scintillating material and a photomultiplier tube. (See Fig. 1-5 in Section I "Survey of Particle Detectors.") A charged particle passing through the

scintillator produces a flash of light. The photomultiplier tube then transforms this signal into an electric pulse. The associated circuitry is not considered a part of the counter itself.

A coincidence refers to the following: Suppose that we have two cables, each bringing separate pulses into a box (see Fig. K-1).

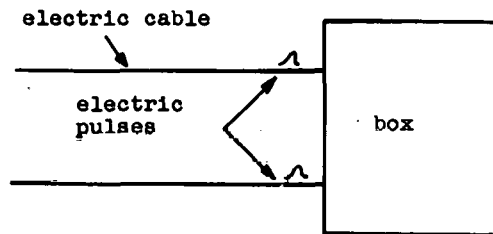


Fig. K-1. A coincidence occurs if the two pulses reach the box simultaneously.

By a coincidence we mean that two pulses, one from each line, arrive at the box simultaneously. This is a coincidence of those two pulses. We could also talk about a coincidence of three pulses, four pulses, five pulses, or of as many as we care to.

A coincidence circuit is the device, or box, into which these cables lead. This device tells us when a coincidence has occurred. If two pulses arrive at the two input plugs at the same time, the coincidence circuit will emit an output pulse (see Fig. K-2).

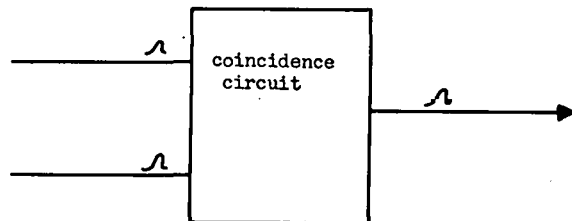


Fig. K-2. A coincidence circuit emits an output pulse when two input pulses arrive simultaneously.

Whenever we see an output pulse we know that two pulses entered at the same time.

An anticoincidence circuit would be another part of the same box as the coincidence circuit. It may be used when there are three (or more) input cables (Fig. K-3), which we will label as C_1 , C_2 , and C_3 .

If input pulses arrive simultaneously on C_1 and C_2 , the coincidence circuit emits an output pulse. However, if pulses arrive simultaneously on C_1 , C_2 , and C_3 , an anticoincidence circuit blocks the output pulse from being emitted. This allows us to do the following logic: If we get an output pulse we know that input pulses on C_1 and C_2 arrived together, but that no pulse arrived on C_3 at that instant. If we get no output pulse, two situations can exist:

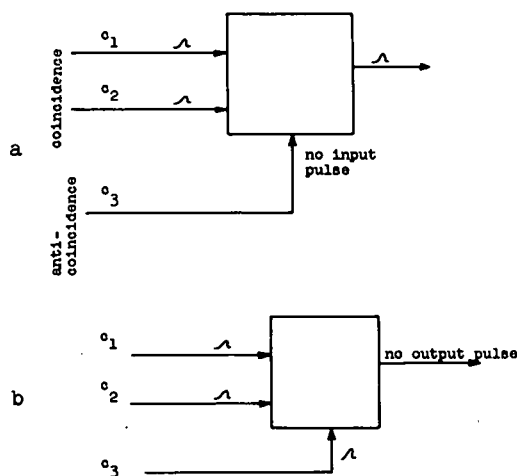


Fig. K-3. An anticoincidence circuit.

(1) input pulses arrived simultaneously on C_1 , C_2 , and C_3 ; or (2) the input pulses on C_1 and C_2 arrived at different times. This technique is very important in many experiments.

Delay time. It will be recalled that a pulse takes a definite length of time to travel from one end of a cable to the other. Suppose that we have an experiment such as that shown in Fig. K-4.

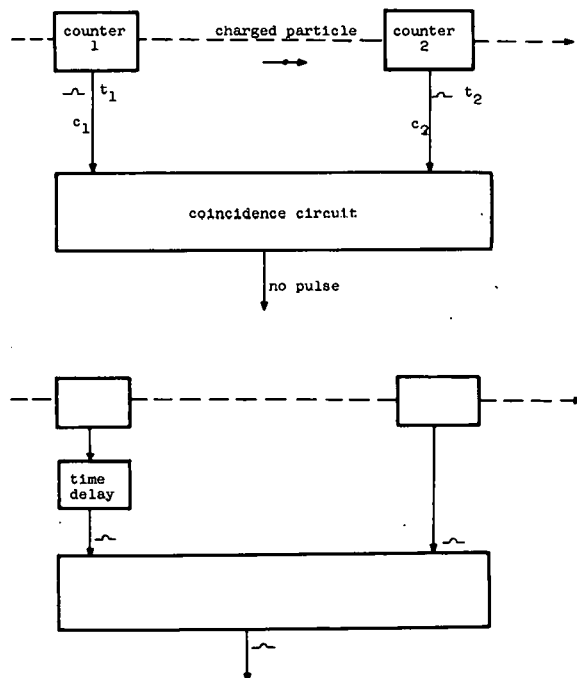


Fig. K-4. Illustration of how a time delay is used.

A charged particle passes through two counters. Because a pulse occurred at counter 1 earlier, it will arrive at the coincidence circuit first, at time t_1 . A certain time later, at t_2 , a pulse arrives from counter 2. Therefore, no output pulse occurs. Assume that we want to know whether a particle of a given velocity passed through both counters. We could do this by inserting a time delay in cable C_1 corresponding to the time it would take the particle to travel from counter 1 to counter 2. If the particle

then passed through the two counters, input pulses would arrive simultaneously at the coincidence circuit, causing an output pulse. This example shows how we can use a coincidence circuit to detect events that occur at two different places at different times. In practice, a time delay consists simply of inserting a certain additional length of cable in one transmission line, since the speed with which a pulse travels down a cable is a well determined number.

A scaler is nothing but an electronic circuit that counts pulses. Thus, a scaler actually counts pulses, whereas a counter produces pulses. This terminology may seem confusing, but nevertheless it is part of the jargon that has grown up among physicists. Physically, a scaler is a box which has dials on the front. After it is turned on it counts the number of electric pulses that arrive until it is again turned off. Thus, a scaler is connected to the output of a coincidence or anticoincidence circuit as a recording device.

The basic elements that are used in a counting system are represented in Fig. K-5.

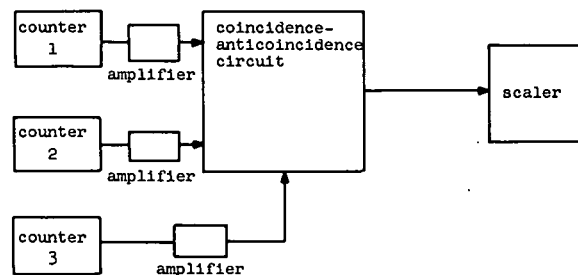


Fig. K-5. A simple counting system.

One device which has not been mentioned so far is the amplifier. If a pulse is too small an amplifier is used to make it bigger. If one were to walk around and look at some of the counting experiments that are done, most of the boxes he would see are amplifiers, but the ones that are doing the logic and actually solving the problem are coincidence circuits and scalars.

The word "logic" has been used a number of times in this lecture. This is perhaps indicative of one of the main advantages of counting equipment. That is, that we can devise electronic equipment to do most or all of the logic that is involved in a physics experiment. For example, one can set up counting devices which will tell that a certain combination of events occurred but that a different combination did not.

Dr. Wallace described the various counters—the ion chamber, the proportional counter, the Geiger counter, the scintillation counter, and the Cherenkov counter. The only ones that we will discuss are scintillation and Cherenkov counters, because these are the ones that are almost universally used in high-energy physics. The other counters have uses in low-energy physics, in radiochemical techniques, and certainly the Geiger counter is a standby for all uranium prospectors.

The counters that we will talk about—the scintillation and Cherenkov counters—produce an electric signal when a charged particle passes through their sensitive volume. In both cases the electric signal comes from a photomultiplier tube when it observes a flash of light. However, the

method for producing this light is different in each type of counter.

Plastic scintillators are by far the most popular at this Laboratory and in most high-energy physics laboratories because they are easily fabricated into all sorts of shapes and sizes; also plastic is quite stable. Some of the other scintillators are not stable in air; being hygroscopic, they pick up water and dissolve. Moreover, it is difficult to grow large crystal scintillators. But this plastic is nothing more than polystyrene which can be machined by ordinary techniques. A small amount of organic material called terphenyl is dissolved in it before it is polymerized.

It is sometimes necessary to use a light pipe to carry the light from the scintillator into the phototube. Lucite is commonly used for this.

The procedure in assembling a plastic scintillation counter is first to fasten the light pipe to the plastic scintillator with a transparent cement. Aluminum foil is then wrapped around the scintillator and light pipe except where the light pipe contacts the phototube. This foil is to keep room light from reaching the phototube. The phototube is next joined to the light pipe. The whole assembly is wrapped with black tape to make it lighttight.

The completed counter is then placed inside an iron tube (Fig. K-6), which provides magnetic shielding. This shielding is necessary because the electrons traveling through the phototube are very sensitive to magnetic fields.

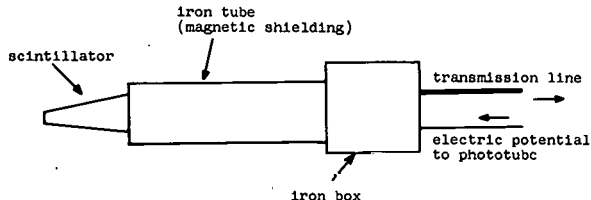


Fig. K-6. A scintillation counter assembled inside an iron housing, which provides magnetic shielding.

The electrons could easily be bent away from the dynodes. The box that surrounds the end of the phototube has connectors for the transmission line (the output signal) and for the potential that must be supplied to the dynodes.

The amount of light that is given off when a particle passes through the scintillator depends on the amount of energy lost in the counter. The energy lost by a particle in passing through a counter is proportional to $1/v^2$, where v is the velocity of the particle. Therefore, the pulse size is an approximate indication of particle velocity. The more light that is produced, the greater will be the number of electrons cascading down the phototube. Consequently the voltage pulse emitted by the counter will be larger. Pulse-height analysis is a technique which is quite frequently used to indicate the velocity of a particle. For various reasons this is not always a very good technique, but it gives a sufficient indication for many purposes. However, there is a velocity-sensitive counter which has come into use in the last five years in high-energy physics. This is the Cherenkov counter. Cherenkov, a Russian, very recently received the Nobel Prize for discovering the effect on which the counter is based.

This effect is explained as follows: Suppose that a charged particle passes through a transparent medium such as glass, Lucite, or water. Its velocity can be referred to as βc , where β is the particle velocity divided by the velocity of light.

The velocity of light inside a transparent medium is the velocity of light in vacuum divided by the index of refraction, η . Thus the velocity of light inside of the transparent medium is equal to c/η . The Cherenkov effect occurs when the particle velocity βc is greater than the velocity of light in this medium, that is, when $\beta c > c/\eta$. You may object to this statement, thinking that nothing can go faster than the velocity of light. However, note that we stated "the velocity of light inside of this medium," because light goes slower in a transparent medium than it does in vacuum. It is the velocity of light in vacuum that cannot be exceeded.

When a particle exceeds the velocity of light inside of a transparent material, what happens? It turns out that a cone of light is given off behind the particle similar to the bow wave produced by a speedboat. (See Fig. A-14 in Section A-2 "Trigonometry.") Suppose that the particle is now at point P and is traveling with a velocity v . At some earlier time it was at point P'. The light emitted at point P' travels outward with a circular wave front. By the time the particle has reached point P this circle has grown to the size indicated in Fig. A-14. However, the light emitted when it was at point P' has not had time to travel as far. Therefore, a smaller circular wave surrounds point P'. Similarly, the waves from other points between P' and P establish a conical wave front. We can draw a straight line from point P tangent to these circles. This wavefront, or shock wave, is visible as light, which we call Cherenkov radiation.

We can readily calculate the angle α between the direction of motion of the coherent wavefront and the direction of particle motion. (See Fig. A-15). The cosine of α will be

$$\cos \alpha = \frac{c/\eta}{\beta c},$$

which reduces to

$$\cos \alpha = 1/\eta\beta.$$

Thus, we see that this angle α is inversely proportional to the velocity of the particle.

This fact has been used in many novel and different ways to determine the velocity of particles passing through Cherenkov counters. One of the most useful implications is this: if the velocity of the particle is lower than the velocity of light inside the medium, no Cherenkov radiation is produced. That is, for Cherenkov radiation to be emitted, the particle must be traveling faster than a certain critical velocity. This critical velocity is the case where the angle α is zero, i.e., where $\cos \alpha = 1$. For this condition, the relationship $\beta c = 1/\eta$ obtains. Therefore, as long as the β of the particle is less than $1/\eta$, particles passing through the counter will not be recorded.

Cherenkov counters use a photomultiplier tube to observe the light that is given off. If the photomultiplier were located in such a position that it would detect light only at a given angle, then it would detect

particles of one only velocity. In practice, this must be a small angular interval, so that the detector counts particles of a certain velocity interval. These are called velocity-interval counters, of which a number of different types have been made. They differ only in the optics that transmit light from the sensitive cell to the photomultiplier. A Cherenkov counter of this type was used in the original antiproton experiments at this Laboratory in 1955 (Fig. K-7).

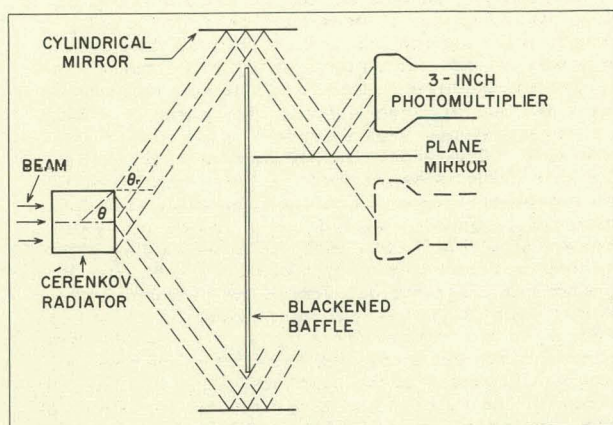


Fig. K-7. A velocity-selecting Cherenkov counter used in the antiproton experiment.

A beam of particles from the Bevatron enters a Cherenkov radiator, which is a block of glass. A baffle made of an opaque material prevents light from going straight through and striking any of the three photomultipliers that are arranged symmetrically about the axis. However, that light which is emitted at a certain angle strikes a cylindrical mirror. This reflected light then strikes three plane mirrors and is reflected into the photomultiplier tubes. In this way, only that light emitted in a certain small angle will enter the photomultiplier. Thus, only particles of a given velocity interval are counted. We then take these pulses and correlate them with other pulses coming from other counters; this allows us to interpret certain pulses as being special events arising from particles of a given velocity interval.

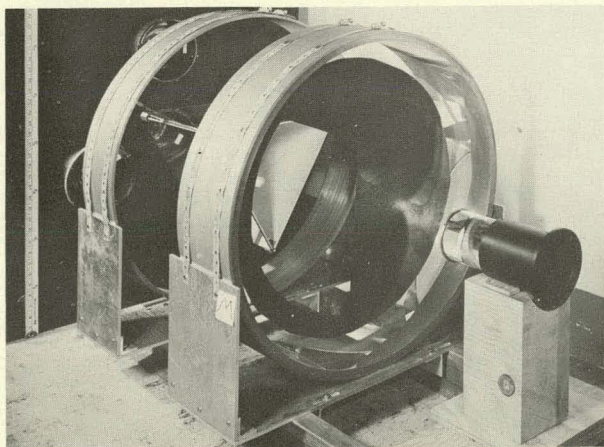


Fig. K-8. Photograph of the Cherenkov counter represented in Fig. K-7.

Figure K-8 is a photograph of the Cherenkov counter that was used in the antiproton experiments. We can see the Cherenkov radiator and the cylindrical mirror. The baffle has been removed so that we can see through. The plain mirrors and the photomultipliers are visible at the rear. This is quite a large device, being about three feet in diameter.

Thus, we have seen that the Cherenkov counter can count particular velocity intervals or count particles having a velocity greater than a certain value. Scintillation counters, on the other hand, count all charged particles; the pulse height to some extent indicates the velocity of the particles.

A very famous counting experiment was one that was performed around the turn of the century by Lord Rutherford (Fig. K-9).

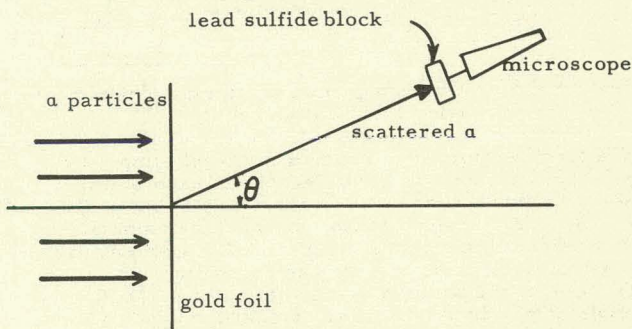


Fig. K-9. Lord Rutherford's counting experiment.

He allowed a beam of alpha particles from a radioactive source to strike a very thin gold foil. Then at a certain angle θ he placed a piece of lead sulfide. The alpha particles in passing through the gold foil were scattered. Those that hit this lead sulfide would give a small flash of light. Rutherford observed these flashes with a microscope, counting each one that he saw. By varying the angle θ he was able to measure the scattering distribution of alpha particles from gold. From this he concluded that atoms are made up of a very small central core with a rather diffuse cloud of electrons around it. In other words, an atom is made up mainly of empty space.

Rutherford's experiment is very typical of experiments in nuclear physics. That is, we send a beam of particles through a target; they do something, and by observing what they do we try to infer what the force is and to determine the shapes and sizes of the particles that were interacting. In high-energy physics we are trying to discover the properties of the particles involved in nuclei and in the particles which are produced in high-energy collisions.

Now, we will look at a rather simple experiment that we could do with counters. Suppose we want to measure the forces that exist between two protons. We use a target of protons, i.e., a liquid hydrogen target; we then send in a beam of protons, say from the cyclotron (Fig. K-10).

We know that when two particles of the same mass collide they bounce off at 90 degrees to each other. For example, two billiards balls will always do this, providing they are not given any "English." It may happen that one goes straight forward and the other just dribbles off sideways with practically no energy, but still the angle is 90 degrees between the two particles. Thus, we can tell when two protons

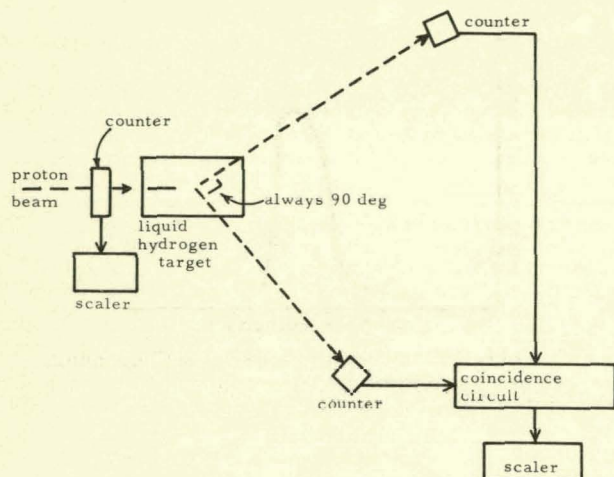


Fig. K-10. A proton-proton scattering experiment using scintillation counters.

scatter from this liquid hydrogen target simply by placing two counters at 90 degrees to each other. We then connect these with cables to a coincidence circuit and a scaler. Also, we may use some sort of a device in the proton beam to tell how many protons came in; from this we can determine the fraction of the protons that scatter at a particular angle. We can rotate these counters around to some other angle and do it again. It was this kind of a simple experiment that started out counter experiments in high-energy physics. This is a rather simple example of how we use the logic of counters to pick particular events that we want to measure and then count how many times they happen out of a certain number of tries.

Now, let us look at another possibility (Fig. K-11). Suppose we have a pi-meson beam. We set up

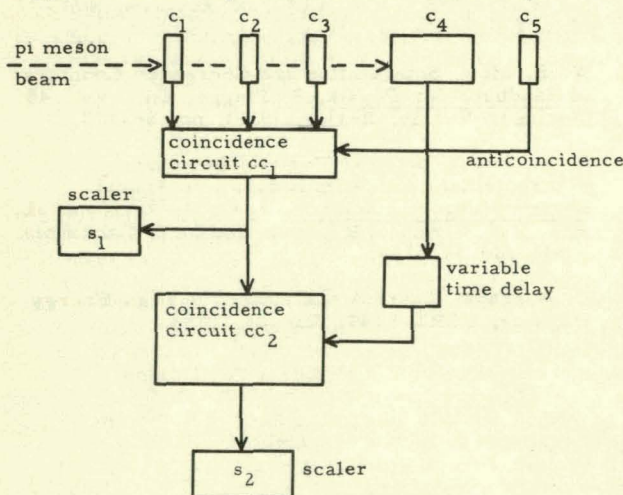


Fig. K-11. A counter experiment for determining the half life of the pi meson.

several counters (c_1 , c_2 , and c_3) which say, "Yes, a particle came along here," because we connect them to a coincidence circuit (CC_1). We then put in another counter (c_4)—a rather big one—and adjust the energies of the incoming pi mesons so they will stop in this large counter. How do we know that they stop? We put in another counter (c_5), which we

connect in anticoincidence. Coincidence circuit (CC_1) will then give an output pulse when counters c_1 , c_2 , and c_3 count and c_4 does not count. This means that a particle entered c_4 but did not come out. A scaler (S_1) records the number of particles that stop in c_4 . We also connect c_4 into a coincidence circuit (CC_2), inserting a variable time delay in the line. We set this delay so that the pulse that corresponds to a pi meson coming to rest in c_4 arrives at the second coincidence circuit (CC_2) at the same time as the pulse from the first coincidence circuit (CC_1). At this setting we get a pulse at S_2 . Now, what happens if we set the delay for a shorter time? If we set the delay for a shorter time, the only way we can get a signal at S_2 is if a pulse occurs in counter c_4 after the pi meson came in. We know that pi mesons decay radioactively. When a pi meson stops, it gives rise to a mu meson after a time Δt . Therefore, we get a pulse at S_2 if the variable delay is set so that the pulse caused by the mu meson arrives at CC_2 at the same time as the pulse from CC_1 . Thus, scaler S_1 tells us how many pi mesons stop in c_4 . Scaler S_2 tells us how many mu mesons were created after a certain time, Δt . From this we can learn the half life of the pi meson. To do this we plot the ratio of the number of counts from S_2 to the number from S_1 for a given Δt . When we do this we get a curve which corresponds to the exponential decay of the pi meson (Fig. K-12). This is how the half life of the pi meson was measured by Drs. Chamberlain and Wiegand of this Laboratory.

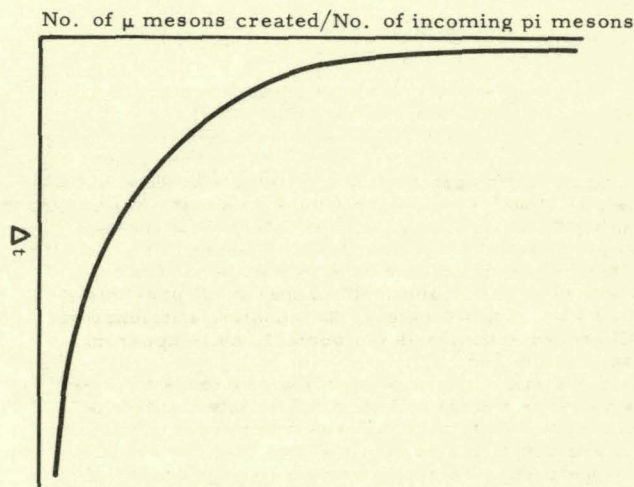


Fig. K-12. Exponential decay of the pi meson.

Another way of using variable delays and time intervals is shown in Fig. K-13.

We flip a metal target into the proton beam of the Bevatron to produce a beam of secondary particles. Some of these particles are then led out of the magnetic field of the Bevatron. Those of a given momentum pass through a focusing magnet and then proceed through two counters (c_1 and c_2) that are, say, 100 feet apart. Suppose for simplicity that the beam traversing the two counters consists of only two kinds of particles, say protons and pi mesons. The proton is six times the mass of the pi meson; since they have the same momentum they must have different velocities. Because they have different velocities it will take a different time for them to go from counter c_1 to c_2 . We will then connect these counters to a coincidence circuit, but insert a time delay in the cable from c_1 . This delay will correspond exactly to the time that it takes, say,

a pi meson to go from c_1 to c_2 . In this way, pulses from c_1 and c_2 will arrive simultaneously only when a pi meson traverses the two counters. The scaler will then count the pi mesons only.

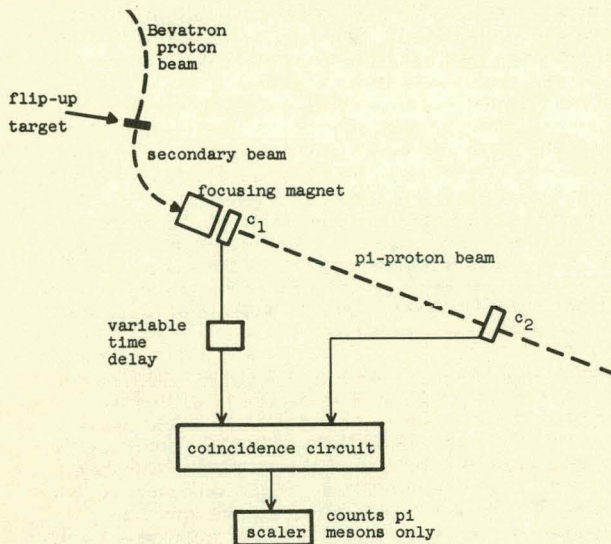


Fig. K-13. An experiment for counting particles of a given type using the time-of-flight technique.

About the best time resolution we can get (that is, the smallest time difference we can measure) is of the order of 10^{-9} second in a time-of-flight experiment such as that just described. As we go to higher energies, the particles go farther in a given time, so that their velocities cannot be measured as accurately. However, this problem can be helped somewhat by using a Cherenkov counter in conjunction with scintillation counters, as was done in the antiproton experiment previously alluded to. Unfortunately, the counting efficiency of a Cherenkov counter is not perfect, as is apparent from Fig. K-14.

We see that the curve of counting efficiency vs velocity has a little tail on it. For some reason a Cherenkov counter will always count some particles of a slightly different velocity than that for which it was designed. This may amount to only a tenth of a percent or a hundredth of a percent. But if we have 100,000 pi mesons for every antiproton and are looking for antiprotons, then we have to worry about these tails.

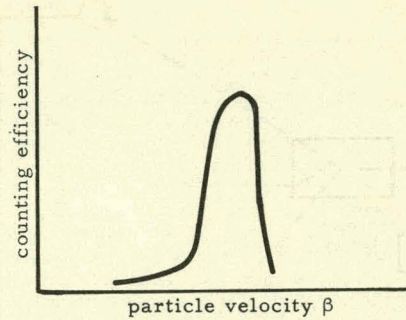


Fig. K-14. Counting efficiency of a Cherenkov counter.

Bibliography

1. R. E. Bell, Nuclear Particle Detectors (Fast Electronics), in Annual Review of Nuclear Science, J. G. Beckerley et al., Eds., Vol. 4 (Annual Reviews, Stanford, California, 1954), pp. 93-110.
2. S. de Benedetti, The Coincidence Method, in Handbuch der Physik, S. Flugge, Ed., Vol. 45 (Springer-Verlag, Berlin, 1958), pp. 222-259.
3. G. B. Collins, Scintillation Counters, Scientific American, 36 ff. (Nov. 1953).
4. S. A. Korff, Counters, Scientific American, 40 ff. (July 1950).
5. S. A. Korff, Electron and Nuclear Counters, Second Edition (Van Nostrand, Princeton, 1955).
6. J. Marshall, Nuclear Particle Detectors (Characteristics of Cherenkov Counters), in Annual Review of Nuclear Science, J. G. Beckerley et al., Eds., Vol. 4 (Annual Reviews, Stanford, California, 1954), pp. 141-156.
7. W. E. Mott, Scintillation and Cherenkov Counters, in Handbuch der Physik, S. Flugge, Ed., Vol. 45 (Springer-Verlag, Berlin, 1958), pp. 86-173.
8. R. K. Swank, Nuclear Particle Detectors (Characteristics of Scintillators), in Annual Review of Nuclear Science, J. G. Beckerley et al., Eds., Vol. 4 (Annual Reviews, Stanford, California, 1954), pp. 111-140.
9. C. Wiegand, Cherenkov Counters in High-Energy Physics, UCRL-8148, Jan. 21, 1958.

L. BUBBLE CHAMBERS

James D. Gow

Particle accelerators are probably the most publicized instruments used in high-energy physics today. One reads that Brookhaven is building a 25-Bev (billion electron volt) synchrotron or that the Russians already have a 10-Bev machine and are currently working on a 50-Bev version. Accelerators have all the glamor, but they make up only one half of a working team. The detectors are an equally important partner. An accelerator gives speed to atomic particles, but without the detectors with which to observe them, an accelerator would serve no useful purpose. If one looks back into the history of high-energy physics, it is noted that most of the fundamental discoveries were made not with artificially accelerated particles, but rather with cosmic rays.

One might then ask, "What is the advantage of the accelerator?" The answer is twofold. In the first place, particle beams can be produced giving much larger intensities than are available in cosmic radiation; and secondly, there is the advantage of separation and selection. In using cosmic rays one must take what comes - protons, mesons, hyperons, etc. With an accelerator the type of particle being accelerated is known, as well as the energy or momentum at which the particle enters an experimental system. For example, in the last year or two it has become practical to produce a beam of nearly pure π mesons, K mesons, or antiprotons.

Perhaps the next question that arises is, "What types of detectors are proving most useful in physics research?" To answer this, it is necessary to review briefly the two general classes of detectors. The first class is comprised of counters. Such devices are represented by the familiar Geiger and scintillation counters, which count the number of charged particles passing through their sensitive volume. The second class is made up of the track-forming detectors. Examples of these are the Wilson cloud chamber, the bubble chamber, and nuclear emulsions.

Aside from signaling the passage of a particle through its sensitive volume, a counter can give little more information. In some cases, as with a Cherenkov counter, it is possible to determine whether the particle has more than a certain minimum velocity. Counters generally have very poor spatial resolution, so that it is impractical to determine with high precision the direction a particle has taken. On the other hand, counters have one invaluable feature. This is their time resolution. With modern devices and techniques one can tell when a particle goes through a counter to an accuracy of better than 10^{-9} second. This time resolution has proven very important in unscrambling the particles that come from the Bevatron.

The track-forming detectors have advantages over the counters in many situations. With this type of detector one can visualize what happened to the particle in a nuclear reaction. Also it is often possible to observe an event that would be

completely missed by counters. The track-forming detectors have therefore been invaluable in the discovery of new particles. Counters have served to collect rapidly large amounts of statistically accurate data, but only rarely can a counter experiment be designed for finding a new particle. A classical exception to this was the discovery of the antiproton in 1955 at the Lawrence Radiation Laboratory.

The first and most familiar track-forming detector is the Wilson cloud chamber. It was developed by C.T.R. Wilson, of England, in 1912. A mixture of a vapor and a suitable gas, contained in a chamber, is suddenly cooled by expansion so that it becomes supersaturated. Vapor condenses around any charged particles that are present, leaving a track of tiny droplets which can be photographed and later examined.

A second track-forming detector is the nuclear emulsion. Nuclear emulsion is ordinary photographic film which is made very thick and loaded heavily with chemically sensitive material. Charged particles passing through an emulsion expose it in a manner similar to light exposing ordinary film. When the emulsion is developed, the particle path is seen as a dark track against a light background. These tracks are then examined under a compound microscope equipped with a precision measuring stage.

Unfortunately, both the cloud chamber and the nuclear emulsion have serious limitations. As particles are accelerated to higher energies, the probability of an interaction occurring decreases. Energy regions investigated prior to World War II involved cross sections of the order of one barn (10^{-24} cm²). In the energy region now being studied (1 Bev and up) the cross sections are measured in millibarns and sometimes in microbarns. For this reason, and also because of the fact that the density of the gas in a cloud chamber is low, only a small percentage of cloud chamber photographs contain interesting events. If, however, the effective density of the gas is increased by putting it under high pressure, the cycling rate falls off sharply. As much as 20 minutes may be required for the chamber to recover after an expansion. Thus, all that has been gained from greater density has been lost owing to the "dead time" between each expansion.

Emulsions are extremely good from the viewpoint of density, but they too have their shortcomings. In the cloud chamber one can measure particle momentum (a measure of particle energy) by bending the particle in a magnetic field as it traverses the chamber. The amount of curvature is inversely proportional to the momentum. The sign of the charge on the particle is given by the direction of curvature. In the emulsion it is impossible to determine the sign of the charge and difficult to determine momentum. Momentum measurement can be accomplished only by a tedious statistical process involving measuring the small-angle deviations of the particle path caused by scattering. Another problem with the emulsion is that it is a complicated chemical assembly, con-

taining silver, bromine, chlorine, oxygen, nitrogen, and hydrogen. This makes it difficult to identify the target nucleus. In a collision the nucleus may be torn to bits with as many as 15 protons flying off. One can tell that they are protons only by performing a very laborious energy-balance calculation. In spite of these limitations, the emulsion technique has produced very valuable results. Even so, this technique is tending toward obsolescence as newer devices and methods come into use.

It was against this background of difficulties that Donald Glaser of the University of Michigan invented in 1953 the bubble chamber. It now has become one of the detectors most used in high-energy nuclear physics. In a systematic search for a new detector, Dr. Glaser listed in advance the desirable features that it should have. Such a detector should be simple, have a high-density sensitive medium, have a fast cycling rate, and it must contain an amplification mechanism in order to render visible the extremely small energy associated with a single particle.

In the case of emulsions, the amplification results from a chemical reaction involving the developer. For counters, the amplification occurs in a photomultiplier tube and the associated electronics. In a cloud chamber, the amplification results from the thermodynamic instability of a super-saturated gas. The energy that nature provides to correct this instability is triggered by the passage of a charged particle.

A bubble chamber also takes advantage of a thermodynamic instability. Instead of a super-saturated gas, it utilizes a superheated liquid. The thermodynamics of this situation are illustrated in Fig. I-4 of the lecture entitled "Survey of Particle Detectors." Here we have plotted curves of pressure vs volume for various temperatures. These are generalized curves applicable to any material. The top curve ($T > T_K$) represents a substance that is above its critical temperature (the temperature at which a liquid and its saturated vapor coexist in equilibrium). As the volume increases, the pressure drops according to the familiar $PV = NRT$ rule. However, for a material which is at or below its critical temperature the $PV = NRT$ rule no longer applies. In this case, as the volume increases the pressure drops, so that the line AB is traced out. However, when the liquid reaches the state at B, the pressure no longer drops. Instead, the line BC is followed. The dashed portion (BFC) represents a superheated liquid, an unstable state. This superheated condition can be maintained for a brief interval, sometimes up to a minute, if the liquid is contained in a very clean smooth-walled vessel, and providing there are no other causes of disturbances. During this period a charged particle traversing the liquid causes boiling in its immediate vicinity. This produces visible tracks which can be photographed. What is responsible for this boiling is that the vapor pressure is greater than the actual pressure, and some of the liquid would like to turn into gas. It does so when triggered by the passage of a charged particle. When boiling occurs, the line BC results; this is the stable condition where the liquid and gas are in equilibrium.

The operating point for a bubble chamber is set by the pressure-volume curve for the particular type of liquid used. For hydrogen it is about 270°K (-443°F), the critical temperature being 33°K.

The first bubble chamber built by Glaser was 1 in. long and 1/2 in. in diameter. It used diethyl ether heated to 140°C and was pressurized to 20 atmospheres to prevent boiling. He obtained a superheated condition by suddenly lowering the pressure. The chamber was placed between two counters. When a charged particle traversed the counters and the chamber, an electric signal was produced which caused a photograph to be taken of the particle track.

Other liquids besides ether can be used in bubble chambers. The most common ones are propane, hydrogen, and deuterium. Hydrogen is an ideal substance. Since it contains only protons, there can be no ambiguity as to the identity of the struck nucleus. Then again, because the atomic number of hydrogen is low, it is possible to use the magnetic-deflection method for determining particle momentum and charge.

Deuterium, the heavy brother of hydrogen, is ideal for the study of neutrons. Where the energy of the incoming particle is high compared to the binding energy, deuterium may be considered a chemical compound - "neutronium hydride." If one performs an experiment in hydrogen and then repeats it in deuterium, the effect due to neutrons can be determined by subtracting the effect of hydrogen.

Building a liquid hydrogen bubble chamber presented many formidable engineering problems. For example, a glass-to-metal seal had to be made that would remain tight when the apparatus is cooled to 27°K. The differential shrinkage of the materials would cause the contacting surfaces to slide, tending to open the seal.

Another problem involved illumination. Liquid hydrogen has a very low index of refraction (1.08). A bubble of its vapor in liquid hydrogen scatters light only slightly, about 6°; for this reason, a bubble in hydrogen is almost invisible when viewed from the side. One method used in earlier chambers for solving this problem was to illuminate the bubbles from the side opposite the camera, installing a set of vanes to keep the direct light out of the lens. A much better system was devised later for the 15-inch bubble chamber. (This is discussed below.)

Another problem in building bubble chambers of an appreciable size is that of the spontaneous boiling which occurs at any discontinuity in the chamber. Such a discontinuity might be the point where a glass wall joins a metal one. In the tiny all-glass chambers, liquids could be kept sensitive for about one minute or until a cosmic ray triggered boiling. However, larger chambers could not be made entirely of glass, owing to strength considerations. Luckily, it was found that by expanding the chamber very quickly and taking a photograph within a few milliseconds, the spontaneous boiling induced at surface imperfections does not generate enough vapor to repressurize the chamber and thus destroy sensitivity.

The first chamber built at Berkeley was a 1-in.-diameter glass chamber. This device demonstrated that hydrogen is a suitable medium for track formation. Later, a 2-in. chamber was built of glass and metal. It showed that "dirty" chambers were feasible. The term "dirty" indicates a chamber having some surface irregularities. This is a convenient if not an esthetic term.

The first chamber built at Berkeley as a research instrument was the 4-in. liquid hydrogen bubble chamber (Fig. L-1). It is suspended inside a tank which is evacuated to provide thermal insulation. In between the chamber and the vacuum vessel is a copper shield cooled by liquid nitrogen (at 77°K). This acts as a heat barrier to keep down the thermal radiation to the chamber. The temperature must be regulated to maintain chamber sensitivity. This is done by connecting the chamber to a flask of liquid hydrogen through a heat leak. Slightly too much refrigeration is supplied, and the difference is compensated for by using a heater. In the past four years nearly a million and a half pictures have been taken using this chamber. Currently it is being operated in connection with the 300-Mev electron synchrotron, but it is not large enough to be useful at Bevatron energies.

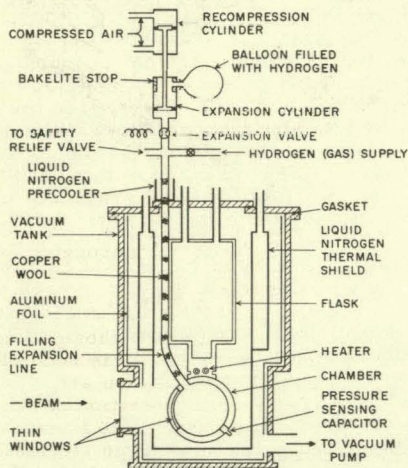


Fig. L-1. The 4-inch liquid hydrogen bubble chamber.

As larger bubble chambers are built, the refrigeration demands can increase inordinately if the 4-in. system of expansion and recompression is used. The only way the pressure can be reduced suddenly is to increase the volume rapidly by moving a piston or opening a valve. If we wait until the liquid boils, the pressure again rises until it reaches the vapor pressure of the liquid. At this point equilibrium is re-established. (See Fig. L-2)

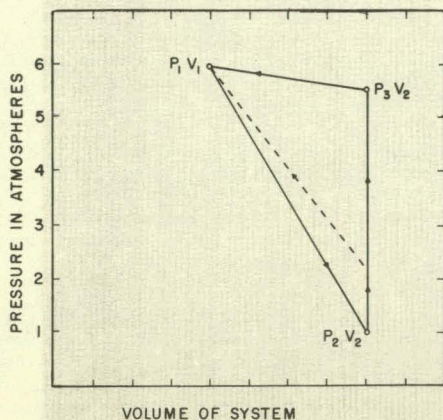


Fig. L-2. Pressure-volume relationships during the expansion-recompression cycle of a bubble chamber.

If the chamber is then recompressed, the pressure returns to its original point. This cycle is represented by the triangle $P_1 V_1$, $P_2 V_2$, $P_3 V_2$. The work done on the system is the area of this triangle. For small chambers this is satisfactory, but for a large chamber the refrigeration load is very heavy. What is done then is to use fast recompression. That is, the pressure is not allowed to rise normally along the line $P_2 V_2$, $P_3 V_2$, but is increased sharply by exposing the chamber to a tank of high-pressure hydrogen gas. As a consequence, the small triangle formed below the dashed line in Fig. L-2 represents the work done. Thus, a great saving has been effected. All chambers built since the 4-inch use this principle.

The next chamber built in Berkeley was the 10-inch (Fig. L-3). It is similar in construction to the 4-inch, except that it is turned so that the windows are on the top and bottom. The illumination system is at the bottom, inside the vacuum system. This design permitted the bubble chamber to fit within a magnet already on hand. This chamber used fast recompression. The expansion tank is maintained at a low pressure. A valve releases the chamber pressure into the expansion tank. About 20 milliseconds later the valve moves to another position, connecting high-pressure gas from the re-compression tank to the chamber. This stops the boiling. In between cycles (an interval of about 5 to 6 seconds) a compressor pumps the gas from the expansion tank to the recompression tank.

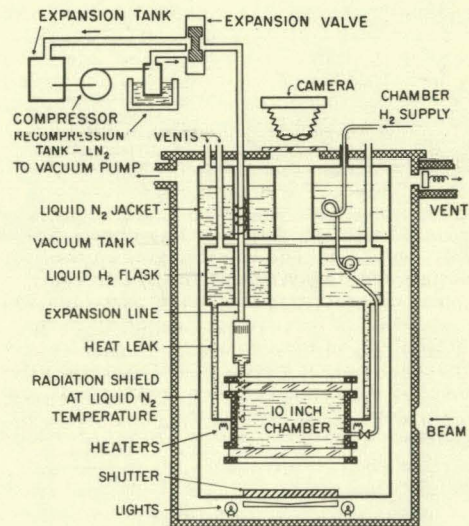


Fig. L-3. The 10-inch liquid hydrogen bubble chamber.

Next to be built was the 15-inch bubble chamber. It was to serve as a 1/5-scale model for the large 72-inch chamber then being planned. The 15-inch chamber contains several improvements. For one thing, it has a larger stereo angle. That is, the two camera lenses are placed farther apart, allowing greater accuracy in measuring the depth of a particle track in the liquid. Another improvement is the use of a variable heat leak, which means that the heat load is not fixed. This cuts down on the consumption of liquid hydrogen. Also, the 15-inch chamber has a single window mounted on top of the chamber.

The single-window design created certain optical problems. The light must be put in through the same window from which photographs are taken, yet the light must not be reflected directly into the lenses. The solution to this was the use of special plastic reflectors called "coat hangers" (Fig. L-4). Each coat hanger has a narrow silvered strip on the bottom, whereas the rest of the bottom and sides are painted black. Rays from the light source come to a focus on the silvered strip and return through the liquid unattenuated. However, scattered light coming from bubbles or other sources does not focus on the strip and is absorbed. This scheme prevents double images which would be confusing and would therefore restrict the number of tracks that could be examined on a single exposure.

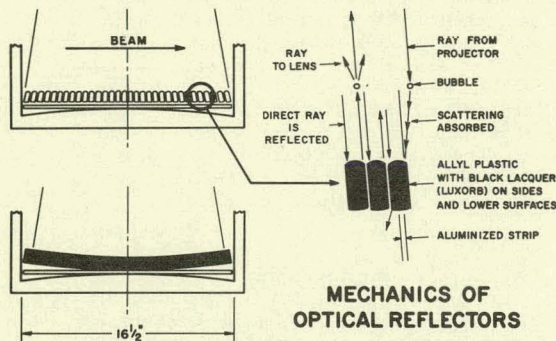


Fig. L-4. Retrodirective reflectors ("coat hangers") used with the illumination systems of the 15-inch and 72-inch bubble chambers.

There is an optimum size for a bubble chamber, depending upon the energy of the accelerator with which it is to be used. What defines this size is the nuclear interaction being studied. Consider a 6-Bev π meson interacting with a proton (Fig. L-5). The reaction products illustrated are a lambda (Λ^0) and a theta (θ^0), which are shown emerging in the forward direction and decaying one mean life away from the point of production. Those particles which leave the scene of an interaction traveling in the same direction as the incident particle will receive more momentum. As a result of relativistic effects, the lifetime of the particle is then lengthened, and the decay which we wish to observe will occur at a greater distance from the scene of the interaction.

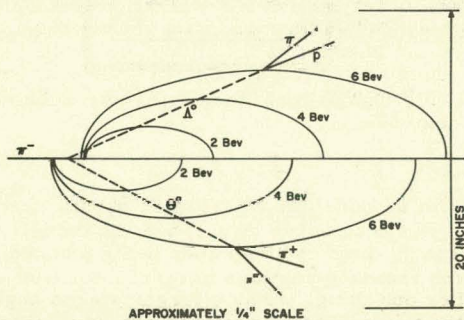


Fig. L-5. Decay contours for a Λ^0 and θ^0 produced from the interaction of an incoming π^- meson with a proton. The ellipses are loci of the points of most probable decay of the reaction products (i.e., the Λ^0 and θ^0). The size of the ellipses depends on the energy of the incoming particle.

Those particles which go off at large angles receive less momentum, and their decays occur nearer the point of interaction. This consideration dictates an elliptical shape for the chamber. Consideration of the lifetimes and energies involved in the study of the strange particles at Bevatron energies permits a rough choice of an optimum chamber size.

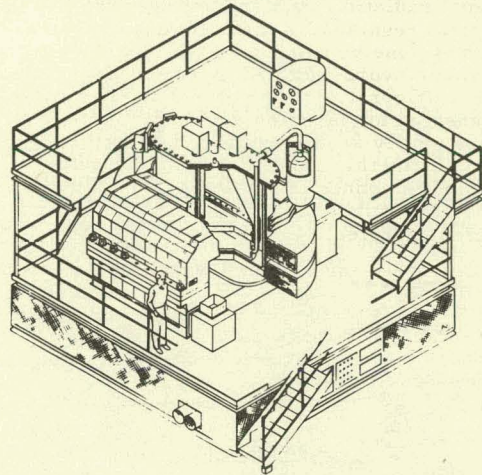


Fig. L-6. The 72-inch liquid hydrogen bubble chamber.

The 72-inch bubble chamber, whose design was started nearly four years ago, was recently put into operation. In Fig. L-6 we see an artist's conception of it. The chemical composition of the stainless steel casting that contains the 520 liters of liquid hydrogen was chosen so that the steel will remain nonmagnetic at low temperatures. Should it become magnetic, it would distort the magnetic field inside the chamber. Assembled to the chamber is the hydrogen shield, a protective canopy which is to contain the hydrogen gas in the event of a glass-window failure. Impurities in the hydrogen must not exceed one part per million; any impurity (except helium) freezes at 27°K and clouds the glass and the plastic coat hangers. If this coating becomes as thick as a



Fig. L-7. A photograph taken with the 72-inch bubble chamber.

wave length of light it renders the bubbles invisible. A 3-lens stereoscopic camera is provided. In performing a momentum balance on a nuclear event, we must reconstruct the directions of the particle tracks in space. Two stereoviews would be adequate, except for those tracks that lie along the line joining the two lenses. Figure L-7 is a photograph taken with the 72-inch chamber.

Safety

As is well known, hydrogen is violently inflammable. In the 72-inch chamber 520 liters of this material are used. Careful study has therefore been given to the matter of safety. For example, in the building that houses the chamber all the wiring is explosion-proof. No one is permitted to wear nylon clothing in the vicinity of the chamber. The air is continuously sampled at several points around the chamber, and a manual sampling is performed every half hour. Outside the building (over the side of an embankment) there is a 22-ft.-diameter Horton sphere into which the hydrogen can be dumped in case of an emergency.

Data Analysis

In cosmic ray research with cloud chambers it was common to obtain one or two interesting nuclear events each day. The techniques of analysis were such that it took a man a full day to do a momentum-balance calculation and properly classify the event. The bubble chamber has increased the incidence of interesting events by a factor of a thousand. If the previous techniques of analysis were still used, 1000 physicists would be kept continually busy with the accumulation of data. Consequently, highly automatic techniques have been developed. One of the more novel is a device called "Frankenstein" (Fig. L-8). It is a projection microscope used for measuring coordinate

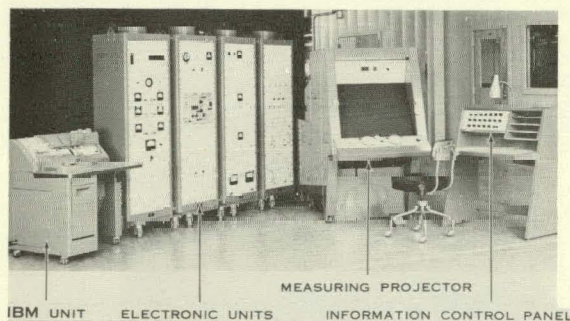


Fig. L-8. The Frankenstein film-measuring projector.

points of particle tracks on the film. An operator aligns the crosshairs with the track in an approximate fashion and the instrument "locks on" the track. The operator then drives along the track, periodically punching a button to record the coordinate information automatically on IBM cards or punched tape.

All the calculations are subsequently made by an IBM computer using various programs. The first program makes the geometrical corrections, yielding a set of directional cosines and the momentum for each particle. Next, a number of sorting programs may be used. The machine assumes that an event is of a certain type, performs certain calculations based on this assumption, and then prints out the probability that it actually was the type assumed. If the probability is poor, the machine makes another assumption and proceeds as before. Occasionally, we find events that do not fit our known categories, and in these there is always the chance of finding completely new processes or particles. The major emphasis of most of our experiments lies in the direction of making specific measurements on interaction processes.

Bibliography

1. W. B. Fretter, Nuclear Particle Detectors (Cloud Chambers and Bubble Chambers), in *Annual Review of Nuclear Science*, J. G. Beckerley et al., Eds., Vol. 5 (Annual Reviews, Stanford, California, 1955), pp. 145-178.
2. D. A. Glaser, The Bubble Chamber, *Scientific American*, 46-50 (Feb. 1955).
3. D. A. Glaser, The Bubble Chamber, in *Handbuch der Physik*, S. Flugge, Ed., Vol. 45 (Springer-Verlag, Berlin, 1958), pp. 314-341.
4. S. Glasstone, *Sourcebook on Atomic Energy*, Second Edition (Van Nostrand, Princeton, 1950). Ch. VI.
5. J. D. Gow, Development and Operation of Liquid Hydrogen Bubble Chambers, in *Proceedings of the Second International Conference on the Peaceful Uses of Atomic Energy*, Geneva, 1958 (United Nations, Geneva, 1958), Vol. 30, pp. 166-173.

M. COMPUTER PROGRAMMING

John C. Caris

This lecture will be devoted to considering how one writes a simple program for the IBM-650 computer. A program is that set of detailed instructions by means of which the computer carries out the various operations required of it to complete a computation. To make the lecture more meaningful, a specific example will be discussed and a program written for it.

The sample equation to be solved is

$$S = (1/2)gt^2 + v_0t, \quad (1)$$

where S is the distance that a freely falling object travels in time t if it has an initial velocity v_0 .

What we wish to do is obtain various solutions to Eq. (1) corresponding to different values of the independent variables t and v_0 . That is, we want the machine to punch out an IBM card giving a value of the distance S , and on the same card we would like to have punched the particular combination of t and v_0 that gave this value. Thus, the output data will consist of S , t , and v_0 .

The input data that must be supplied to the machine are the various combinations of t and v_0 for which we desire solutions. It should also be noted that in Eq. (1) there are two constants, $1/2$ and g , which must be supplied. These are called machine constants.

It is necessary to decide what locations on the IBM cards will be used for the input data, machine constants, and output data. Figure M-1 shows the locations chosen for each class of card. These choices are of course arbitrary. The programmer designs the card formats for his own convenience.

"DIST" Program

$$S = 1/2 gt^2 + v_0t$$

Machine constants: $1/2, g$

Input data: t, v_0

Output results: t, v_0, S

Input-data cards

t	v_0	zeros
sec	ft/sec	→

Machine-constants cards

0000020101	$1/2$	g	zeros
			→

Output cards

t	v_0	S	zeros
sec	ft/sec	ft	→

Fig. M-1. IBM cards, showing the locations chosen for each quantity.

The machine must be told explicitly how to proceed at every point of the program. This sequence is outlined below:

a. The first card to be fed to the computer is a memory-clear card, which clears the machine of previous data.

b. Next, a deck of IBM cards called a load-punch subroutine is fed into the computer. This deck merely tells the computer how to load in cards and punch out answers.

c. Then the machine-constant card bearing the values $1/2$ and g is loaded in.

d. Next come the program cards, a series of cards telling the machine how to proceed step-by-step through the computations of $S = (1/2)gt^2 + v_0t$. Figure M-2 is a flow chart showing the steps in this computational cycle. The machine goes through this cycle once for each value of S obtained. Preparing the flow chart and writing the program will be dealt with in detail below.

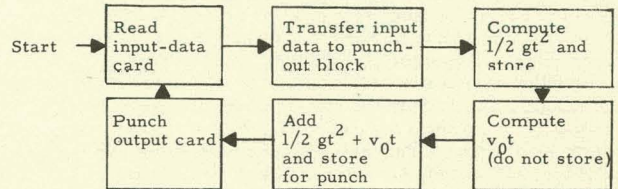


Fig. M-2. Flow chart of program.

e. Next the transfer card is loaded in. This card tells the computer to go to the beginning of the program and proceed through it.

f. Next are fed in the input data cards, each containing a combination of v_0 and t .

The first program card contains a code word that tells the computer to read one input-data card. This starts the computer going around the loop on the flow chart. When the first cycle is completed, the machine punches out one output-data card containing a value for S and alongside it the values of v_0 and t on which it was based. The computer then proceeds to read the next input-data card and a second cycle begins. This cycling continues until all the input-data cards have been read, at which time the computation is completed and the computer stops since there are no more input-data cards to be read.

The IBM-650 System

Before we can discuss how to prepare the program we must first describe some of the basic parts of an IBM-650 computer (see Fig. M-3). Lack of space makes it necessary to refer the reader to the IBM 650 Data Processing Bulletins for many interesting details of operation and construction.

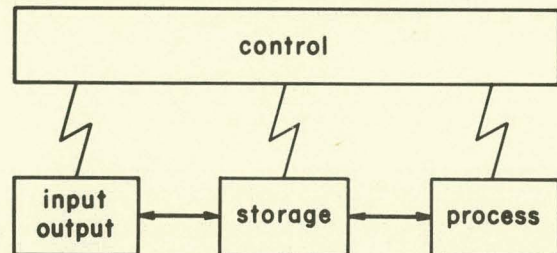


Fig. M-3. Block diagram of the IBM-650 computing system.

Input Block

The input block is used to enter data, instructions, etc., into the storage block of the system. The input unit is capable of reading IBM cards and translating decimal input information into the biquinary system used by the computer.*

Output Block

The output block translates the results obtained by the computer from the biquinary code back into a decimal form, and punches IBM output-data cards or prints a report on paper. The operation of the input-output units is supervised by the control block.

Storage

The storage block is used to store machine constants, input data, instructions, intermediate results, final results, and any other pertinent information. The IBM-650 system uses four types of magnetic storage: drum, core, tape, or disk. However, only the magnetic-drum type will be discussed here.

The computer's memory is a magnetic drum rotating at 12,500 rpm. Physically, it is a cobalt-nickel-plated cylinder approximately 4 in. in diameter and 16 in. long. Information is stored on the drum in "words," each of which consists of 10 digits and an algebraic sign. Two thousand 10-digit words may be stored on the drum at a given time.

Each specific word is assigned a 4-digit number (address), which locates the word on the drum. One should consider drum addresses merely as locations of certain information on the drum, much the same as you consider your house number as the

* The biquinary digit representation (or code) is used internally by the 650 (not the binary representation). The computer translates input decimal representations into this code and translates the code back into decimal representation for the output results. The biquinary code represents a number from 0 through 9 by two groups of units: binary units of which there are two and quinary units of which there are five.

The first binary unit designates 0 and the second designates 5. The quinary units designate one of two possible numbers (i.e., 0 or 5, 1 or 6, 2 or 7 etc.). The choice between possible numbers is made by the binary units. Reference to the table below will make this clear.

Binary Units	Quinary Units
0 5	0 1 2 3 4
0 0	0 0 0 0 0
	5 6 7 8 9

Decimal number	Bi-Quinary number
0	● 0 ● 0 0 0 0
1	● 0 0 ● 0 0 0
2	● 0 0 0 ● 0 0
5	0 ● ● 0 0 0 0
9	0 ● 0 0 0 0 ●

etc.

location of your house. These addresses range from 0000 to 1999 and are arranged in twenty 50-word bands (see Fig. M-4, the drum-layout form on which a record of the information in each address is kept). Each band lies circumferentially on the drum. (Note: The drum-layout form has space for only 1000 addresses. One must usually use two forms when programming.)

On the drum-layout form there are certain regions designated as "reading" and "punching." Information may be put into the machine in the read-in blocks only. Similarly, information may be extracted from the punch-out blocks only.

A digit is identified on the drum by means of a pattern of magnetic spots. These spots remain on the drum until erased by recording new information in the same location. As the drum rotates past a group of stationary coils of wire (one for each band), the magnetic spots are read or written. The time required for one word-location to pass one of these read-write heads is called a word-time. Since there are 50 word-times around the circumference of the drum, and since the drum rotates at 12,500 rpm, one word-time is 0.096 msec (or 0.000096 sec). The average time required to locate a given word is 25 word-times (about 2.4 msec). This access time can be reduced to one word-time by optimizing the locations of the various addresses.

Process

In the process block are performed arithmetic or logical operations, decimal points are aligned, information is shifted right or left, editing is performed, and data are rearranged.

The process block consists of an accumulator, a distributor, and an adder (see Fig. M-5). The accumulator is an electronic storage unit having a capacity of two words (20 digits and a sign). This unit temporarily stores the results of a computation while the calculation is being performed. The accumulator is divided into two halves, an upper and a lower section. The address of the lower accumulator is 8002, while that of the upper is 8003.

The distributor is also an electronic storage unit, but it has a capacity of only one word (10 digits and a sign). It is used to store data employed in arithmetic operations. The distributor acts as an intermediate link (or office boy) between the accumulator and drum memory. Any information transferred between the accumulator and memory must pass through the distributor. Its address is 8001.

Arithmetic operations are performed by combining the contents of the distributor with those of the accumulator and storing the results in the accumulator. This combining is performed by the adder.

Control

The control block receives each instruction of the program, determines the operation to be performed, and then executes the operation. While the operation is being carried out, the control block monitors the flow of data to check for any error. In addition, this unit notifies the operator when manual intervention is needed.

The program register, although part of the control block, is an electronic storage unit having a

program

	0	1	2	3	4	5	6	7	8	9
00	50	00	50	00	50	00	50	00	50	00
01	51	1/2	51	01	51	01	51	01	51	01
02	52	q	52	02	52	02	52	02	52	02
03	53	03	53	03	53	03	53	03	53	03
04	54	04	54	04	54	04	54	04	54	04
05	55	05	55	05	55	05	55	05	55	05
06	56	06	56	06	56	06	56	06	56	06
07	57	07	57	07	57	07	57	07	57	07
08	58	08	58	08	58	08	58	08	58	08
09	59	09	59	09	59	09	59	09	59	09
10	60	10	60	10	60	10	60	10	60	10
11	61	11	61	11	61	11	61	11	61	11
12	62	12	62	12	62	12	62	12	62	12
13	63	13	63	13	63	13	63	13	63	13
14	64	14	64	14	64	14	64	14	64	14
15	65	15	65	15	65	15	65	15	65	15
16	66	16	66	16	66	16	66	16	66	16
17	67	17	67	17	67	17	67	17	67	17
18	68	18	68	18	68	18	68	18	68	18
19	69	19	69	19	69	19	69	19	69	19
20	70	20	70	20	70	20	70	20	70	20
21	71	21	71	21	71	21	71	21	71	21
22	72	22	72	22	72	22	72	22	72	22
23	73	23	73	23	73	23	73	23	73	23
24	74	24	74	24	74	24	74	24	74	24
25	75	25	75	25	75	25	75	25	75	25
26	76	26	76	26	76	26	76	26	76	26
27	77	27	77	27	77	27	77	27	77	27
28	78	28	78	28	78	28	78	28	78	28
29	79	29	79	29	79	29	79	29	79	29
30	80	30	80	30	80	30	80	30	80	30
31	81	31	81	31	81	31	81	31	81	31
32	82	32	82	32	82	32	82	32	82	32
33	83	33	83	33	83	33	83	33	83	33
34	84	34	84	34	84	34	84	34	84	34
35	85	35	85	35	85	35	85	35	85	35
36	86	36	86	36	86	36	86	36	86	36
37	87	37	87	37	87	37	87	37	87	37
38	88	38	88	38	88	38	88	38	88	38
39	89	39	89	39	89	39	89	39	89	39
40	90	40	90	40	90	40	90	40	90	40
41	91	41	91	41	91	41	91	41	91	41
42	92	42	92	42	92	42	92	42	92	42
43	93	43	93	43	93	43	93	43	93	43
44	94	44	94	44	94	44	94	44	94	44
45	95	45	95	45	95	45	95	45	95	45
46	96	46	96	46	96	46	96	46	96	46
47	97	47	97	47	97	47	97	47	97	47
48	98	48	98	48	98	48	98	48	98	48
49	99	49	99	49	99	49	99	49	99	49

NOT USED FOR TABLES

Fig. M-4. Drum-layout form.

capacity of one word. It stores each instruction of the program while the instruction is being analyzed and executed. Associated with the program register are the operation register and the address register (Fig. M-5). The former receives the operation code (to be described below), analyzes it, and determines the operation to be performed. The latter receives the address portion of the instruction.

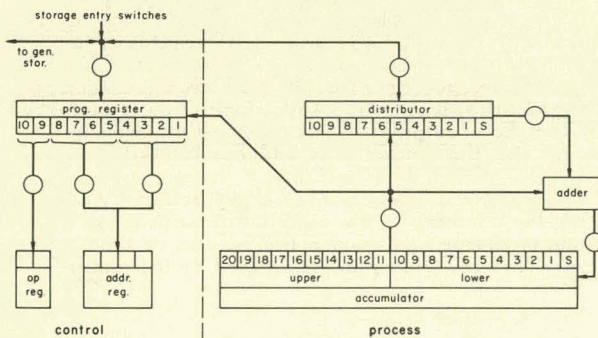


Fig. M-5. Process block of the IBM 650.

Program Instructions

Each instruction in the program must indicate:

- the operation to be performed;
- the storage location of data to be used in the operation; and

c. the storage location of the next instruction of the program. Momentary reflection shows that this three-part instruction is sensible. This is exactly how one teaches or supervises. First one must explain what is to be done (a.); second, one must explain where the student finds the data or material necessary to accomplish the operation (b.); and, finally one must explain where or how the student obtains additional instructions to proceed (c.) to the next step.

An instruction consists of ten decimal digits and sign. These ten digits are divided into three groups (Fig. M-6):

a. operation code — two digits designating the operation to be performed, e.g. addition is given the code designation 32; multiplication is 39; etc. The operation code is punched in positions 10 and 9 in the instruction word.

b. data address — four digits (punched in positions 8, 7, 6, and 5 of the instruction word) designating a drum address, with a meaning that depends on the nature of the operation code. Some of these meanings are:

- drum address of the information to be used in the operation;
- drum address in which information is to be stored by the operation;
- drum addresses into which information is to read from a card; or

4. drum addresses from which information is to be punched onto a card.

c. Instruction address — four digits (punched in positions 4, 3, 2, and 1 of the instruction word) which normally designate the drum address of the next instruction to be performed. Each step of the program consists of one of these instruction-type words.

op code		data address				instruction address				units
10	9	8	7	6	5	4	3	2	1	0

Fig. M-6. Instruction grouping.

Writing the Program

The programmer begins by assigning a code name to the problem to be solved. A code name is simply a convenient shorthand for specifying a given program and is used for such things as record keeping. Since the equation $S = (1/2)gt^2 + v_0 t$ describes distance, we shall call this program "DIST," which is short for program to compute distance by $S = (1/2)gt^2 + v_0 t$.

The programmer next draws up a form such as that shown in Fig. M-1, which gives the locations of the input data, machine constants, and output data on their respective IBM cards.

The programmer then draws the flow chart (Fig. M-2) showing the steps in a computational cycle.

Next, he reserves certain regions of the drum for certain information to be stored. These choices are made for the programmer's convenience. He records these locations on the drum layout form (Fig. M-4) so he does not inadvertently use the same addresses for two functions. In this case, band 9 is reserved for the load-punch subroutine. Addresses 1900-1949 are assigned to the "load" portion, while addresses 1950-1999 are reserved for the "punch" section. Similarly, band 2 is reserved for the program instructions. The machine constants $1/2$ and g are also entered: $1/2$ is put in address 0101 and g is 0102. The input variables t and v_0 are assigned addresses 0151 and 0152, respectively. Note that both the machine constants and the input variables are given addresses in the "read-in band." However, the specific address may be chosen arbitrarily. From the flow chart (Fig. M-2), one observes that the quantity $(1/2)gt^2$ has to be stored temporarily while the remainder of the equation is being calculated. Therefore, $(1/2)gt^2$ is put into address 0161. Finally, the output data, v_0 , and S , are assigned addresses 0177, 0178, and 0179, respectively. We should repeat that these choices are arbitrary for our convenience.

The programmer must now translate the flow chart into machine instructions that are compatible with the locations adopted on the drum. To do this, he uses the "650 Program Sheet" reproduced here as Table M-I, and a list of operation codes (Table M-II). The steps for each instruction number are summarized below:

Instr. no. 1 (program card no. 1)

a. Enter 0200 as the drum location of the first instruction.

b. Consulting the flow chart, one finds that the first step in the cycle calls for the machine to read an input-data card. From Table M-II, the operation code for "read a card" is found to be 70, which is entered in the column headed "OP." Also the abbreviation RD1 is entered under "OPERATION ABBRV;" this abbreviation is not actually punched on an IBM card, but is a convenient mnemonic device. (The "70", of course will subsequently be punched into columns 10 and 9 of the first IBM program card.) Note that the Operation Abbreviation is used as the operation code in methods of programming other than the most simple type we are using.

c. The first input-data card, which bears a value of t , is to be read into address 0151; therefore, 0151 is entered in the column headed "DATA." (These four digits are subsequently punched into columns 8, 7, 6, and 5 of the first IBM program card.) A unique feature of the IBM 650 is that when operation code 70 is followed by any data address in the block 0151 to 0160, all other input-data words on the card are automatically read into the drum without any further instructions. Thus, t and v_0 are now entered.

d. The machine must be told where to find the next instruction, which in this case, is in address 0201.

Table M-I									
650 PROGRAM SHEET									
PROBLEM: $S = (1/2)gt^2 + v_0 t$									
INSTR NO.	LOCATION OF INSTRUCTION	a	b	c	d	REMARKS			
		OP	DATA	INSTRUCTION	OPERATION ABBRV.				
1	0200	70	0151	0201	RD1	read in data cards t and v_0	step 1		
2	0201	69	0151	0202	LDD	load distributor with t			
3	0202	24	0177	0203	STD	store distributor	step 2		
4	0203	69	0152	0204	LDD	load distributor with v_0			
5	0204	24	0178	0205	STD	store distributor			
6	0205	60	0151	0206	RAU	reset and add t to upper			
7	0206	39	0151	0207	FMP	multiply $t \times t$			
8	0207	37	0101	0208	FMP	multiply $(1/2) \times t^2$	step 3		
9	0208	39	0102	0209	FMP	multiply $g \times (1/2)t^2$			
10	0209	21	0161	0210	STU	store $(1/2)gt^2$ in 0161			
11	0210	60	0151	0211	RAU	reset and add t to upper	step 4		
12	0211	39	0152	0212	FMP	multiply $v_0 \times t$			
13	0212	32	0161	0213	FAD	add $(1/2)gt^2 + v_0 t$			
14	0213	21	0179	0214	STU	store $(1/2)gt^2 + v_0 t$ in 0179	step 5		
15	0214	71	0177	0200	WR1	punch out S , t , and v_0	step 6		

Table M-II

Operation	Codes for the IBM 650	
Operation	Code	Abbreviation
Floating Add	32	FAD
Floating divide	34	FDV
Floating multiply	39	FMP
Floating subtract	33	FSB
Load distributor	69	LDD
Punch a card	71	WR1
Read a card	70	RD1
Reset and add to upper accumulator	60	RAU
Store distributor	24	STD
Store upper accumulator	21	STU

Instr. no. 2 (program card no. 2)

- a. Enter 0201 as the location of the second instruction.
- b. The next step on the flow chart is to transfer the input data to the punch-out block. This is done because we want to identify each value of S with the values of t and v_0 used to compute it. We recall that one must transfer to a punch-out block any information we wish the machine to punch out for us. Of course the values of t and v_0 remain in addresses 0151 and 0152 as well as in 0177 and 0178. The distributor is used to effect this transfer. The distributor must be told to go first to address 0151, pick up t , and carry it to the punch-out block. From Table M-II we find that operation code 69 is used to load the distributor. Thus, on the program sheet "69" is written under the heading "OP," and "LDD" is entered as the operation abbreviation.

c. The address of the data to be loaded into the distributor is 0151, which now entered under "DATA."

d. The next instruction is located in address 0202.

Instr. no. 3 (program card no. 3)

a. Enter 0202 as the location of instruction no. 3.

b. We now wish to store t in the punch-out band in address 0177. From Table M-II we find that the operation code for this is 24, which is entered under "OP." The symbol STD (store the distributor) is entered as the operation abbreviation.

c. The address in which t is stored in the punch-out band is 0177, which is now entered under "DATA."

d. The next instruction is located in address 0203.

Instr. no. 4 (program card no. 4)

a. Enter 0203 as the location of instruction no. 4.

b. We now want to load the distributor with v_0 so that it can be transferred to the punch-out block. Therefore, we again write "69" under "OP" and "LDD" as the operation abbreviation.

c. The address of v_0 is 0152, which is entered under "DATA."

d. The location of the next instruction is 0204.

Instr. no. 5 (program card no. 5)

a. Enter 0204 as the location of instruction no. 5.

b. To store v_0 in the punch-out block we enter operation code 24 under "OP" and STD as the operation abbreviation.

c. The address in which v_0 is to be stored in the punch-out block was chosen as 0178, which is next entered under "DATA."

d. The location of the next instruction is 0205.

Instr. no. 6 (program card no. 6)

a. Enter 0205 as the location of instruction no. 6.

b. From the flow chart we see that the next step is to compute $(1/2)gt^2$ and store it. First, t^2 will be computed (we will use simplest programming procedure to compute t^2 although it is not the fastest from the viewpoint of the machine). The first step is to reset the upper accumulator to zero and then enter t in the upper accumulator. The operation code for this is 60, which is entered under "OP." The abbreviation RAU (reset and add to upper) is also entered under "OPERATION ABBRV."

c. The address of t is 0151, which is entered under "DATA."

d. The location of the next instruction is 0206.

Instr. no. 7 (program card no. 7)

a. Enter 0206 as the location of instruction no. 7.

b. At this point t is in the upper accumulator. We will multiply t times t , but in a manner by which the computer keeps track of the decimal point for us. This is called the "floating-decimal-point" method, in which the position of the decimal point is designated by a two-digit number as shown in Table M-III. The manner of representing a number and the position of the point on an IBM card is shown in Table M-IV. From Table M-II we find that the operation code for floating multiply is 39, which is now entered under "OP." The operation abbreviation for this is FMP.

Table M-III

IBM-650 Method of keeping track of the decimal point

Number	Floating-decimal-point notation
839.67838	53
83.967838	52
8.3967838	51
0.83967838	50
0.083967838	49 etc.

Table M-IV

Method of representing a number and its decimal point on an IBM card by the floating-decimal-point system

10	9	8	7	6	5	4	3	2	1	0
Number with 8 significant figures									Floating decimal point	(Sign)

The numbers designate digit positions in the word only. For example, the floating decimal point occupies digit positions 1 and 2. The first number from Table M-III would appear as

8	3	9	6	7	8	3	8	5	3	+
---	---	---	---	---	---	---	---	---	---	---

c. The address 0151 is now entered under "DATA."

d. The location of the next instruction is 0207.

Instr. no. 8 (program card no. 8)

a. Enter 0207 as the location of instruction no. 8.

b. At this point we have t^2 in the upper accumulator. Now we wish to multiply $1/2$ times t^2 . Therefore, we again enter 39 under "OP" and FMP under "OPERATION ABBRV."

c. The address of $1/2$ is 0101, which is now entered under "DATA."

d. The location of the next instruction is 0208.

Instr. no. 9 (program card no. 9)

a. Enter 0208 as the location of instruction no. 9.

b. We now have $(1/2)t^2$ in the upper accumulator, and wish to multiply this quantity by g . We write 39 under "OP" and FMP under "OPERATION ABBRV."

c. The address of g is 0102, which is now entered under "DATA."

d. The location of the next instruction is 0209.

Instr. no. 10 (program card no. 10)

a. Enter 0209 as the location of instruction no. 10.

b. We now have $(1/2)gt^2$ in the upper accumulator and want to store it in address 0161. From Table M-II we find that the operation code for "store the upper" is 21 and the abbreviation is STU; these are entered in the appropriate columns.

c. The data address 0161 is now entered.

d. The location of the next instruction is 0210.

Instr. no. 11 (program card no. 11)

a. Enter 0210 as the location of instruction no. 11.

b. The quantity $(1/2)gt^2$ is now stored on the drum. Next, we want to compute $v_0 t$ but not store it. We must first reset the accumulator to zero and then add t to the upper accumulator. The operation code for this is 60 and the abbreviation is RAU, which are now entered.

c. The address of t is 0151, which is now entered as the data address.

d. The location of the next instruction is 0211.

Instr. no. 12 (program card no. 12)

a. Enter 0211 as the location of instruction no. 12.

b. We now floating multiply v_0 times t . Enter the operation code 39 and the abbreviation FMP.

c. Enter the address of v_0 , which is 0152.

d. Enter 0212 as the location of the next instruction.

Instr. no. 13 (program card no. 13)

a. Enter 0212 as the location of instruction no. 13.

b. At this point $(1/2)gt^2$ is stored on the drum and $v_0 t$ is in the upper accumulator. Now we wish to floating add these two quantities. From Table M-II we find that the operation code for this is 32 and the abbreviation is FAD.

c. The address of $(1/2)gt^2$ (which we want to add to $v_0 t$) is 0161, which is now entered under "DATA."

d. Enter 0213 as the location of the next instruction.

Instr. no. 14 (program card no. 14)

a. Enter 0213 as the location of instruction no. 14.

b. At this point $(1/2)gt^2 + v_0 t$ is in the upper accumulator. This must now be transferred to address 0179 in the punch-out block. Therefore, enter operation code 21 and its abbreviation STU (store the upper).

c. Enter the address in which the above is to be stored, i.e., 0179.

d. Enter 0214 as the location of the next instruction.

Instr. no. 15 (program card no. 15)

a. Enter 0214 as the location of instruction no. 15.

b. We now want the machine to punch out the values for S , t , and v_0 on an output-data card. From Table M-II we find that the operation code for "write a card" is 71 and its abbreviation is WR1, which are now entered in their respective columns.

c. The words that we want punched out are in three different addresses, namely, 0177, 0178, and 0179. A feature of the 650 allows all three to be punched out on the same card when any address in the punch-out block is given. Therefore, we enter 0177 under "DATA" to effect the punch-out of S , t , and v_0 . The punch unit of the 650 punches the results in floating-point word form on a card which comes out of the machine.

d. One loop of the program cycle has now been completed.

We now want the computer to process the second input-data card in the same fashion. Therefore to commence the second cycle we write 0200 as the location of the next instruction. This seemingly complicated procedure is performed in much less than one second from the time the input card goes into the machine to the time the output card is punched.

Conclusion

The foregoing has been an explanation of how to write a simple program. In many cases there are more elegant methods which would allow quicker

processing of the data. However, this basic approach should be mastered before considering the advanced techniques.

Bibliography

1. IBM 650 Data Processing Bulletins: No. G24-5000-0 1/59, General Information, Console Operations, Special Devices; No. G24-5002-0 1/59, Basic Operation Codes, Program Optimizing, Program Loading (International Business Machines Corporation, New York, 1958).

N. CHROMATOGRAPHY

Wilson Johnson*

Chromatography is a very appropriate topic for high school study. It is readily demonstrated with inexpensive equipment, and the physical principles upon which it is based are easy to comprehend. Also, the development of chromatography is an interesting story of progress in jumps, with long periods of inactivity in between. Today it is one of the most useful separation techniques in industry. Yet ten years ago gas chromatography, the main topic of interest at present, was unknown. No major instrument maker would dare omit it from his repertoire; if he did, he would face a certain business failure because so much money now goes into this type of equipment.

Chromatography was developed between 1903 and 1904 by the Russian botanist Tswett. He was attempting to separate such similar and complicated chemical compounds as root and plant extracts. He knew that existing methods such as distillation were not capable of doing this. Therefore, he experimented until he discovered the technique represented in Fig. N-1. This turned out to be a remarkable solution to a very difficult problem.

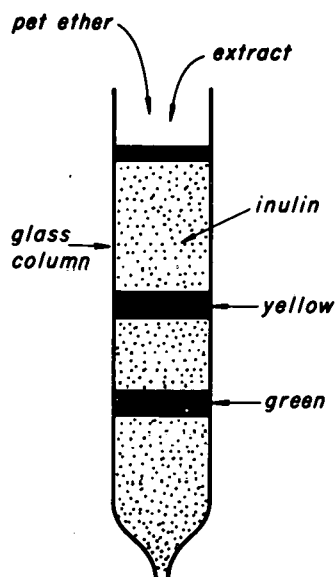


Fig. N-1. Tswett's original experiment—the earliest chromatographic separation.

He placed a powder called inulin into a column that was constricted at the bottom so that it could not run out. Inulin is a material very similar to starch, both in appearance and chemical composition. Over this he poured petroleum ether, a highly volatile liquid, and waited until some of it began dripping out the bottom. Next, he put onto the column a layer of the extract which he wished to analyze. This had previously been dissolved in pet ether. He then kept flowing more pet ether over this layer.

After a while he noticed a yellow color in the effluent; also on the column there appeared two distinctly separate bands. The lower one was green and the upper one yellow. He allowed the material to dry, and then pushed it up out of the surrounding glass tube. Next, he sliced the material into discs, each of a different color. When he washed each disc and analyzed the solution, he found that it was composed of a single constituent of the original material.

In this way Tswett achieved the separation of plant pigments, something that had never before been done. Because he used colored materials and obtained colored bands, he named this technique "chromatography." This has turned out to be a misnomer, since color is not involved in some of the more recent types of chromatography.

The surprising thing is that chromatography was not immediately put to use. Although it was reported in the literature in 1910, this novel method lay idle until 1931 when R. Kuhn employed it in connection with his work on carotenes. It is true that Tswett's article was in Russian, but it had been reviewed in American journals soon after it was published.

The technique originated by Tswett is now called liquid-solid chromatography. Since 1931 several other types of chromatographic separation have arisen. The first of these, liquid-liquid chromatography, was developed in 1941 by Martin, an Englishman. This opened a whole new field of investigation. At first, one may find it difficult to imagine how a liquid can be held in a column while another liquid is poured through it.

Instead of an active solid such as inulin, Martin used an inert, free-flowing, porous solid called Celite, a diatomaceous earth. He then impregnated the pores of this powder with a liquid such as water. The result was still a free-flowing solid. Over this Martin then poured pet ether and proceeded as in the case of liquid-solid chromatography.

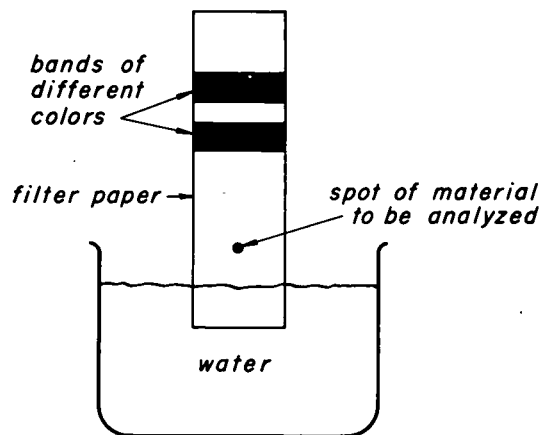


Fig. N-2. Paper chromatography.

* Guest lecturer from Shell Development, Emeryville, California.

A few years later, in 1946, Martin developed paper chromatography (see Fig. N-2). A drop of the solution to be analyzed is placed on a porous paper (for example, a drop of ink on filter paper). This paper is next dipped into a vessel containing water. As the water seeps upward by capillary attraction distinct horizontal bands appear on the paper. These colored bands are characteristic of the materials in question.

In his 1941 paper Martin also suggested the technique of gas-liquid chromatography. The procedure for preparing the column is the same as in liquid-liquid chromatography. But instead of pouring a liquid over the column, gas (such as helium) is passed through it. The sample is injected into this gas stream by means of a hypodermic syringe. Martin did not try this method until 1952, nor did anyone else. Thus, one of the best ideas of the century was available for 11 years before anyone applied it, and then Martin himself had to do the work!

Gas-solid chromatography has also been developed, but it has such a limited application that we will not discuss it. In practice, when we use the term "gas chromatography" we mean "gas-liquid chromatography."

The techniques mentioned so far are collectively called "elution chromatography." There is another type, named "displacement chromatography." This differs from elution chromatography in that the solvents which are poured over the column are much more active. In this technique a particular component in the mixture moves down the column only because it is pushed by the component immediately following it. The last component, in turn, is being pushed by an active mobile phase, called the displacer. Since one component is pushing another, adjacent components can never be completely separated. Therefore, displacement chromatography is seldom used and we will not consider it further in this lecture.

By contrast, the principle of separation utilized in elution chromatography is capable of completely separating adjacent components. To explain this principle we can make an analogy to a conveyor belt. Suppose that this conveyor belt is rotating at a constant speed and that there is a stationary platform alongside it. A person could either stand on the conveyor belt or the platform. If he were to spend all his time on the belt, he would travel at the same velocity as the belt. Likewise, if he were to stay on the platform he would have zero velocity. But by hopping back and forth between the belt and platform he would travel at some intermediate average speed, depending on the fraction of time spent on the belt. For instance, if he spent one-fourth of his time on the belt, his average velocity would be one-fourth that of the belt.

This is exactly what is happening in a chromatographic column. There is one phase that is moving and one that is stationary. A molecule in the mobile phase travels at the speed of that phase, whatever that happens to be. But its average depends on the fraction of the time spent in the mobile phase. Thus, the material which spends the greater portion of its time in the mobile phase will be carried faster, and hence come out first in the effluent.

This explanation accounts for the vacant spaces between the bands: There is no molecule present that spends the correct fraction of its time

in the mobile phase to be deposited in that particular vacant region.

By noting where a given material is deposited, we can calculate backwards to find what fraction of the time it spent in a mobile phase. All materials, if they are in dilute solution, have their own unique value for any pair of materials, whether it is a liquid in a solid, a liquid in a liquid, a gas in a liquid, or a gas in a solid. We call this value which characterizes a material its "partition coefficient."

Tables of partition coefficients have been compiled for many materials. Suppose that we have performed a separation. We would then consult these tables to find what materials deposit at a given point. Perhaps we find that five or six different materials can. If so, we would repeat the experiment using a different type of column. For instance, if we first used pet ether and inulin, we might next use water and pet ether. We would then find that the unknown material deposited at a different point. Again consulting the tables, we would find several materials that could do this. The material in question would be the one that appeared in both lists. Assume, however, that all but two materials were eliminated. We would then have to repeat the experiment until finally all the materials except one have been eliminated.

Unfortunately, at present our tables are not complete for all known materials. This is not surprising, since chromatography has only recently attracted wide attention. Most of the data that are now being collected pertain to gas chromatography, since this is the most important area. In the case of gas chromatography, the term "retention volume" is sometimes used rather than partition coefficient.

The lack of tables is not a disadvantage providing one knows that the material in question is one of several possibilities. Suppose that a biochemist knows that the material he wants to identify is one of five substances. He would run five tests, each with a different one of the known materials, and note where they came off on the column. The one that came off in the same place as the unknown sample would be the same as the unknown sample.

There are a number of advantages to gas chromatography. One of them is that samples can be more easily placed on the column because the flow rate is much faster. With the other methods, all the excess liquid must be removed from the top of the column before the sample can be placed on. Before pet ether can be poured over it, the material must once more be dry. This is a tedious process and also an uncertain one. Sometimes air gets trapped in the column and ruins the experiment. However, with gas chromatography it is a simple matter to place the sample on the column. A hypodermic syringe is used to squirt the sample into a rapid stream of gas as it enters the column. Even though some of the sample diffuses out, most of it hits the top of the column practically instantaneously.

Other advantages of gas chromatography are that the columns are easier to make, they work better, and they last longer. Most important, gas chromatography is much more sensitive in detecting trace amounts of a material. Also, the degree of its response does not depend on the type of material but only on the amount. Thus, we can use gas chromatography to tell not only what material was initially present but also in what quantity.

Before describing the detection methods used with gas chromatography, we will first explain the technique used with liquid-liquid chromatography, since the two detection methods are somewhat similar in principle. Suppose that the sample consists of several fatty acids which we wish to separate. The effluent is then run into a vessel containing sodium hydroxide. Since the pet ether (the carrier) is lighter than any of the fatty acids, it floats on top of the sodium hydroxide. But the acid settles to the bottom where it neutralizes some of the sodium hydroxide. An automatic titrator then senses this unbalance and automatically replaces enough sodium hydroxide to restore the balance. At the same time a recording pen is moved, indicating the amount of sodium hydroxide that was replaced. In this way a graph is obtained. Whenever one of the fatty acids comes through in the effluent, the recorder pen is displaced on this graph (Fig. N-3).

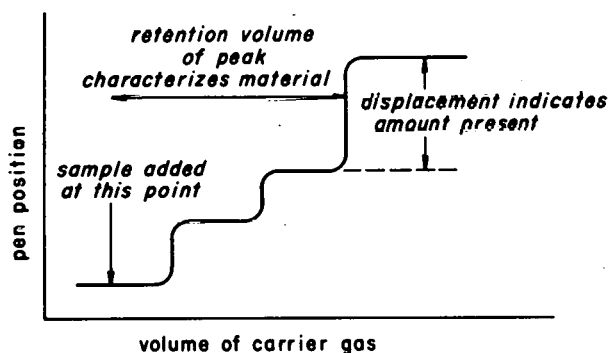


Fig. N-3. A chromatogram obtained for liquid liquid chromatography.

The distance between pen positions before and after a component has been neutralized is a measure of the amount of the material present in the original sample.

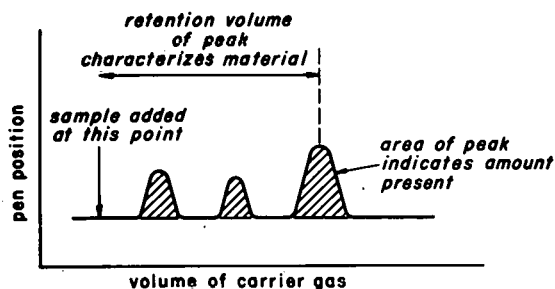


Fig. N-4. A gas chromatogram.

The detection method used with gas chromatography involves passing the gas effluent through a detector. In a similar fashion a chromatogram is obtained (Fig. N-4). The distance between peaks again characterizes the type of material. However, in this case the amount originally present is indicated by the area under the peak, not just its height as before.

Only two types of detectors used in gas chromatography will be described in this lecture: the thermal-conductivity cell and the flame-ionization chamber (see Figs. N-5 and N-6). The former

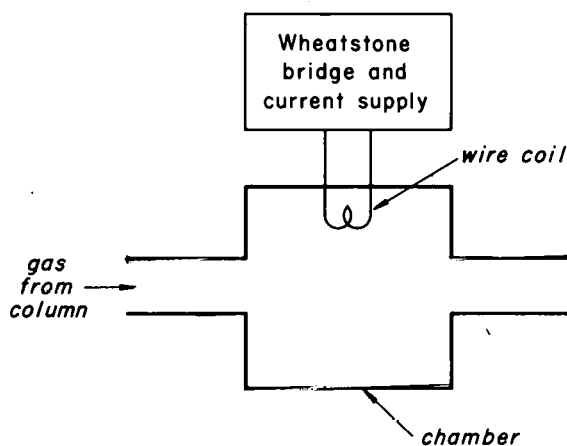


Fig. N-5. Thermal-conductivity cell.

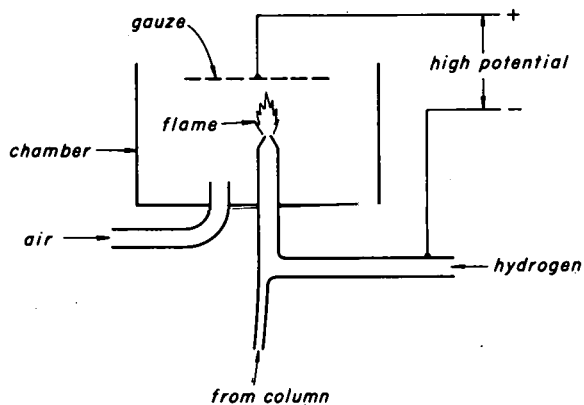


Fig. N-6. Flame-ionization chamber.

is the most commonly used but the latter is the most sensitive. Under the best conditions the thermal-conductivity cell can detect 0.0000005 gram of gas as it emerges from the column. But with flame ionization one-fiftieth of this amount can be detected.

The thermal-conductivity detector consists of a small chamber which is connected to the column. A wire coil inside this chamber is heated by a constant current. But the wire is continually cooled by the helium carrier gas coming from the column. This is because heat can transfer in the helium gas between the hot wire and the wall of the detector, which is cooler. However, when foreign molecules come along in the helium they impede this cooling of the wire and it becomes hotter. As it becomes hotter, its electrical resistance changes. This change in resistance is then measured by a Wheatstone bridge

and is automatically recorded on a graph, giving a chromatogram.

With flame ionization, nitrogen is used as the carrier gas. As it leaves the column it enters a chamber in which hydrogen and oxygen are burning. A high potential exists across the chamber, so that a small current (perhaps a trillionth of an amp) flows across the chamber in the flame. When a sample component comes along in the nitrogen stream it becomes ionized when it strikes the flame. This changes the electrical conductivity of the flame. The resulting change in current is then measured by an electrometer amplifier and automatically recorded, giving a chromatogram.

In closing, we should mention the one limitation of gas chromatography: there must be an appreciable volatility of the sample in the gas phase. To use our analogy, the material must spend some of its time on the conveyor belt in order to come off the column. There are multitudes of materials that cannot be used in gas chromatography because they

will not move down the column. Roughly, one might say that any substance that boils below 300° C can be used with gas chromatography. In special cases this limit might be 400° C or slightly higher. Of course, the material must not break down under the heat necessary to encourage it to spend some of its time in the mobile phase.

Thus, there is still a challenge, in liquid-liquid chromatography especially, to find detectors and sample injectors that will enable us to perform separations that are now impossible.

Bibliography

1. E. Lederer and M. Lederer, Chromatography; A Review of Principles and Applications, Second Edition (Elsedier, New York, 1957).
2. W. H. Stein and S. Moore, Chromatography, Scientific American, 35 ff. (Mar. 1951).

O. THE TRANSURANIUM ELEMENTS

Richard M. Diamond

It is appropriate that the subject of the trans-uranium elements be included in this lecture series, because all ten of the ones that are presently known were either discovered or co-discovered at this Laboratory. However, our story starts not here in Berkeley, but rather in Rome about 25 years ago. Enrico Fermi and his distinguished collaborators were irradiating many elementary materials with neutrons, and by neutron capture, creating new radioactive isotopes. Fermi suggested that they should bombard uranium 238, element 92, the heaviest element then known. The neutrons might be captured, giving uranium 239. This would be unstable and might beta-decay into element 93. In this way, a new element might be obtained, one which does not exist in nature.

When Fermi's group irradiated uranium 238 with neutrons they indeed found several new activities and ascribed one of them to element 93. A few years later Fermi won the Nobel Prize mainly for this discovery. As later developments showed, this was a case where the right man got the right prize for the wrong reason. Element 93 had not yet been discovered.

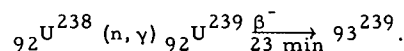
In the next four or five years the situation became very puzzling. In irradiating U^{238} , Fermi and others found not just two or three new activities, but dozens of them. They were led to ascribe some of these to elements such as radium and actinium. This was very difficult to believe, because no place else in the periodic table had anyone found that when a nucleus captures a slow neutron it could then eject three or four protons to lower its charge. Yet the uranium nucleus was apparently lowering its charge from 92 to 89 or 88.

To identify these as radium and actinium activities, they first dissolved the irradiated material; then they added a few milligrams of barium or lanthanum as carrier materials and did a precipitation analysis. For example, they might precipitate barium sulfate by adding a few milligrams of barium and then some sulfate. Barium sulfate is insoluble, as is radium sulfate. If some radium activity were present it would carry on the barium sulfate. When the precipitate was separated from the solution, it was found that one of the activities had indeed carried on the barium, and during additional tests the activity followed barium chemistry, which is very much like radium chemistry. However, barium and radium are not the same element. They are in the same family in the periodic table and their sulfates are insoluble, but they do have differences. But, it was not possible to separate the radium from the barium sulfate precipitate. In a similar fashion, lanthanum carries actinium. But, again, actinium could not be separated from the lanthanum sulfate precipitate. These results were baffling.

In 1939 Hahn and Strassmann cleared up this mystery. Uranium does not just capture a neutron and give off a gamma ray, as was previously thought. Rather, the uranium nucleus splits into two fragments, each of which is an element belonging somewhere in the middle of the periodic table. In other words,

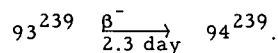
these activities that were taken to be radium and actinium because they followed barium and lanthanum chemistry were in fact barium and lanthanum! Obviously, an element cannot be separated from itself. The fission hypothesis also solved the difficulty as to why there was such a multiplicity of activities, because very many elements are made in fission. However, this hypothesis left open the question as to whether uranium will capture a neutron and go to element 93. That is, the original Fermi suggestion had not yet been proved or disproved.

In the following year, 1940, E. M. McMillan and P. H. Ableson at this Laboratory were studying the fission process. They were looking at uranium samples that had been bombarded with neutrons and observed that most of the fission-fragment pairs came off the target in opposite directions as they repelled each other coulombically. However, one of the numerous activities did not recoil away from the target, but instead stayed on the target plate. This made them suspect that they had seen the following reaction, in which there is little recoil of the resulting product:



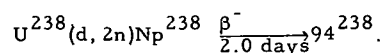
They found that this activity (${}_{93}^{239}$) had a half life of 2.3 days. They then did some chemistry to show that this activity was not uranium or any of the neighboring elements such as radium or actinium. In fact, it was not any known element. Therefore, they ascribed this activity to element 93. They named this new element neptunium after the planet Neptune, because it is outside of Uranus in the solar system.

McMillan and Ableson knew that Np^{239} should be a beta-minus emitter and should decay into the next higher element, element 94, as given by



But they were not able to find any activity that could be ascribed to element 94. The reason for this became clear later when element 94 was found. The isotope ${}_{94}^{239}$ has a very long half life—23,000 years. With the very small amount that they made they were not able to observe the few counts that were obtained.

However, element 94 was discovered later in the same year (1940) by means of charged-particle irradiation rather than neutron capture. Seaborg, McMillan, Wahl, and Kennedy used the 60-inch cyclotron to produce high-energy deuterons, which were then shot at U^{238} . The deuteron was captured by the uranium nucleus, and in the process two neutrons were given off, yielding neptunium 238. This in turn beta-decayed into element 94, as given by



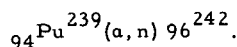
Element 94, with mass number 238, has a half life of 90 years, and so was easier to observe than ${}_{94}^{239}$. Element 94 was given the name plutonium because

Pluto is the next planet beyond Neptune.

Since 1940, almost a dozen different isotopes of neptunium and another dozen of plutonium have been discovered. The one that is best known is Pu^{239} which is now made by the neutron-capture method in large amounts; it is the fissionable material of the atomic bomb. When fission was discovered in 1939 no one foresaw its tremendous importance. What started out as a purely academic investigation proved to have rather startling consequences.

Several problems made it difficult to discover elements heavier than plutonium. Of these, the most important was that in order to obtain elements 95 and 96 by the methods already described, one must have weighable amounts of neptunium or plutonium as a target material to be bombarded by neutrons or deuterons. However, neptunium and plutonium do not exist naturally, but must be made. By contrast, uranium is fairly abundant, so that this problem did not exist in making elements 93 and 94. To make sufficient plutonium or neptunium in a cyclotron is a time-consuming task. However, during the war the first nuclear reactors were built to produce plutonium for the bomb.

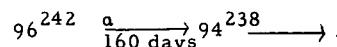
By 1944, about 10 milligrams of plutonium became available to investigators so that they could begin looking for further elements. This time alpha particles were used instead of deuterons to bombard the target material. Deuterons have only one proton, so that they can raise the charge on the target nucleus by only one. But if alpha particles (which can be accelerated just as easily in the 60-inch cyclotron) are used, the charge can be raised by two. Plutonium 239 was therefore irradiated with alpha particles to give the following reaction:



Element 96 was identified by Seaborg, James, and Ghiorso by means of a microscale qualitative

analysis for all the elements in the periodic table. This consisted mainly of precipitations of various carriers, such as insoluble sulfides and hydroxides, AgCl , LaF_3 , etc., and observing whether the activity carried. If it did not carry on silver chloride, for example, the activity could not be lead, silver, or mercurous mercury. In this way they gradually eliminated all the elements in the periodic table. Since the unknown activity was no element in the periodic table, including element 93 and uranium, they were fairly certain that it was element 96.

The complete identification of this element was made in a very simple way, one that is quite often used. This particular isotope of element 96 was an alpha-emitter. In emitting an alpha particle, element 96 decays into plutonium 238:



Since the plutonium-238 activity was already known, the parent nucleus obviously had to be plutonium 238 plus one alpha particle, or ${}_{96}^{242}$. This new element was named curium in honor of the Curies.

It was recognized that the transuranium elements were behaving like a new rare earth series. Previously, it was believed that elements 90, 91, 92, and 93 should be similar to elements 72, 73, 74, and 75. However, as early as 1940 McMillan and Ableson pointed out that neptunium did not behave like it belonged in the family with manganese and rhenium. Rather, neptunium behaved like uranium. They suggested that a new rare earth series might exist in which the elements were more similar to each other horizontally than vertically. By the time that element 96 was made, this fact became evident. Figure O-1 shows how this new rare earth series fits into the periodic table, beginning with element 90 and extending to element 103, which has not yet been discovered. We call this new series the actinides, since they follow actinium, element 89. Likewise, the light rare earths, elements 58 to 71, are called lanthanides because they follow lanthanum, element 57.

Fig. O-1. The periodic table.

Fig. O-1. The periodic table.

1 H 1.0080																	2 He 4.003				
3 Li 6.940	4 Be 9.013															5 B 10.82	6 C 12.011	7 N 14.008	8 O 16.000	9 F 19.00	10 Ne 20.183
11 Na 22.991	12 Mg 24.32															13 Al 26.98	14 Si 28.09	15 P 30.975	16 S 32.066	17 Cl 35.457	18 Ar 39.944
19 K 39.100	20 Ca 40.08	21 Sc 44.96	22 Ti 47.90	23 V 50.95	24 Cr 52.01	25 Mn 54.94	26 Fe 55.85	27 Co 58.94	28 Ni 58.71	29 Cu 63.54	30 Zn 65.38	31 Ga 69.72	32 Ge 72.60	33 As 74.91	34 Se 78.96	35 Br 79.916	36 Kr 83.80				
37 Rb 85.48	38 Sr 87.63	39 Y 88.92	40 Zr 91.22	41 Nb 92.91	42 Mo 95.95	43 Tc 101.1	44 Ru 101.1	45 Rh 102.91	46 Pd 106.4	47 Ag 107.880	48 Cd 112.41	49 In 114.82	50 Sn 118.70	51 Sb 121.76	52 Te 127.61	53 I 126.91	54 Xe 131.30				
55 Cs 132.91	56 Ba 137.36	57-71 La Series	72 Hf 178.50	73 Ta 180.95	74 W 183.86	75 Re 186.22	76 Os 190.2	77 Ir 192.2	78 Pt 195.09	79 Au 197.0	80 Hg 200.61	81 Tl 204.39	82 Pb 207.21	83 Bi 208.99	84 Po	85 At	86 Rn				
87 Fr	88 Ra 226.03	89-103 Ac Series	(104)	(105)	(106)	(107)	(108)														

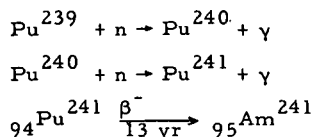
Lanthanide Series

57 La 138.92	58 Ce 140.13	59 Pr 140.92	60 Nd 144.27	61 Pm	62 Sm 150.35	63 Eu 152.0	64 Gd 157.26	65 Tb 158.93	66 Dy 162.51	67 Ho 164.94	68 Er 167.27	69 Tm 168.94	70 Yb 173.04	71 Lu 174.99
--------------------	--------------------	--------------------	--------------------	----------	--------------------	-------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------

Actinide Series

89 Ac 227.04	90 Th 232.05	91 Pa 231.05	92 U 238.04	93 Np 237.06	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 E	100 Fm	101 Mv	102	(103)
--------------------	--------------------	--------------------	-------------------	--------------------	----------	----------	----------	----------	----------	---------	-----------	-----------	-----	-------

After element 96 was discovered, Seaborg, James, and Morgan found element 95. Their method was to irradiate plutonium with neutrons. Most of the plutonium nuclei fissioned, but some captured a neutron, producing plutonium 240. Plutonium 240 captured another neutron, giving plutonium 241; plutonium 241 is a beta-minus emitter and goes to element 95, with mass number 241. These reactions are as follows:



Because it was realized that element 95 was a member of the new rare earth series, it was named americium in analogy to element 63, europium, which is right above it in the periodic table.

There ensued a period of battling with difficulties. One difficulty has been mentioned already, that of having enough target material. To go on to make elements 97 and 98 by the methods so far described, one would need weighable amounts of elements 95 and 96 as target materials. To obtain these targets, uranium could be placed in a pile and irradiated to produce plutonium 239. The plutonium 239 could then be removed and separated, the pure Pu^{239} being put back into the pile. This would undergo two successive neutron captures to form Pu^{241} which would beta-decay to Am^{241} . It might take several years to get just a few milligrams or micrograms of americium to use as a target.

In addition to the difficulty of obtaining the target material, there is the problem of separating it from the many grams of plutonium that were put in the reactor. These materials, being highly radioactive, are very dangerous to handle. Therefore, new equipment and techniques for handling radioactive materials had to be developed which would allow very complicated chemistry to be done without a human every touching the materials. The materials must be placed in a large lead-enclosed cave where they can be handled by remote control. Special precautions had to be taken to prevent the alpha particles that are emitted from being ingested.

Another handicap to purifying the target material is that it is a member of the same chemical family as plutonium, from which it must be separated. Moreover, the target material, such as americium, may constitute only one part in 10^{12} of the total material. This means that the separation techniques must be extremely good. Prior to the time of the Manhattan Project the light rare earths, or lanthanides, had never been well separated. Chemists had spent whole lifetimes trying to get pure samples by means of fractional crystallization and precipitation. Fortunately, the ion-exchange resin separations were developed about this time (see Section N "Chromatography"), to solve this very formidable problem. With this method, the lanthanides and actinides can be separated quite well in a matter of hours, even when one material is present only in trace amounts. Professor Cunningham of this Laboratory was the first to use this method for separating the transuranium elements. Today the ion-exchange column is one of the two techniques (and perhaps the best) used for this purpose.

Still another problem connected with heavy-element discovery is that of knowing something about the characteristics of the nucleus with which one is

dealing. All of these nuclei are, of course, radioactive, but it would be helpful to know something about the nature of their decay. All the elements heavier than lead are alpha-active, but some of them decay predominantly by beta decay, by electron capture, or by spontaneous fission. It would help very much in designing the experiment to find them if one knew what to expect. It would be helpful to know the approximate energy of the decay, particularly when using an alpha counter. It would also help to know what half life to expect. Fortunately, about this time there began to develop a heavy-element systematics based on the following sort of reasoning: All of the elements above lead are unstable toward alpha decay. It is relatively easy to calculate empirically what the alpha-decay energy should be. For instance, a given element Z of mass number A will alpha-decay as follows:

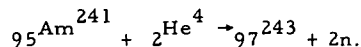
$$Z^A \rightarrow (Z-2)^{A-4} + (Z-4)^{A-8} \rightarrow \text{etc.}$$

This may not be the predominant decay mode, because at some point it must beta-decay, but nevertheless, one can estimate the energy of these alpha decays. Since the mass of an alpha particle is known, one can then calculate the mass of the succeeding nucleus. This can be done for each succeeding nucleus formed in the chain.

There can be only four such possible families of alpha-decay chains. This is because an alpha particle has a mass of four. That is, the starting material can be Z^A , Z^{A-1} , Z^{A-2} , or Z^{A-3} . At Z^{A-4} one is back in the Z^A chain again. Thus, one can calculate empirically the masses of all the isotopes above lead in terms of four known nuclei. These four chains can be joined together by any three well-measured beta-decay or neutron-capture processes which go from one chain to another. In this way, one can calculate all the isotopes in terms of a particular nucleus. This is precisely what was done. Long before the heavier elements were discovered their masses were approximated in this manner. Once the masses of all the elements are known, one can make a big two-dimensional grid of all the elements. From this chart one can look at pairs of isotopes which have the same mass number but differ in charge by one unit. The difference in mass is the beta-decay energy. By knowing the beta energy, one can then estimate the half life, i.e., the time it takes for the isotope to decay by beta emission. Also, by knowing the alpha energy one can estimate the alpha-decay half life. The two half lives can be compared to see if the isotope will decay predominantly by beta emission or alpha decay. As a general rule, the more energy that is available for a particular type of decay, the faster the decay occurs. Consider an element that can decay by either alpha or beta emission. If it has a high energy for beta decay, then its partial half life for beta decay is very small, say, one second. But if its partial half life for alpha decay were a year, which type of decay would one mainly see? The answer, of course, is the 1-sec beta decay. The total half life of a particular sample is the rate at which it disappears. Alpha and beta decay may be going on at the same time, but beta decay might be 10,000,000 times faster than the alpha decay, or vice versa. Consider, for example, U^{239} and neptunium 239. Although they have the same mass number, they do not have identical masses. It would be quite a coincidence if their actual masses were identical. They differ in charge by one unit. The heavier one is going to beta-decay to the other. If neptunium 239 were the heavier, it would electron-capture (which is a type of positive beta decay) to U^{239} . However, if U^{239}

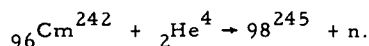
were the heavier, it would beta-minus-decay to neptunium 239. Of course, U^{239} is actually heavier so that it beta-minus-decays to neptunium 239, and this process is much more likely than the alpha decay of U^{239} . The partial alpha half life is longer than the partial beta-decay half life.

In the four or five years prior to 1949, complete charts were made of the heavy elements and estimates were made of their half lives, the type of decay, and the energy of the decay. By the end of 1949 Ghiorso, Thompson, and Seaborg were ready to look for elements 97 and 98. The first reaction looked for in 1949 was



An extremely small amount of americium was bombarded with alpha particles. The new activity that was found was named berkelium as an analogy to element 65, terbium, which is just above berkelium in the periodic table. Terbium was named after Ytterby, a town in Sweden where rare earths were first found.

The next reaction looked for by the same group (plus Dr. K. Street) was



It was predicted that ${}_{98}\text{Cf}^{245}$ would be an alpha emitter, that its half life would be between 30 minutes and 2 hours, and that the energy of the alpha would be between 7.0 and 7.3 Mev. As it turned out, ${}_{98}\text{Cf}^{245}$ is a 45-min alpha emitter with an energy of 7.1 Mev. Also, it was predicted that element 98 would come off at a certain place in the ion-exchange elution, and it did! In fact, the first two drops that were looked at were a new activity with the expected properties. Element 98 was named californium. It was not named by analogy. If it had been, it could have been called eupsorium, after dysprosium (element 66), which is above element 98 in the periodic table. Dysprosium means "hard to find." By contrast, element 98 was a cinch after the experiment on element 97, since so much about it could be predicted. The first time it was looked for, it was found. Therefore, a logical name might be eupsorium, meaning "easy to find."

By the time element 98 was discovered it was becoming quite obvious that to make any more new elements by the same procedures would be increasingly difficult. The target material could be obtained only after several years of waiting while the material was being irradiated in a pile. Then a very careful separation would have to be made. Therefore, it was decided to start with a more readily available target material and use a projectile with a higher Z, so that a bigger jump could be taken at one time. For example, element 98 could be made by bombarding uranium with carbon ions, which have a charge of six rather than two, as in the case of alpha particles. In fact, several new isotopes of element 98 were subsequently made in this way, using carbon ions from the 60-inch cyclotron. However, it is fairly difficult to accelerate carbon and get the conditions just right in the 60-inch cyclotron. Therefore, a new machine was built, a heavy-ion linear accelerator, whose purpose would be to produce heavy ions. This machine, which was completed in 1957, can accelerate ions as heavy as argon, element 18.

But this is getting ahead of our story. In the meantime, investigators used the old method, although it was time-consuming and difficult. They put plutonium in the Arco pile and left it there for several years. Gradually the plutonium captured neutrons and went to a very heavy plutonium, which then beta-decayed to the next element (95). This in turn continued to capture neutrons till it got to a very heavy isotope of americium, which finally beta-decayed forming element 96. This then captured neutrons until a beta-unstable isotope was produced which decayed to an isotope of element 97. This captured neutrons until it formed a beta-unstable isotope of 97 which decayed to element 98, and so forth until small amounts of some isotopes of elements 99 and 100 were produced. Figure O-2 shows this process, which took several years.

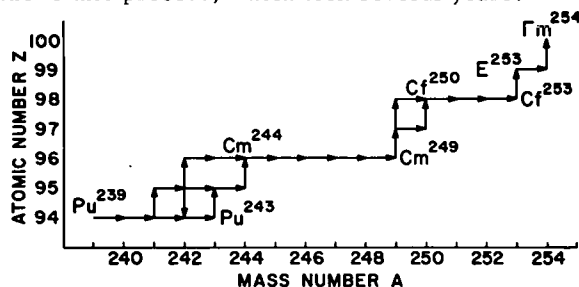


Fig. O-2. Production of heavy elements in a pile.

It does not appear that this process can be carried much further in a pile. An equilibrium is established between the production and decay of the nuclei, since their half lives become so short in this region. However, if the neutron flux could be greatly increased, heavier elements might be created. For example, if it takes 14 neutrons to get up to a certain heavy-element isotope, then the probability of making the element is proportional to the neutron flux to the 14th power. Thus, if the neutron flux is doubled, the chance of making the heavy element is improved by a factor of 2^{14} .

Now, where is the biggest neutron flux that one can think of? In a thermonuclear explosion, of course. Thus, elements 99 and 100 were discovered by accident in the big Mike explosion in 1952 in the Pacific. Drone planes flying through the clouds of the explosion gathered samples of the radioactive dust. These samples were brought back to the United States for chemical analysis. This was done at the Argonne National Laboratory, the Los Alamos Scientific Laboratory, and here at the Radiation Laboratory. Some very heavy plutonium isotopes, heavier than had ever been seen before, were found. Plutonium 244 was found, for example. In a bomb explosion a terrific neutron flux occurs for a very short period so that the neutrons all get captured before any beta decays can occur. Thus, one finds a distribution of heavy uranium isotopes ranging from normal uranium 238 to uranium isotopes in the 250's; these are very unstable and beta-decay quite rapidly. When they do beta-decay, they go to the heavy plutoniums that have never been seen before. From calculations, chemists suspected that some new heavy elements might also have been created. Therefore, hundred-pound amounts of debris were collected and analyzed. In this way elements 99 and 100 were discovered before they were created in laboratory experiments, but declassification procedures did not allow this fact to be published. Element 99 was named einsteinium in honor of Albert Einstein, and element 100 was named fermium in honor of Enrico Fermi.

By 1955 a very small amount (only a billion atoms) of element 99 became available as a target. It was made by long irradiation in the Arco pile. Such a tiny amount is not visible nor is it weighable. It was an extremely valuable sample, having a 20-day half life. To avoid wasting any, the einsteinium was plated as a very thin layer onto a thin gold foil. This was exposed to a very intense alpha-particle beam from the 60-inch cyclotron (Fig. O-3). After going through the gold foil, the alphas then hit the atoms-thick layer of einsteinium 253. In the process, the einsteinium picked up an alpha and went to element 101. The nuclei of element 101 that were formed then recoiled out of the target, owing to the momentum of the incoming alpha. These nuclei of element 101 were then collected on a second gold foil placed behind the target. After a few minutes of bombardment, this collector foil was taken out of the assembly and dissolved in aqua regia. The gold was solvent-extracted away and the remainder cleaned up completely on an ion-exchange column. An ion-exchange elution was then made in which it was possible to predict the position of element 101. From the dissolved foils of eight bombardments a total of 17 atoms of element 101 were found. This new element was named mendelevium in honor of Mendeléef, the father of the periodic table.

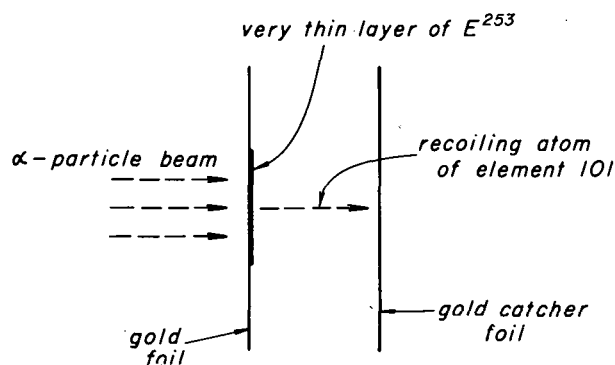


Fig. O-3. Target arrangement for producing element 101.

The final chapter of our story started in 1957 when a group of workers at the Nobel Institute in Sweden, working with colleagues from the Atomic Energy Research Establishment of England and the Argonne National Laboratory of the United States, reported element 102. They bombarded a curium target with carbon 13, the rarer form of carbon. They reported that element 102 has a half life of 10 minutes, and they identified it by an alpha decay of 8-1/2 Mev. They did this experiment several times, but only once were they able chemically to identify the activity as a new element. This experiment has been repeated many times at this Laboratory, but without being able to find this particular activity. The Swedish workers have also been unable to reproduce their results.

Last year (1958) Ghiorso, Sikkeland, Walton, and Seaborg of this Laboratory announced the discovery of 102. They made the isotope 102^{254} by bombarding curium 246 with carbon-12 ions from the heavy-ion linear accelerator. They identified it in a variety of ways, in particular, by chemical identification of the daughter. Element 102 alpha-decays to element 100, fermium 250. By chemically identifying the daughter and showing that it was, in fact, the alpha daughter of the unknown activity, they definitely proved the existence of 102. Its half life is 3 seconds, so that one cannot do too much chemistry with 102 itself.

This brings us to the end of our story and to the last difficulty, namely, that the half lives are becoming very short as we go to heavier elements. The reason for this is the coulombic repulsion of the protons in the nucleus. So many protons piled into a very small space repel each other very strongly, and the nucleus tends to blow apart. The two methods of disintegration available to the nucleus are (a) the emission of an alpha particle, and (b) a spontaneous fission. The time scale for spontaneous fission becomes increasingly short as more charges are added to the nucleus. For example, the partial half life of thorium (element 90) for spontaneous fission is of the order of a billion-billion years. For fermium (element 100) it is on the order of one year. For elements 101 and above, it becomes seconds and then microseconds. This means that one can create a heavier element, but it blows apart as fast as it is made. Thus, the chase is ending. A few more elements will be made and their properties will be interesting. For instance, element 103 will be the last of the heavy rare earths; it will correspond to element 71 in the rare earths. Element 104 will not be a rare earth, but instead will be the homolog of the plus-four family. That is, 104 will lie under hafnium (element 72), as indicated in Fig. O-1. Element 105 will be like tantalum (73), 106 like tungsten (74), 107 like rhenium (75), 108 like osmium (76), etc.

Bibliography

1. G. Friedlander and J. W. Kennedy, Nuclear and Radiochemistry (Wiley, New York, 1955).
2. A. Ghiorso and G. T. Seaborg, The Newest Synthetic Elements, Scientific American, 66 ff. (Dec. 1956).
3. S. Glasstone, Sourcebook on Atomic Energy, Second Edition (Van Nostrand, Princeton, 1958), Ch. XVI.
4. G. T. Seaborg and E. G. Valens, Elements of the Universe (Dutton, New York, 1958).
5. G. T. Seaborg, The Transuranium Elements (Addison-Wesley, Reading, Mass., 1958).
6. H. F. Wilson, Ion Exchange, Scientific American, 48 ff. (Nov. 1950).

P. HEALTH PHYSICS

Wade H. Patterson

The first radiation injury, which was later identified as an x-ray burn, occurred a few months before the discovery of x-rays was announced by Becquerel in 1895. However, not until 1928 were procedures established for insuring adequate protection. Prior to that, most of the regulations were concerned with the number of hours a person could work per day or per week, but not with the actual exposure he could safely receive or with the shielding necessary to protect him.

In 1928 the International Commission of Radiation Protection was formed, growing out of the International Congress of Radiologists. A year later the National Commission on Radiation Protection was created. From these commissions came rules concerning safe levels of radiations exposure, as well as protective procedures to be followed.

During World War II when the Manhattan District first began producing large quantities of radioactive materials with the chain-reacting pile, it became necessary to have an organized group within the project responsible for radiation control. Originally, there were a health division, a health instrument section, a health biology section, and a health physics section. Eventually these coalesced into one group called health physics.

Today health physics is the term most often used to describe personnel at Atomic Energy Commission laboratories engaged in radiation safety. The Public Health Service generally has the responsibility for protecting the public at large from the harmful effects of radiation. At this Laboratory there is a unique division of responsibility between the health physics section and the health chemistry section. The former is responsible for the personnel monitoring program and for controlling accelerator-produced radiation. The health chemistry group is responsible for the control of radiation from radioactive sources, for decontamination, for waste disposal, and for radioisotope control. This division of responsibility exists at no other laboratory.

Units of Radiation Measurement

Certain units of radiation measurement must be defined in order to discuss radiation exposure quantitatively. The units most commonly used are:

a. The curie, which is defined as 3.7×10^{10} disintegrations per second (d/sec). Originally this definition was different, being the disintegration rate of 1 gram of radium in equilibrium with its decay products. This was an unsatisfactory definition owing to certain small corrections that had to be applied.

b. The roentgen, which is defined as one electrostatic unit of charge (+ or -) in 1 cm^3 of air at standard conditions of temperature and pressure. The roentgen has also undergone a modification since its first definition, which originally was in terms of x-ray radiation only. The present definition is also applicable to other types of radiation.

c. If one is concerned with the effect of radiation on tissue, an appropriate unit is the roentgen equivalent physical (rep), since the roentgen is a measure of charge in air, not tissue. The rep is defined as a dose of 90 ergs of radiation absorbed or liberated per gram of tissue. The figure 90 ergs/g is in dispute, since it is based on a calculation involving the energy required to produce an ion pair in air, this value being uncertain.

d. Consequently, another unit, the rad, has been defined as the amount of radiation that results in the absorption in tissue of 100 ergs/g. The rad is now used almost exclusively in place of the rep. Neither the rep nor the rad indicates that for a given energy some radiations are more effective than others in producing the same biological effect.

e. Therefore, it is necessary to introduce a another unit called the RBE, or relative biological effectiveness. The RBE is the ratio of a γ -ray (or x-ray) dose to the dose required to cause the same biological effect by the radiation in question. Both doses are measured in either rep or rad. For example, consider a case where a certain weight loss is produced in a mouse, first by x-rays and next by neutrons. With both exposures measured in rad (or rep), the ratio of doses would turn out to be about 2 to 1, giving the relative biological effectiveness of neutrons as $\text{RBE} = 2$. In general, the values of RBE lie between 2 and 10 for various radiations, depending on the particular organism being studied and on the length of exposure.

f. Another unit, the roentgen equivalent man (rem), is that quantity of radiation which, when absorbed by man, produces an effect equivalent to the absorption of 1 roentgen of x or γ radiation. The rem is found by multiplying the rep by the RBE. There is considerable argument as to whether one can really determine what the rem is, i.e., what the true biological effect of a dose is, since the value of the rep itself is uncertain.

g. Flux is defined as the number of particles incident on an area of one square centimeter each second.

Types of Radiation

The biologically important radiation from radioisotopes consists of α , β , and γ rays, while that from high-energy accelerators is largely neutrons. Figure P-1 shows how the first three of these radiations are influenced in a magnetic field. Gamma rays have a long range and are undeflected; β particles have a shorter range and are deflected only to a certain degree; α particles have a very short range, are deflected to the greatest degree, and are deflected in the opposite direction to β particles (indicating that α and β particles have opposite charges).

The range of alpha particles from a natural source is so short that they cannot penetrate one's skin. They are stopped by a piece of paper or by a few centimeters of air. However, if alphas are

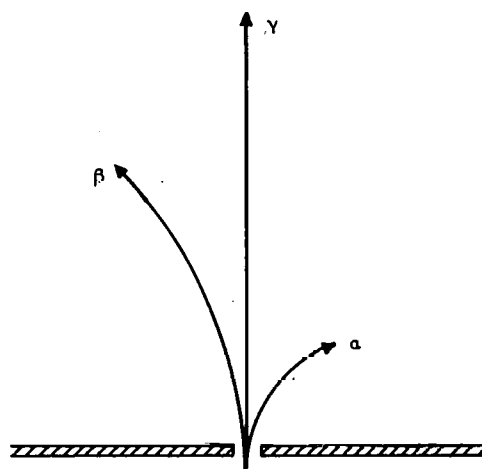


Fig. P-1. Behavior of α , β , and γ radiations in a strong magnetic field, the field being directed into the page.

ingested they are quite damaging, owing to their intense ionization. Beta particles from natural sources have a range of about 1 meter in air, or 1-2 cm in aluminum. They can penetrate a few millimeters of tissue, producing burns. Gamma rays, x-rays, and neutrons, on the other hand, can penetrate quite far into tissue, causing serious consequences.

Radiation Levels

Table P-I gives the levels of radiation to which people in this country are normally exposed. The wide variation in the radium (Ra^{226}) exposure is due to the fact that the concentration of radium is not uniform throughout the country. Also the cosmic ray exposure depends on the altitude at which one lives. At an altitude of 10,000 feet the dose is four times that at sea level.

	Internal (milli-rem/yr)			External (millirem/yr)
Natural sources	K^{40}	25	cosmic rays radioactivity in environment	35 (sea level) - 60 (5000 ft)
	C^{14}	1		35-70
	Ra^{226}	5-350		
Man-made sources	Sr^{90}	1-5	luminous watch	40
			medical and dental x-rays	100
			fallout	1
			TV screen	<1

In regard to man-made radioactivity, we see from Table I that the exposure caused by medical or dental x-ray machines is the largest. That due to weapons testing may reach 6 or 7 millirem per year, because fallout is occurring somewhat faster than originally expected.

Maximum Permissible Exposure

The permissible weekly occupational exposure, as set by the Atomic Energy Commission for its workers, is 0.1 rem. The yearly exposure must not exceed 5 rem. These are values for radiation

received over the whole body. The permissible exposures to the general population are 1/10 of these values. The amount of radiation, as measured in rep, to produce a 0.1-rem exposure depends upon the type of radiation. Table P-II lists the equivalent exposures in rep and rem for various radiations.

Radiation	rep	x	RBE	= rem
X- and gamma rays	0.1		1	0.1
Betas	0.1		1	0.1
Protons	0.01		10	0.1
Alphas	0.005		20	0.1
Fast neutrons ^a	0.01		10	0.1
Thermal neutrons ^a	0.03		3	0.1

^aThe maximum permissible flux of neutrons varies with neutron energy.

The permissible occupational exposure levels are generally set so that about ten times that amount of radiation is required to produce an observable effect. This explains why, as detecting schemes have become more sensitive, the permissible levels have been revised downward. In 1928 the safe occupational exposure was set at 72 roentgens per year; in 1940 this was revised to 30 r/yr, and in 1945 to 15 r/yr. It is now set at 5 rem/yr.

1 rem	No detectable change
10 rem	Blood changes are detectable
100 rem	Some injury - no disability
200 rem	Injury and some disability
300 rem	Injury and disability
400 rem	50% deaths in 30 days ^a
600 rem	100% deaths in 30 days ^a
10,000 rem	50% deaths in four days ^a
100,000 rem	Quick death ^a

^aThe deaths at 30 days are due to intestinal damage, the deaths at 4 days are due to blood-cell damage. The quick deaths are due to nerve damage.

Table P-III shows the biological effect of a given full-body exposure received in a few days. It is evident that a 0.1-rem exposure in one week is well below the level required to produce any observable effect. The biological experience on which this evaluation of radiation damage is based comes from four sources:

a. There is a large body of human experience with x-ray and gamma-ray treatments. There have been numerous cases of radiation injuries, especially in the earlier days. The earlier data are often difficult to evaluate because the dose given was usually not measured. In recent years much careful work has been done with x-rays and gamma-rays under conditions where the dose is known. It would seem that in view of the large number of x-ray films exposed for diagnostic purposes and with a large amount of therapeutic data available that the effect of x- and γ -rays on humans would be well known. This is partially true for fairly large doses where immediate damage or injury result. However, because of other factors in the environment which affect human health, it is not easy to evaluate long-term effects caused by small chronic doses.

b. There are good data available on a particular specialized type of occupation involving radiation exposure: that of painters of radium watch and clock dials. During the first World War when luminous dials were first introduced, many of the individuals who painted these dials got radium paint into their mouths. This radium lodged in their bones, and (in some cases) bone cancers eventually developed. It was possible after autopsy to determine the amount of radium present and to relate this to the amount of damage. This information applies especially to radioactive materials whose chemistry causes them to be deposited in the bones, and it applies principally to alpha-particle emitters. Although many died of bone tumors, it is interesting to note that there were no cases of leukemia among the radium-dial painters, since leukemia is usually thought of as a radiation-induced disease of the blood-forming tissue in the bone marrow. At present, because of this unfortunate experience with radium, adequate precautions are taken, and injuries of this type are quite rare.

The miners in Schneeberg and Joachimsthal in Central Europe who mine uranium ore in tunnels provide another example of occupational radiation exposure. They are known to have suffered lung damage owing to breathing too high a concentration of radon gas, which is emitted by the radioactive ores. The data obtained from these mining operations are again quite specialized, since it is very unusual for a situation to arise where a person can breathe a high concentration of a radioactive gas.

c. There is a fair body of information that comes from Hiroshima and Nagasaki. Fifteen to twenty percent of the injuries suffered in the Hiroshima and Nagasaki bombings were caused by gamma rays or neutrons. In spite of the large number of people involved, there is a relatively small number for whom the dose is known with sufficient accuracy to make it possible to evaluate the effects of a large dose delivered in a short time. The data from Hiroshima and Nagasaki, meager though they may be, appear to agree with data from other sources. It is very difficult to separate blast damage and radiation damage in the Japanese cases, and it is also difficult to obtain a satisfactory control population for comparison purposes.

d. In addition to the above sources of human information, there is a large body of experience that has been gained with animals. The most useful animals have been the mouse, rat, and dog, although work has been done with some of the larger domestic animals. The principal limitation of this work is that it applies to higher dose rates than those for which we wish to have data. Even though animals have a much shorter life span, it is still very laborious to conduct an experiment with low dose rates and sufficient precision to give any useful data.

Shielding

Various materials, such as lead, concrete, or earth, are commonly used as shielding. However, probably the simplest medium to interpose between oneself and a source is distance, since the intensity of the radiation falls off as the square of the distance. A useful relation is

$$\frac{I_0}{I} = \frac{d^2}{d_0^2} \quad (1)$$

where I_0 is the radiation intensity measured at a distance d_0 , and I is the intensity at a distance d , assuming that there is no absorption or scattering in the medium. Consider a standard radiation source whose intensity I_0 at 1 cm is 8.4 roentgens/hr. A detector 1 meter from this source should indicate an intensity I of

$$I = \frac{I_0 d_0^2}{d^2} \\ = \frac{8400(100)^2}{1}$$

$$I = 0.84 \text{ milliroentgen / hr.}$$

ignoring such things as scattering and absorption. In practice, one cannot ignore these, but must build the detector in such a way that it gives a true counting rate.

Another relation, which is useful when dealing with neutron flux, is

$$\phi = \frac{Q}{4\pi r^2} \quad (2)$$

where ϕ is the flux at a distance r from a source whose emission is Q . A neutron detector placed 50 cm from a plutonium-beryllium source having an emission of 1.75×10^6 neutrons/sec should indicate a flux ϕ of

$$\phi = \frac{1.75 \times 10^6}{4\pi (50)^2}$$

$$\phi = 50 \text{ neutrons/cm}^2/\text{sec.}$$

In considering the thickness of shielding required for personnel protection, one must determine the range of the most penetrating particle expected from the particular accelerator or radioactive source, by the method outlined in Section C, "Nuclear Physics." The thickness of shielding is generally increased by a factor of at least two, to provide an adequate margin of safety.

Bibliography

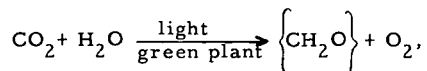
1. Biological Effects of Atomic Radiation. A Report to the Public (National Academy of Science, National Research Council, Washington, 1956).
2. S. Glasstone, Sourcebook on Atomic Energy, Second Edition (Van Nostrand, Princeton, 1958), Ch. XIX.
3. Miner, Shackleton, and Watson, Teaching With Radioisotopes (U.S. Atomic Energy Commission, Washington, 1959). (Available from Superintendent of Documents, U.S. Government Printing Office, Washington 25, D.C.; price 40 cents.)

4. Radioisotopes — Uses, Hazards, and Controls
(Isotope Extension, U.S. Atomic Energy
Commission, Oak Ridge, Tenn.) (Free upon
request.)
5. E. Spoerl, The Lethal Effects of Radiation,
Scientific American, 22ff (Dec. 1951).

Q. PHOTOSYNTHESIS*

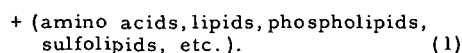
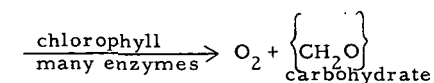
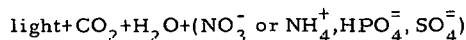
J. A. Bassham

The process of photosynthesis in plants is often defined by the over-all equation



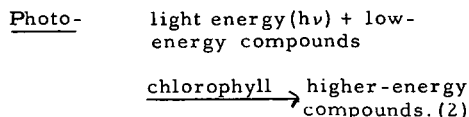
which stated in words says that carbon dioxide and water "under the influence of light" are converted by green plants to carbohydrate and oxygen. Such a definition, when used to introduce the subject of photosynthesis to a general chemistry class, provides a much too incomplete description of this important process. How might the definition be improved?

First, the essential energy-storing character of the reaction should be emphasized by showing light energy as a reactant in the equation, thus distinguishing photosynthesis from other photochemical reactions in which the role of light is essentially catalytic. Secondly, in view of present day knowledge of the direct photosynthesis of products other than carbohydrates it should be stated that the reactants include nitrate (or ammonia), sulfate, and phosphate; and the products include amino acids, fatty acids, phospholipids, and sulfolipids. In addition, a number of other inorganic substances (Fe, Mg, etc.) are required in varying amounts for photosynthesis and may in some cases be incorporated into the products. Finally, if catalysts are to be indicated, then chlorophyll and the many enzymes required should be shown. An equation for the over-all process of photosynthesis might then be given as

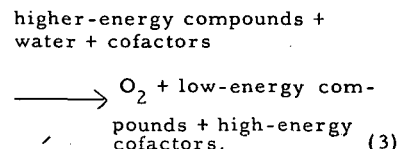


Any generalized formula for photosynthesis is a compromise between simplicity and accuracy, but this one at least hints at the complexity of this important process.

For the purpose of discussion, the photo-synthetic reaction may be conveniently broken down into three stages. The first of these is the "photo-" part which consists of the absorption of light by the plant pigments and the conversion of the light energy into the stored energy of new chemical bonds of "high-energy" compounds:

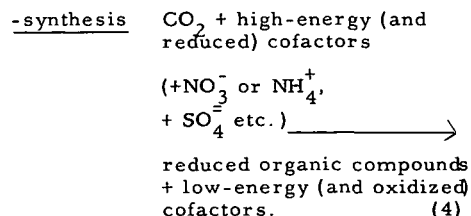


An intermediate stage then follows in which the chemical energy of these primary products of the photochemical reaction is utilized in the breaking of the O-H bonds of water and the conversion of two enzymatic cofactors to their more energetic and (for one of them) more reduced forms:



At the same time molecular oxygen is liberated. We know very little of the detailed mechanism of these reactions and all or part of this so-called intermediate stage may be in fact a part of the primary photochemical reactions. Indeed, the first stages of photosynthesis may best be considered as the sum of one primary photochemical act in which light energy is absorbed by chlorophyll and a series of non-photochemical steps in which this energy is efficiently transferred through the production of increasingly stable chemical species.

The "-synthesis" stage of photosynthesis is the sum of a great number of enzymatic reactions in which the two cofactors supply the reducing power and energy necessary to bring about the conversion of carbon dioxide (together with nitrate, phosphate, and sulfate) to carbohydrates, amino acids, and lipids:



Let us consider in more detail these several stages in photosynthesis. All light energy that is utilized for photosynthesis appears to be employed in one early stage to raise the energy content of chlorophyll a from its ground state to a higher energy level or "excited state." This is true whether the light energy is absorbed directly by chlorophyll a, or by other plant pigments, in which case the energy appears to be transferred to chlorophyll a.

Then the energy of the excited chlorophyll a must somehow be used for the formation of stable compounds with a high content of energy stored in their chemical bonds. There is no unanimity of opinion as to how this is accomplished. One view of the mechanism is illustrated in general terms by the scheme shown in Fig. Q-1.

* A longer version of this paper is to appear in the Journal of Chemical Education.

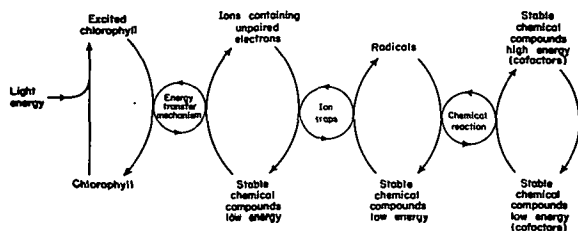


Fig. Q-1. Mechanism of conversion of light energy to chemical energy.

In this scheme the energy of the excited chlorophyll is transferred by some mechanism, which probably involves its migration from one chlorophyll molecule to another, through a closely packed sheet of chlorophyll molecules. This sheet of chlorophyll molecules, which is thought to be only one molecule thick, may exist at the interface between lipid and protein layers, the presence of which has been revealed by electron microscope pictures of the chloroplasts. (Chloroplasts are the subcellular units of green cells responsible for photosynthesis.) Eventually, (after a millionth of a second or so) the energy packet or "exciton" reaches a point where it is used to supply the energy of ionization of some unknown compound, perhaps chlorophyll itself. This ionization is believed to consist of transfer of electrons in such a way that the resulting ionic species contain "unpaired electrons."¹ In turn, the ions undergo reactions leading to the formation of somewhat more stable radicals, in this theory. Finally, these radicals may form yet another stage in the transfer of stored energy of chemical bonds.

Eventually, stable chemical compounds of high energy content must be formed. In this scheme they would be made by the interaction of stable chemical radicals. The resulting chemical compounds may be the cofactors required for the reduction of carbon dioxide and nitrate, or they may be other energy-containing compounds which subsequently give rise to the necessary cofactors by chemical or enzymatic reaction. In addition, molecular oxygen is evolved from the products formed by the breaking of the O-H bond of water.

The two energy-carrying cofactors, triphosphopyridine nucleotide (TPNH) and adenosine triphosphate (ATP), are shown in Fig. Q-2.

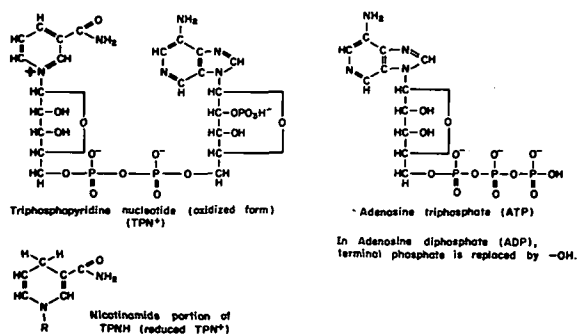
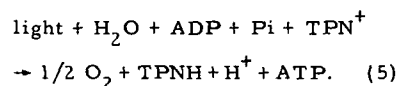


Fig. Q-2. Formulae of TPNH and ATP.

The requirement for these coenzymes was predicted from studies of the carbon reduction cycle, certain

steps of which can only be carried out by the enzymes if the appropriate cofactors are present. TPNH is an electron carrier, transporting electrons obtained by splitting water to the reactions in which they are used for the reduction of carbon dioxide. Each time a molecule of TPNH is oxidized to TPN^+ , two electrons are transferred. The energy stored in ATP is stored through the formation of an anhydride from ADP and inorganic phosphate.

An important stoichiometric relationship between the formation of ATP, TPNH, and O_2 during photosynthesis has been discovered by Arnon and coworkers,^{2,3} as given by



This relationship was discovered in studies with isolated chloroplasts from spinach leaves when oxygen evolution, TPNH formation, and ATP formation were measured simultaneously. This stoichiometric relationship does not tell us what the mechanism of formation of these substances is but does perhaps hint at a pathway that is simpler than we formerly suspected.

We will now turn to the "synthesis" part of photosynthesis and see how these cofactors which have been formed by means of the light reactions are used. The studies of the pathway of carbon reduction during photosynthesis were in a large part carried out in the Lawrence Radiation Laboratory at the University of California by Professor Melvin Calvin and co-workers.⁴ These studies provide an interesting example of the use of radioactive tracer atoms in the elucidation of a complex biochemical pathway. Since carbon dioxide is the sole source of carbon for the photosynthetic reaction, radiocarbon (carbon fourteen) may be introduced very easily into photosynthesizing plants, in the form of C^{14}O_2 or, for aquatic plants, $\text{HC}^{14}\text{O}_3^-$ ion.

Let us consider a simple experiment with a suspension of the algae, *Chlorella pyrenoidosa*, which have been very extensively used in these studies. These green, microscopic unicellular plants, suspended in water containing a few inorganic substances (nitrate, phosphate, etc.) with a stream of C^{12}O_2 (ordinary carbon dioxide), photosynthesize at a rapid rate if illuminated from each side in a thin glass vessel. The CO_2 is continually taken up from the solution (where it is in equilibrium with bicarbonate ion) and converted by the photosynthetic plant through a series of biochemical intermediates to various organic products.

A solution of radioactive bicarbonate, $\text{HC}^{14}\text{O}_3^-$, is suddenly introduced into the algae suspension. The plant does not distinguish in any important way between the C^{12} and the C^{14} (which are chemically identical), and immediately some of the C^{14} is incorporated into the first of the biochemical products leading from CO_2 to end products. As time passes, the C^{14} would be passed on to subsequent intermediates in the chain. After a few seconds of exposure to the

¹ Sogo, Pon, and Calvin, *Proc. Nat. Acad. Sci.* **43**, 387 (1957).

² Arnon, Whatley, and Allen, *Science* **127**, 1026 (1958).

³ Arnon, Whatley, and Allen, *Nature* **180**, 182 (1957).

⁴ Bassham, Benson, Harris, Wilson, and Calvin, *J. Am. Chem. Soc.* **76**, 760 (1954).

C^{14} , the suspension of algae is run into methanol to a final concentration of 80% methanol. This treatment denatures all the enzymes instantly and freezes the pattern of intermediates toward further change. Now all that remains to be done is to analyze the dead plant material for radioactive compounds to see what are the first products of carbon reduction during photosynthesis.

The first step in this analysis is to prepare an extract of the soluble compounds, since the early products of carbon reduction have been found to be simple soluble molecules. This extract is then concentrated and analyzed by the method of two-dimensional paper chromatography.

The first step in this useful method of analysis consists of drying a small quantity of the concentrated extract on the corner of a large sheet of rectangular filter paper.⁵ The edge of the paper next to this dried extract is then placed in a trough filled with a suitable chromatographic solvent which travels down the paper by capillarity. As the solvent passes over the dried extract, it dissolves the various compounds and carries them along with it at various rates, depending upon the particular compound.

The substances are thus separated in a row along the edge of the paper. The paper is dried, turned 90°, and the edge next to the compounds placed in a second solvent, where the process of separation is continued with the compounds moving in a direction at right angles to their former direction of travel. Upon drying the paper again, the compounds are found to be scattered about the paper as pure substances, each in its own unique location. The importance of the method for these studies can be seen from the fact that it permits the analysis of a few micrograms or less of dozens of different substances in a single simple operation.

Of these many compounds, those into which the plant incorporated carbon fourteen during its few seconds of photosynthesis with $HC^{14}O_3$ are radioactive and emit the particles resulting from radioactive decay of the C^{14} . In this case these are β^- particles, which may be detected by the fact that they expose x-ray film. Thus, if a sheet of x-ray film is placed in contact with the paper chromatogram, subsequent development of the film will show a black spot on the film corresponding to the exact shape and location of each radioactive compound on the paper. A quantitative determination of the amount of radio-carbon in each compound may then be made by placing a Geiger-Muller tube with a very thin window (to permit the particles to pass through) over the radioactive compound on the paper and counting the emitted particles electronically.

The next stage in the method of radiochromatographic analysis is the identification of the radioactive compounds. This identification is accomplished in a variety of ways. Of these, the most important is by elution or washing of the compound off the paper and the determination of such chemical and physical properties of the substance as can be measured with a solution of a few micrograms or less of the material.

The properties then are compared with those of known compounds. The final check on the identity of the compound is frequently made by placing on the same spot on filter paper the radioactive compound and 10 to 100 μg of the pure nonradioactive substance with which the radioactive compound is thought to be identical, and chromatographing the two together. A radioautograph is then prepared to locate the radioactive substance, after which the paper is sprayed with a chemical spray (for example, ninhydrin for amino acids), which produces a color where the carrier compound is located on the paper. Superposition of the paper chromatogram and the radioautograph (x-ray film) will show an exact coincidence between chemically-developed color on the paper and the black spot on the film if the two substances are identical.

Once the identity of the radioactive compounds formed during a short period of photosynthesis had been established, experiments were performed under a variety of conditions and times of exposure of the algae to radiocarbon. The radioautograph from the experiment with *Chlorella* described above is shown in Fig. Q-3.

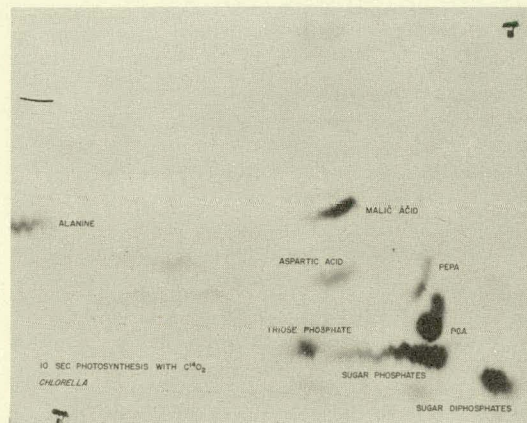


Fig. Q-3. Radioautograph of *Chlorella* extract after 10 seconds' photosynthesis with $C^{14}O_2$.

Even after only 10 seconds of exposure to C^{14} , a dozen or more compounds are found. Some of these (the sugar phosphates) are not separated from each other by the first chromatography and must be subjected to further analysis. When the sugar monophosphates are hydrolyzed to remove the phosphate groups and rechromatographed, separate spots are found of triose (dihydroxyacetone), tetrose, pentoses (ribulose, xylulose and ribose), hexoses (glucose and fructose), and heptose (sedoheptulose). The sugar diphosphates are found to include ribulose, fructose, glucose, and sedoheptulose.

After periods of photosynthesis with C^{14} of less than 5 seconds, 3-phosphoglyceric acid (PGA) was found to be the predominant radioactive product. Chemical degradation of this compound showed that the radioactivity first appears in the carboxyl carbon. Later kinetic studies showed that the rate of incorporation of C^{14} into PGA at very short times was much greater than the rate into any other compound. Therefore, it was concluded that PGA is the first stable product of carbon-dioxide reduction during photosynthesis, and furthermore, that carbon dioxide first enters the carboxyl group of PGA, presumably via a carboxylation reaction.

⁵ Benson, Bassham, Calvin, Goodale, Haas, and Stepka, J. Am. Chem. Soc. 72, 1710 (1950).

⁶ Block, Durrum, and Zweig, *Paper Chromatography and Paper Electrophoresis*, 2nd Ed., (Academic Press, New York, 1958) p. 710.

From this point we borrowed from the already known pathways of the glycolytic breakdown of sugars which lead to PGA as an intermediate. Noticing that the sugar phosphates are important early products of carbon reduction in photosynthesis, we proposed that they are formed from PGA by a reversal of the glycolytic pathway. Degradation of the radioactive hexoses from short experiments showed that they were labeled in the two center carbon atoms (numbers 3 and 4) just as one would expect if two molecules of PGA were first reduced to triose and then linked together by the two labeled carbon atoms to give hexose (Fig. Q-4).

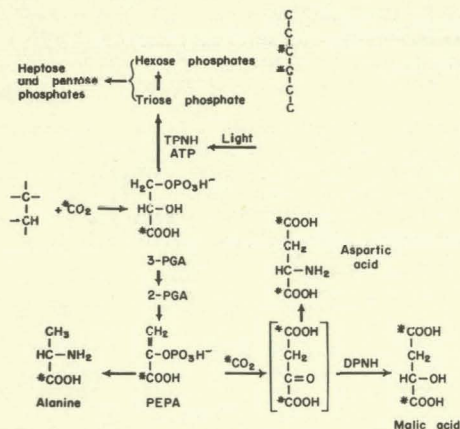


Fig. Q-4. Pathways of $C^{14}O_2$ incorporation in early stages of photosynthesis.

The hexose and triose phosphates may be converted by aldolase or transaldolase and transketolase enzymes to pentose and heptose phosphates. Degradation of these sugars and comparison of the labeling patterns within the molecules showed that this conversion did occur and in such a way that five molecules of triose phosphate were ultimately converted to three molecules of pentose phosphate.

Other known metabolic pathways leading from PGA (Fig. Q-4) give rise first to phosphoenolpyruvic acid (PEPA), which then may undergo further transformations as follows: (1) it may be carboxyllated and transaminated to give aspartic acid, (2) it may be carboxyllated and reduced to give malic acid, (3) it may be dephosphoryllated and transaminated to give alanine. All of these compounds are labeled after short exposures of the algae to $HC^{14}O_3$ in the light. The carboxylation of PEPA, while a second point of entry for carbon dioxide during photosynthesis, accounts for only 5 to 10% of the reduced carbon dioxide under normal conditions in *Chlorella*, according to kinetic measurements. Since this reaction converts three additional carbon atoms from PGA to malic acid and aspartic acid at the same time, the amounts of these compounds photosynthesized are significant.

The enzyme system which brings about the oxidation of triose phosphate to PGA in the glycolytic pathway was known to produce ATP and TPNH (or DPNH). If PGA is to be reduced to triose phosphate during photosynthesis, it follows that ATP and TPNH must be supplied. We have already seen that these two cofactors are produced as a consequence of the light reaction and the splitting of water. It might be expected that if the light were turned off from plants photosynthesizing

in ordinary carbon dioxide at precisely the same time that $C^{14}O_2$ is introduced, PGA would no longer be reduced to sugar phosphates, but would still be formed (if neither TPNH or ATP are required for the carboxylation reaction) and would still be used in other reactions not requiring these cofactors.

In Fig. Q-5, the radioautograph from just such an experiment, we see that this prediction was correct.

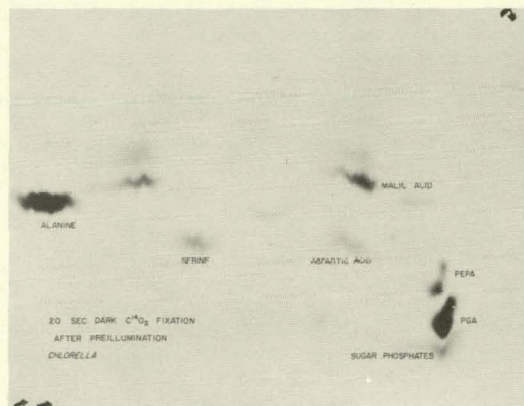


Fig. Q-5. Radioautograph of *Chlorella* extract after 20 seconds with $C^{14}O_2$ in the dark.

Labeled PGA is still formed from $C^{14}O_2$ during 20 seconds in the dark, but only a very little of the PGA is reduced to sugar phosphates. On the other hand, a large amount of alanine is formed from PGA via PEPA in reactions which do not require ATP or TPNH. The small amount of labeled sugar phosphates that does appear is due to the residual ATP and TPNH which was formed while the light was on but which had not yet been used up when the $C^{14}O_2$ was introduced. Some malic acid is still formed in the dark, indicating the presence of some DPNH, either remaining from the light or derived from some other metabolic reaction.

Thus far we have not discussed the identity of the source of the compound which undergoes carboxylation to produce PGA. In order to explain its discovery, we must turn to another type of experiment with $C^{14}O_2$ and photosynthesizing algae. In these experiments, algae are first permitted to photosynthesize for 20 minutes or more in the presence of a constant supply of $C^{14}O_2$. During this time all environmental conditions are maintained constant (temperature, CO_2 pressure, light intensity, etc.). After about 10 minutes of exposure to $C^{14}O_2$, so much radiocarbon has passed through the various biochemical intermediate compounds on its way to end products that each carbon atom of each intermediate compound contains, on the average, the same percentage of carbon-fourteen atoms as the CO_2 which is being absorbed. In other words, the specific radioactivities of all the carbon atoms of all the early intermediates are the same as the specific radioactivity of the entering radiocarbon, which can be measured and is therefore known.

At this point, samples of the algae are removed without disturbing the rest of the algae and these samples are killed and subsequently analyzed by the methods already described. The total radioactivity of each intermediate is measured, and when this is divided by the known specific radioactivity, the total number of carbon atoms of each intermediate com-

pound in the sample can be calculated. Thus, the concentrations of the various intermediates of the actively photosynthesizing system may be determined.

This determination of the concentrations of intermediate *in vivo* is an extremely valuable tool which has many uses, but let us proceed with the experiment which we had started to describe above. Having taken samples of algae for later determination of the concentrations of intermediate compounds, we now turn off the light and proceed to take a series of samples of the algae as rapidly as we can, which is about every three seconds. When the concentrations of intermediate compounds in these samples are determined, any changes resulting from turning off the light will be revealed. The two most striking changes are found to be in the concentration of PGA which increases rapidly (Fig. Q-6) and in concentration of one particular sugar, ribulose diphosphate, which drops rapidly to zero.

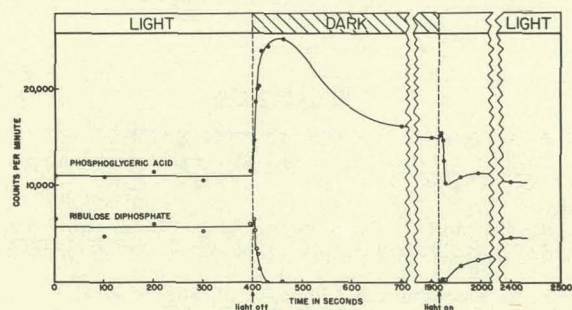


Fig. Q-6. Light-to-dark concentration changes in PGA and ribulose diphosphate.

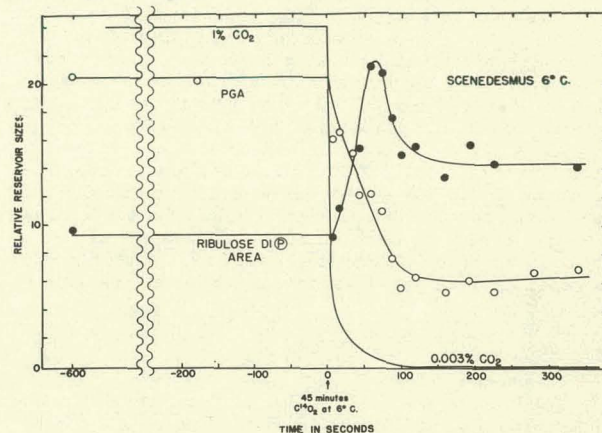
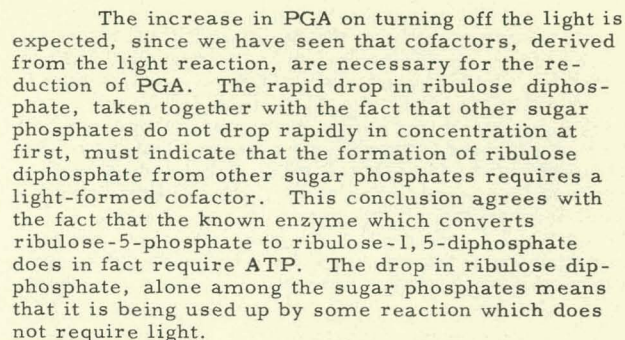


Fig. Q-7. High-CO₂ to low-CO₂ concentration effect on PGA and ribulose diphosphate concentrations.

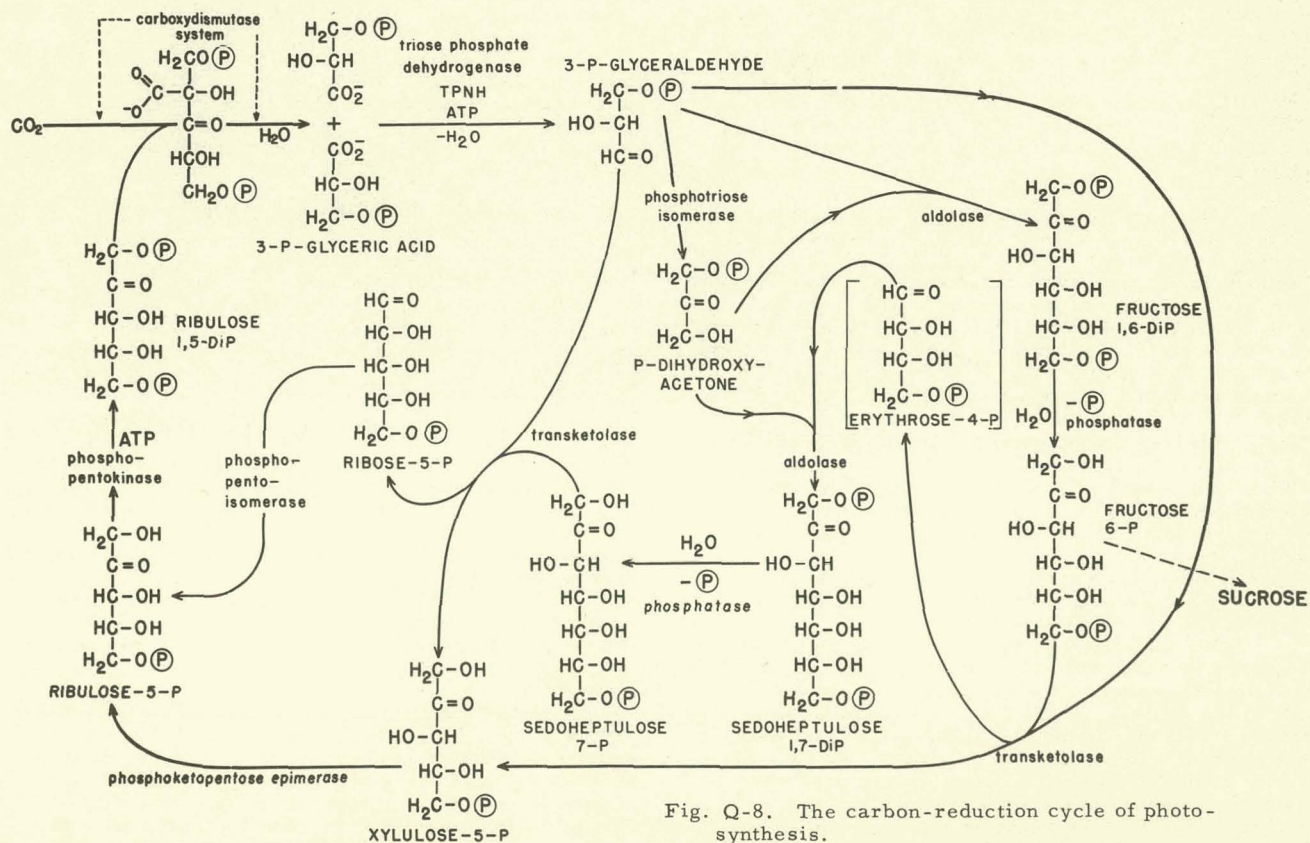


Fig. Q-8. The carbon-reduction cycle of photosynthesis.

If ribulose diphosphate is used by some reaction that proceeds in the dark, and if PGA continues to be formed in the dark, could the carboxylation of ribulose diphosphate to form PGA be the first step in carbon-dioxide reduction? To answer this question, another experiment similar to the one just described was performed. This time, however, instead of turning off the light, the light was left on and carbon dioxide was suddenly removed. The result of this experiment (Fig. Q-7) confirmed the idea of a carboxylation of ribulose diphosphate, for the concentration of ribulose diphosphate now rose rapidly while PGA dropped rapidly.

The carbon reduction cycle was now complete and is shown in Fig. Q-8. The methods and reasoning leading to the elucidation of this cycle have been presented more completely elsewhere.^{4, 7} A complete circuit of the cycle involves the conversion of five molecules of triose phosphate to three molecules of ribulose diphosphate upon which carboxylation produces six molecules of PGA or its equivalent. There is thus a gain of one three-carbon unit corresponding to the three molecules of carbon dioxide introduced for each turn of the cycle. On the average, five and a fraction of the three-carbon units are reduced to triose phosphate, with five of these going to generate ribulose diphosphate while the fraction is

employed (presumably) in the synthesis of sucrose, polysaccharides, glycerol, and galactose (constituents of lipids) and other substances. The remainder of the PGA is converted to alanine and serine or, via PEPA, is carboxylated to make four-carbon compounds (Fig. Q-4).

In conclusion, it may be said that through the employment of tracer elements, particularly carbon fourteen, the pathway of carbon reduction during photosynthesis has been mapped. From the nature of this pathway, the requirements for energetic cofactors, derived ultimately from the light reaction, have been established. The detailed mechanism by which these cofactors are formed in the light reaction, as well as the mechanism by which water is split and O_2 is evolved, is not as yet known, but studies with isolated chloroplasts have demonstrated a stoichiometric relation between the evolution of oxygen and the formation of both cofactors, ATP and TPNH.

Bibliography

1. J. A. Bassham and M. Calvin, The Path of Carbon in Photosynthesis (Prentice-Hall, Englewood Cliffs, New Jersey, 1957).
2. R. Hill and C. D. Whittingham, Photosynthesis, Revised Edition (Wiley, New York, 1957). (general)
3. E. I. Rabinowitch, Photosynthesis, 3 vols. (Interscience Publishers, New York, 1956). (extensive compilation)

⁷J. A. Bassham and M. Calvin, The Path of Carbon in Photosynthesis (Prentice-Hall, Englewood Cliffs, New Jersey, 1957).

R. THE CHEMISTRY AND PHYSICS OF VIRUS

Robley C. Williams

Dr. Stent's lecture on the biology of viruses will cover the problem of how one tests for virus infection, how virus is attached to cells, the intracellular process of virus development, and viral genetics. This lecture will consider viruses essentially as particles, not as infectious agents. That is, we will think of them as being dead rather than alive, if we may use those two distinguishing terms. This may give a better basis for understanding the greater complexity involved in the biology of viruses. In other words, we are going to pretend that we are chemists and physicists working with viruses.

Viruses are characterized according to what type of cells they can inoculate. That is, there are animal, bacterial, plant, and insect viruses. Upon inoculation something happens to the cell. By and large, we say that disease results if the cell is somehow different than it was before inoculation took place. The cell may die; it may die rather dramatically; it may die a lingering death; it may just have a bit of malaise and recover; in some cases the cell may actually be happy about the virus infection. The latter is true of cancer viruses. If one infects a tissue culture with a cancer virus, the cells do not become deranged and sicken and wither away; on the contrary, they enjoy the experience and grow much more rapidly and healthfully. As a result they enlarge and crowd out other cells.

Virus structure is important from the medical point of view. Eventually, if we know enough about viruses and their mechanism of reproduction we may be able to do something about virus diseases. At present (all advertisements to the contrary) there is no therapeutic agent against virus disease. There is nothing comparable to the antibiotics that have proven so successful with bacterial diseases. One can see a doctor if he wants to; but more importantly he should go to bed and rest, hoping for the best as far as specific therapy is concerned. If one can predict that he is likely to come down with a certain virus disease, it is a very good idea to immunize so as to build up the protective antibodies of the body in advance. But we might look forward to the day when we can take a bottle off the shelf and on it will be labeled "anti-measles virus--take one spoonful before retiring and your measles will be gone tomorrow." This is not very likely at the moment because such a development would have to plow through too many acres of ignorance.

Viruses are characterized by the fact that they can grow and develop and increase in numbers only within living cells. Viruses cannot be cultured in synthetic media in the test tube--at least not yet. As a consequence, it is no wonder that specific chemo-therapeutic agents have not been developed against them because any such agent would have to kill the virus when it is within the living cell. We believe that as they multiply viruses transfer from cell to cell without going out into the bloodstream, where they might be exposed to chemo-therapeutic agents. Thus, the problem is to kill the virus or to prevent its multiplication while it is within the cell. This can be done now, but not without destroying the cell.

In this lecture we will consider viruses as they exist outside cells. In a way, this might be considered the least interesting phase of their life cycle. One can think of a virus as a particle. Such a particle (either in whole or in part) enters a cell, and then something happens within the cell. The metabolic machinery is deranged so that the cell makes more viruses and eventually these are spilled out after the cell is sufficiently diseased.

The interesting part, certainly from the biological point of view, is what goes on within the cell. This is when the virus is actually doing something. Unfortunately, this is the area that is most surrounded with mystery and ignorance. More precise detail is known about viruses as they exist outside cells and this information is quite important. To use an analogy, we can learn the most about how a watch is made if we have a chance to take one apart and examine it. If we are unable to do this, looking at its deal and hands is better than not seeing the watch at all. Similarly, understanding the structure of viruses is better than knowing nothing at all about them.

There are literally thousands of different kinds of viruses. Of these, perhaps fifty have been purified to some extent. Of this fifty, perhaps twenty to thirty have been purified fairly well. Purification of course is a necessity, a preliminary to investigating their chemical and physical structure. It turns out that viruses range in size from being smaller than the larger particulate material of cells to being larger than the smaller particulates of cells. Therefore, by a process of differential centrifugation it is usually possible to effect a fairly good job of purification. This is done by first throwing down the bigger particles at low speed in the centrifuge. By increasing the centrifugal force, one can then throw down the viruses in the centrifuge, leaving the smaller material in the supernatant fluid.

If this does not do a satisfactory job, there are chemical methods that can be used. For instance, ammonium sulfate will rather specifically precipitate viruses as distinct from the normal proteins and nucleic acids of cells. By introducing ammonium sulfate into a sample of infected cell material, one can obtain a precipitate which contains largely the viral material. Or, one can take advantage of the fact that the densities of virus particles are likely to be different than the densities of other cell materials. Thus, by doing the centrifugation in a so-called density gradient (where one has a centrifuge tube in which the liquid is much more dense at the bottom than it is at the top) one can find a certain density where the viruses will concentrate. The other cellular materials will either go to the bottom or stay on the top.

We have seen that the methods for extracting viruses from cells and for purifying them are, for the most part, physical, with some use of chemical techniques. One can now make a preparation of viruses which is almost chemically pure, comparable to sugar in its purity. This would be true for two or three of the better known viruses.

Earlier it was stated that viruses are found in plants, in animals, in insects, and in bacteria. (In virological nomenclature, insects seem to be distinguished from animals, because one speaks about animal viruses on the one hand and insect viruses on the other!). Most of the work of a chemical nature has been done with plant viruses. Second in rank is that done with bacterial viruses, and third, with animal viruses. Little is known about the chemistry and physics of insect viruses. Most of the work with the biology of viruses has been done with bacterial viruses. As far as we know, practically all species and orders of animals and plants are subject to virus diseases. However, there are some exceptions. No one knows of a virus disease of a conifer. Also, yeasts, fungi, and molds seem to be immune from viral infection. No one yet has found a snake virus; but we have discovered a frog virus, and recently, a fish virus. Fowls and mammals of course are particularly subject to virus diseases. This is especially true for man.

Viruses are without exception smaller than the bacteria. Except for the one or two largest viruses, viruses cannot be seen with an optical microscope, or at least not discerned very well. In order to photograph them one must use an electron microscope. This means that they can be observed only when they are dry; moreover, their shapes may suffer some of the desiccation artifacts.

To understand the structure of viruses, one should first of all have some picture of their overall shape or morphology. This gives some inkling

as to how they must be put together. Then one should look for any subunits of structure that they might have, both in a chemical way and a physical way.

Figure R-1 shows a composite electron micrograph of eight of the more common viruses. This gives one the idea of the general gamut of sizes and shapes of various viruses. In the upper right-hand corner is the vaccinia virus, which is the virus that is used in smallpox vaccination (and is the same virus the smallpox disease itself has). In the upper left corner we see the particles of influenza virus. Below this is a bacterial virus which has a tadpole shape. Below this is an example of a very interesting virus of animals called rabbit papilloma, one of the animal cancer viruses. When this virus is rubbed onto a rabbit's skin, a lesion or wart-like growth develops. In the course of about a year this wart-like growth will actually turn malignant and the rabbit will die of skin cancer. Across from this we see another form of bacterial virus in which the tadpole shape is not nearly so prominent. This is a plant virus called the tomato bushy stunt virus. Its diameter is on the order of 0.03 micron, or 300 angstrom units (about 2 millionths of an inch). In the lower right-hand corner is the poliomyelitis virus. In the lower left-hand corner is the very well known tobacco mosaic virus, which is the best example of a rod-shaped virus.

Figure R-2 shows schematically the sizes and shapes of viruses. To give some idea of the scale, a human red cell (which one can just about see with a ten-power magnifying glass) is ten times the diameter of the large circle appearing at the top Fig. R-2. One of the smaller of the bacteria, the *B. prodigiosis*, has a diameter corresponding to the solid circle, so that it can be seen in an optical microscope. The largest objects that might be called viral-like in nature are the rickettsiae. Rocky Mountain spotted fever, for example, is a rickettsial disease. Rickettsia have some of the characteristics of viruses and some of the characteristics of bacteria. In other words, they can be cultured in synthetic media and they can be killed by antibiotics. From Fig. R-2 we see that the vaccinia virus has a diameter of about a quarter of a micron. Some of the viruses are quite large and are apparently spherical in shape. For instance, one of the insect viruses is quite large and has a polyhedral shape. Next in size is the vaccinia, which in turn is followed by the herpes. These cause the cold sores that most everyone has at one time or another. The virus that causes Newcastle's disease (one of the more prominent diseases of chickens) has about the same size and shape as influenza virus. There is an insect virus with a rather pronounced long cylindrical shape. Many of the bacterial viruses have tails that are about a tenth of a micron long and heads which are about two-thirds that. Also, some bacterial viruses are smaller, such as the T3 virus. Another familiar virus, the rabbit papilloma, has a diameter of about 0.05 micron. Tobacco mosaic virus is the best example of the linear viruses. There is no known linear virus of mammals; all the rod-like viruses that we know are found in plants. The smallest virus known is the foot-and-mouth disease virus with a diameter of 0.02 micron, which is 200 angstrom units or about 180 atomic diameters. An albumin molecule such as is found in the human serum or in an egg white is approximately a third as long as this smallest virus.

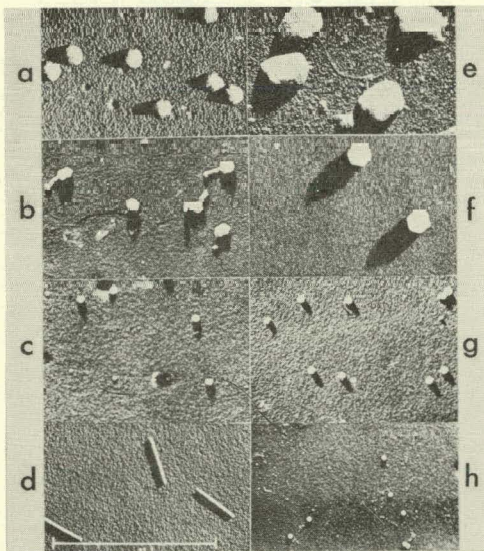


Fig. R-1. A composite electron micrograph showing eight of the better-studied viruses:

- a. Influenza Virus.
- b. T2 Bacteriophage.
- c. Rabbit Papilloma.
- d. Tobacco Mosaic Virus.
- e. Vaccinia Virus.
- f. Tipula Paludosa Virus.
- g. T3 Bacteriophage.
- h. Poliovirus.

X75,000.

Approximate Sizes of Viruses and Reference Objects

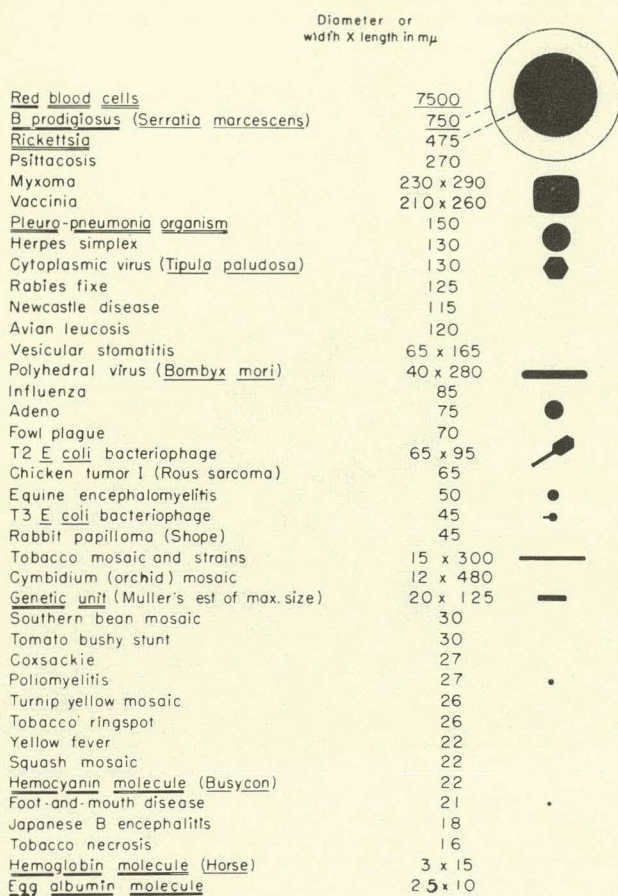


Fig. R-2. A chart of shapes and sizes of viruses and of certain reference objects for comparison.

Except for the largest ones, viruses have in one sense a very simple chemical structure, and in another sense a very complex one. They are simple in that they are composed of only two kinds of chemical--protein and nucleic acid. However, this still means that they are extremely complex chemically, because both protein molecules and nucleic acid are themselves complex. Nevertheless, without too much trouble, one can separate the protein from the nucleic acid portion and analyze them separately. It turns out that the protein portion is not just one big piece of protein with a molecular weight of, say, five million; rather, a virus seems to be made up of protein subunits. If we think of a protein with a molecular weight on the order of 15 to 30 thousand, we have some sort of concept of the size of the protein molecule of which viruses appear to be assembled.

Figure R-3 shows schematically a protein molecule, insulin. The two halves of the molecule together have a molecular weight in the order of six thousand. This begins to be comparable in size to the protein found in viruses. Figure R-3 of course, is schematic and indicates the array of amino acids by their shorthand nomenclature; actually this molecule is in two parts joined together with sulfur bridges. These individual amino acids

are organic compounds with molecular weights on the order of 200. When they are all put together their total molecular weight is about six thousand.

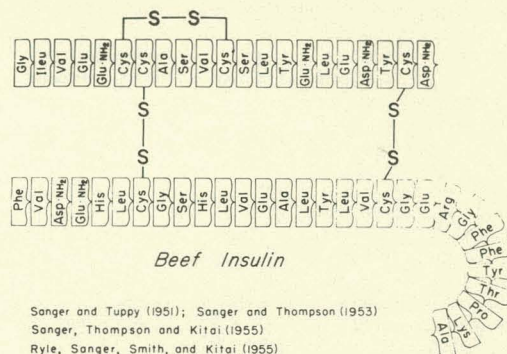


Fig. R-3. A schematic diagram showing the arrangement of the amino acids in the molecule of beef insulin. Note that the molecule is in two halves connected by sulfur bridges.

Figure R-4 shows a protein molecule (myoglobin) in three dimensions. This is the only protein molecule for which there is three-dimensional information in any detail. This structure was just recently worked out by Kendrew at Cambridge, England, and is the molecule of myoglobin, with the heme group of the myoglobin showing as a black area. This molecule is rather amazing because there seems to be no regularity about it. Yet one might have expected regularity. Crystals made up of millions of these molecules when put together in crystalline array show a fantastic degree of order. This means that each one of these molecules has its atoms in the same position as every other molecule. Yet, instead of finding a smooth sphere or a box or a tetrahedron or something, what one finds for an individual molecule is a most convoluted sort of structure.

The model shown in Fig. R-4 is made of clay. This main backbone would be then the backbone of



Fig. R-4. A photograph of a clay model of whale myoglobin, showing its three-dimensional structure as deduced from x-ray analysis (from the work of J. C. Kendrew and colleagues).

the amino acid chain shown in Fig. R-3; in other words, the peptide linkages among the amino acids would lie along this backbone. There appears to be a lot of empty space, but that is because this clay model could not be filled out completely. Some of the amino acids are quite large in diameter compared to their backbone structure. Inasmuch as the x-ray analysis does not indicate which amino acid is which, it would be unwarranted to try to fill up this space, in an attempt to show the true size of the amino acid residues. We are fairly certain that the true molecule is much more space-filling than what is shown here. However, Fig. R-4 is intended to give an impression of a molecule that is probably similar in its shape to one of the protein subunits of a virus particle--a convoluted polypeptide chain amounting to a collection of around 100 amino acids.

The other component of a virus, as mentioned above is nucleic acid. Figure R-5 represents schematically a polymer of nucleic acid, which resembles a double helix. This is a polymer of deoxyribonucleic acid (DNA). Such a polymer would have a length on this scale of perhaps a half mile, and this shows only a little bit of it. The main backbone of the helix is the phosphate sugar groups with an ester linkage tying each one to the next sugar group all the way along the backbone. At right angles to the backbone are the bases, which are characteristic organic components of a nucleic acid molecule. These are presumably held together in the middle of the double helix, one base coming from one chain and one base coming from the other. They are held together in the middle by hydrogen bonds, the whole making a fairly compact and rigid structure.

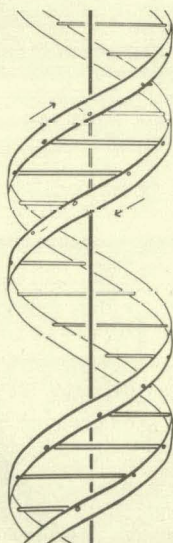


Fig. R-5. A schematic representation of a short section of a polymer of DNA, showing the double-helix nature of the polymer. The ribbon-like bands represent the sugar-phosphate backbone, while the horizontal rods represent the bases. The two strands of the double-helix are held together by hydrogen bonds.

Figure R-6 represents an atomic model of the polymer of nucleic acid shown schematically in Fig. R-5. It appears much more complex, but one can see the double helix. The black circles are the atoms of the sugar chain. The phosphates are along

this chain; they do not show very well because they are dyed black in this model. The bases are shown as planar molecules that run across at right angles to the main chain and hence at right angles to the page.

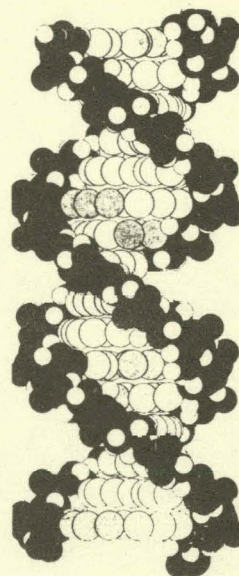


Fig. R-6. An atomic model of a short section of a DNA polymer.

It is believed that DNA has a special type of replication. Suppose that we put a virus or part of a virus into a cell. The cell then makes more viruses. The question arises: "Are the viruses that are then made within the cell genetically connected with the virus that went in? Is there genetic continuity in the virus world as in the animal world? The answer is yes. Dr. Stent will discuss this in more detail. Since there is genetic continuity, we next ask, "What carries the genetic continuity?" We now believe without any question that it is the nucleic acid that is in the viruses. This can be either of the deoxyribonucleic (DNA) type shown in Fig. R-6, or it can be ribonucleic acid (abbreviated RNA). The chemical difference between DNA and RNA is small indeed, being simply the presence of an OH group on the one and not on the other. But physically, the RNA is much different from the DNA; instead of having a double helix as does DNA, it has a single one. This viewpoint, which we now believe is correct, raises certain problems as to how replication occurs. This is not the problem of how a double helix of DNA goes about dividing in two and then forming new double helices. Rather, it is the question of how a single-helix-polymer (RNA) can be caused to grow within a cell as a result of only a single helix of RNA coming in and serving as a template for the new material.

Viruses may contain either DNA or RNA; most, however, contain RNA. Those of the nucleic-acid type seem to have no relation to whether they are plant or animal viruses. It was once thought that plant viruses are all RNA ones and that animal viruses are all DNA ones, but nothing could be further from the truth. For example, tomato bushy stunt virus which makes tomato plants sick, and poliovirus which makes humans sick, are approximately chemically identical. They both con-

tain protein; they both contain RNA to the extent of about 35%; they are the same size and shape; yet it is obvious that they are different viruses. Therefore, any investigation into how viruses are replicated within cells must take account of the fact that there is RNA to be replicated for some viruses and DNA to be replicated for others. There is no good evidence that the same virus ever has both RNA and DNA in it. They seem to have to make up their minds which they are going to have.

We will now consider virus structure in more detail. Figure R-7 is an electron micrograph of the poliomyelitis virus, which has a diameter of about 0.03 micron. This illustrates the fact that, as far as we can tell, every individual of a given type of virus has the same shape and size. This is particularly true of the smaller types. In fact, the x-ray crystallography of viruses indicates that they are indeed identical down to at least two atomic radii.

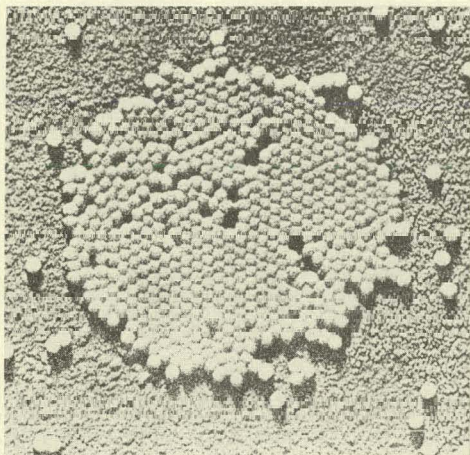


Fig. R-7. An electron micrograph of a purified suspension of poliovirus. X170,000.

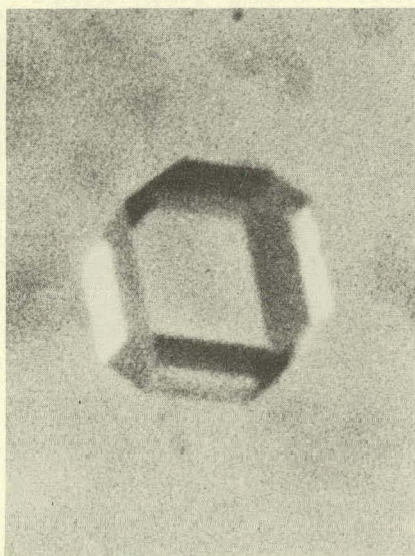


Fig. R-8. A light micrograph of a perfect crystal of poliovirus. The center of the crystal is solid, although it appears transparent here. There are about one billion virus particles within this crystal. X5,000.

Figure R-8 shows a crystal of a poliovirus. It is a full three-dimensional crystal, being about 0.2 mm across and containing about 100 billion virus particles. The virus particles within such a crystal are fully active biologically. One can do x-ray analysis with crystals like this to learn how the virus particles are arranged in their packed structure and how the particles themselves are built up. It is possible to "section" such a crystal; to freeze it, crack it, and make a replica of the cracked surface. Then by electron microscopy one can get some idea as to the arrangement of the virus particles in the crystal (see Fig. R-9). From a chemist's point of view, what is seen in Fig. R-9 could be called molecules. This electron micrograph reveals the internal array of small objects as they pack themselves in a crystal. In this case they have obviously packed with a square array. But the x-ray analysis tells us in much more detail about this than does electron microscopy.

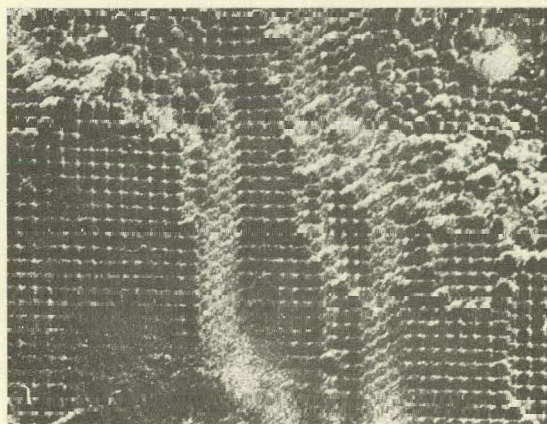


Fig. R-9. An electron micrograph of a section cut through a crystal like that seen in Fig. R-8. The arrangement of the individual virus particles within the crystal is shown. X250,000.

Figure R-10 shows clearly that a virus particle is not just a blob. In this case it is a rather good polyhedron, as indicated by the shapes of the shadows. The virus is that from the disease attacking the insect *Tipula paludosa*.

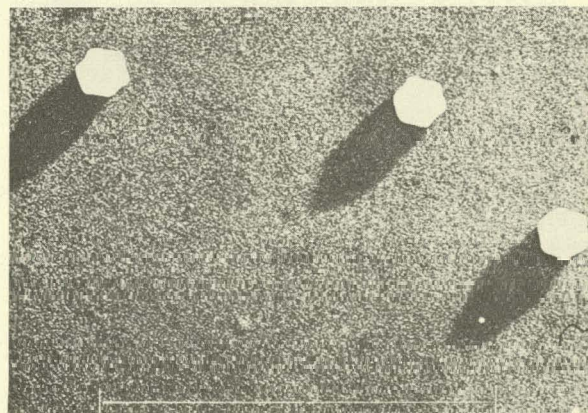


Fig. R-10. The *Tipula paludosa* virus, frozen-dried and shadowed to show the polyhedral nature of the virus. The three-dimensional shape is that of an icosahedron. X160,000.

Figure R-11 is a cardboard model which has been double-shadowed to reveal its shape. This is a model of the virus shown in Fig. R-10. The shape of the model (and the virus) is that of an icosahedron, a figure having twenty equilateral triangular faces. How the virus manages to grow so as to have these 20 planar faces is a mystery.

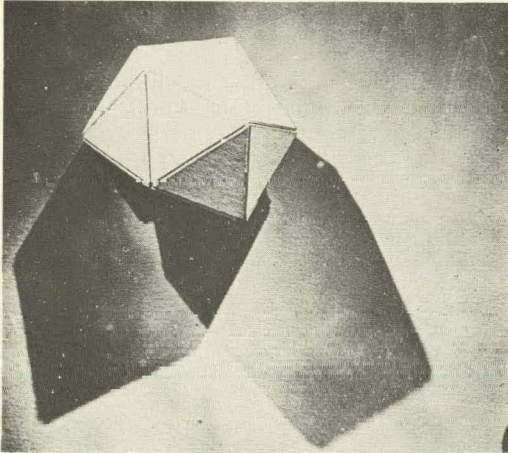


Fig. R-11. A cardboard model of an icosahedron (the shape of the *Tipula paludosa* virus shown in Fig. R-10). An icosahedron is a perfect polyhedron with 20 faces, each of which is an equilateral triangle.

Figure R-12 shows some representative bacterial viruses, the T2 bacteriophage attacking the bacteria of *E. coli*. Here we see that the bacterial viruses have a fairly intricate morphology: a head and a tail. It turns out that all bacterial viruses have tails, and as far as we know, only bacterial viruses have tails. Why they have tails is not known for certain, but it would seem that this structure is related to the job which the virus has in inoculating the host cell. A bacterial cell has a notably tough wall, as do plant cells. Most plant viruses are spread by insects; the insect with its proboscis is able to go through the cell wall. But a bacterial virus must infect the bacterium all by itself. To do this the virus attaches itself to a bacterium by the end of its tail. And then what happens? At first one might think that the whole

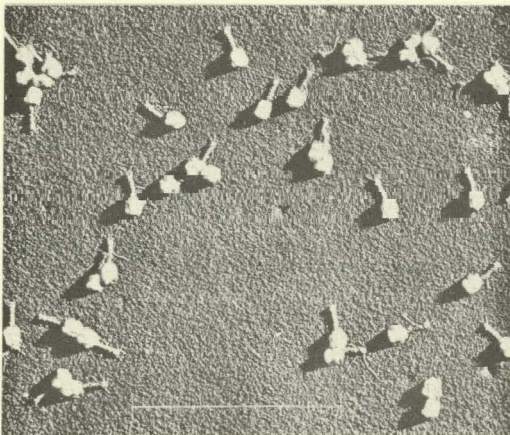


Fig. R-12. An electron micrograph of the T2 bacteriophage, showing the characteristic head and tail structure of all known bacterial viruses. X80,000.

virus goes into the bacterium, but this is not so. Actually only the inside of the head does. The function of the tail is apparently to aid in the transfer of the material inside the head (which is DNA) into the bacterial membrane.

Figure R-13 shows how we know what material is inside the head. In the left corner is a bacterial virus whose head has been broken and is now flat, revealing what looks like a mass of spaghetti. This spaghetti consists of the polymeric molecules of DNA that were represented in Fig. R-6. These molecules have a diameter of about 20 angstrom units only. This mass of DNA is enough to have a length of about a hundred times the diameter of the head and it is all in one piece, making a nice problem of assembly. That is, how does one bring this much DNA all together, ball it up in a package, and put a protein coat around it? This is just one of the many mysteries of virus structure.

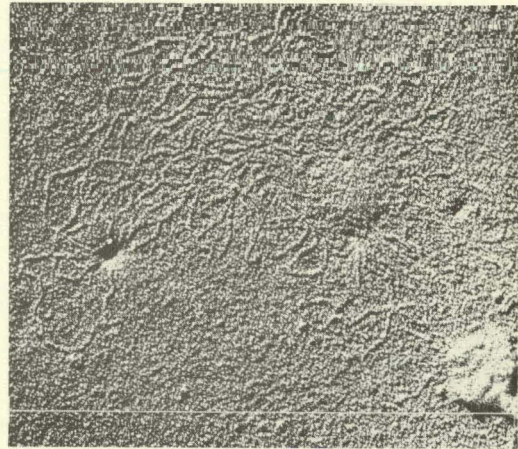


Fig. R-13. In this electron micrograph the remains of an osmotically-shocked bacteriophage is seen in the lower right-hand corner. The strands seen over most of the micrograph are the polymers of DNA that have been liberated from within the head following osmotic shock. X210,000.

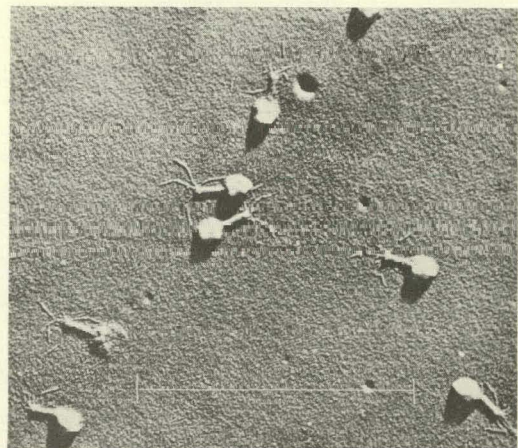


Fig. R-14. Some T2 bacteriophage which have been treated to one cycle of freezing and thawing. The tails have been slightly disintegrated to show the fine fibers which are attached to the ends of the tails. X110,000.

Figure R-14 shows that we also know something about the tail structure. If one degrades the virus slightly, he finds that at the end of the tail there are fibers, which seem to be the immediate organs of attachment. In a solution, bacteria and bacterial viruses move about with Brownian motion. When the virus approaches closely to the bacterium, these tail fibers attach to the bacterial surface, preventing the virus from bouncing off. Then the end of the tail is brought into contact with the surface of the bacterium. Figure R-15 shows schematically, although perhaps not very precisely, what happens next.

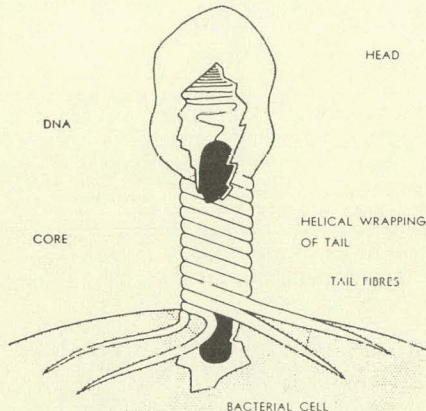


Fig. R-15. A schematic representation of the mechanism of attachment of a bacteriophage to the surface of a bacterial cell. The octopus-like arms on the bacterial surface are the tail fibers, while the black object running along the center of the tail represents the core. Upon injection of the DNA of the virus the core is believed to enter the bacterium.

Since this illustration was made, it has been discovered that following the attachment the tail actually shortens by a considerable amount. The tail has a core, i.e., a sort of solid plug going up the center. Two things happen after attachment: (1) There is an enzyme on the surface of this core that chews a hole in the bacterial surface, and (2) something in the bacterium enzymatically dissolves the core. After this occurs there is a hole down the tail through which the DNA from the bacteriophage head may enter the bacterium. This is a very complicated process, but it is probably necessary because bacterial cell walls are so tough and generally impermeable. Dr. Stent will discuss what happens inside the bacterial cell after the DNA has entered.

Another type of virus, which is illustrated in Fig. R-16, is a spherical virus. This illustration shows us something about subunit structure. The interpretation of this subunit structure is the following: Most viruses are RNA viruses with about one fifth of the total mass as RNA (for most viruses). Suppose now that we are going to make a virus out of nucleic acid and protein; how are we to do it? Let us assume that the nucleic acid acts as a template which controls how the protein is made. Is it possible to devise a scheme whereby all of the protein could be made at one time, yet allowing the nucleic acid to finish up on the inside, as we believe is universally so? The answer seems to be "no." It would seem to be much simpler to

imagine that the nucleic acid acts as a template for making a fairly small protein subunit, and doing this over and over again, so that a given virus particle may have a multiplicity of protein subunits each having a molecular weight about 15 to 20 thousand. These are then all assembled in quite an ordered fashion to make up the final virus. The type of order found implies that there are 60, or a multiple of 60, subunits in each virus particle. In Fig. R-16 the virus particles are those of rabbit papilloma. The electron micrograph at this magnification shows that the surface has an ordered array of bumps which are the protein subunits.

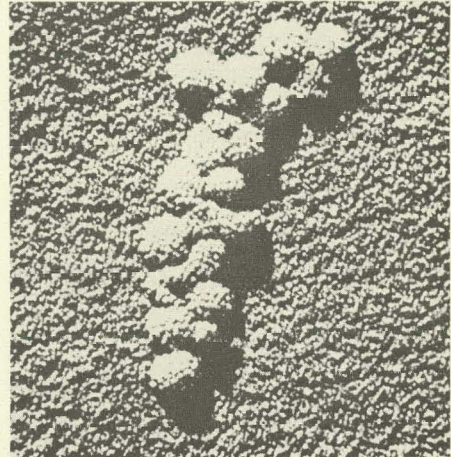


Fig. R-16. Some air-dried, shadowed particles of rabbit papilloma virus. The surface of the virus is covered with an array of knobs separated by about 70 angstrom units. X300,000.

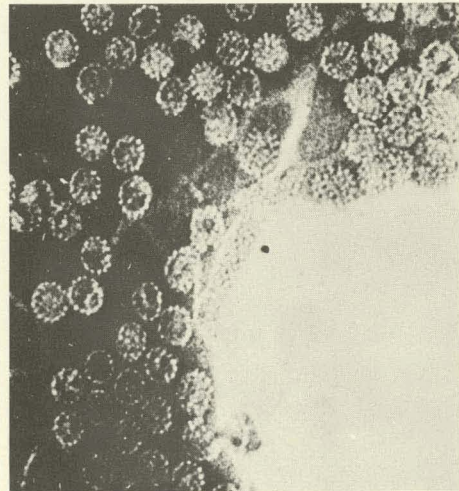


Fig. R-17. An electron micrograph of particles of rabbit papilloma virus which have been pre-treated with phosphotungstic acid. The knobs on the surface of the virus show clearly here. Note that some of the particles appear as hollow rings; these are hollow particles which have lost their content of nucleic acid. X300,000.

Figure R-17 perhaps shows this better. It is a preparation which has been stained with phosphotungstic acid (PTA), so that one can see the protein

subunits as white dots. From this staining technique we know that the PTA discloses the presence of the same kind of bump that was evident in Fig. R-16. Some of these virus particles are hollow. The PTA has entered and made them opaque in the middle. From Figs. R-16 and R-17 it is apparent that this particular virus has some kind of subunit structure on its surface--a set of bumps all of about the same size, which we know from chemical work are proteins.

Figure R-18 shows another example of this. This is a much smaller virus, the turnip yellow mosaic virus. Figure R-18 is an electron micrograph which is designed to show that the surface of these virus particles is not smooth, but rather it is knobby in a fairly ordered way. Because this virus is only 0.03 micron in diameter, it is rather hard to show the surface structure in detail. Nevertheless, this is an example of a very small virus, much smaller than the rabbit papilloma which exhibits a subunit structure on the surface. And we know very well in this case that the nucleic acid of the virus is in the center of the virus.

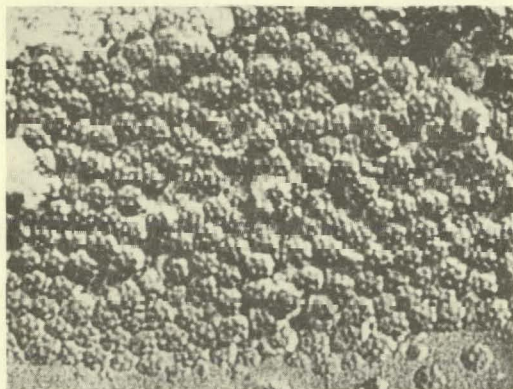


Fig. R-18. A collection of particles of turnip yellow mosaic virus, showing an array of knobs on the surface of the particles. The most characteristic array is one in which there is a central knob surrounded by six others. X400,000.

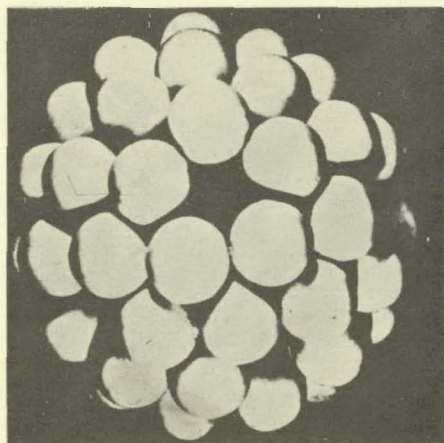


Fig. R-19. A ping-pong-ball model of the way in which the protein subunits are believed to be packed on the surface of a small virus such as poliovirus. The shape of the subunits is shown here as spherical but their actual shape is in all cases unknown.

Figure R-19 is a drawing that shows the structure of a small virus. This model, which is made of ping pong balls, is representative of the outer portion of practically all the small viruses. We do not know how the nucleic acid is assembled, but we do know from chemical and physical evidence (by degradation) that the RNA is in the inside. It is probably just folded back and forth, but we know nothing about this. On the outside there are protein subunits. X-ray analyses of two or three of the viruses show that these protein subunits are arrayed in a very ordered fashion. In these cases there are 60 of the protein subunits; the array makes a geometrical solid known as a snub-dodecahedron.

Figure R-20 shows a section of a larger and more complex virus, an insect virus known as the *Tipula paludosa* virus. A virus of this type has two regions: there is something that looks dark in the center and something which looks lighter on the outside. We know now from chemical work that the central portion is mostly nucleic acid, in this case DNA, and the outer surface is protein. The shape taken by this outer protein has been investigated by electron microscopy and, as will be recalled from Fig. R-10, is found to be a polyhedron.

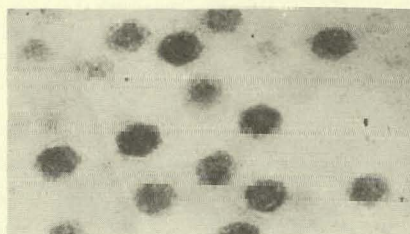


Fig. R-20. An electron micrograph of a section cut through the fat-body of a larva infected with the *Tipula paludosa* virus. Indications of polygonal contours are seen corresponding to the polyhedral shape of this virus as shown in Fig. R-10.

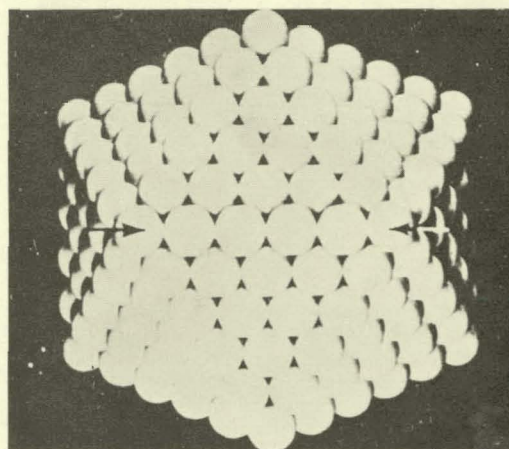


Fig. R-21. A ping-pong-ball model showing the surface structure of a polyhedral-shaped virus known as the adeno virus. It is believed that the surface of the *Tipula paludosa* virus is like this, although evidence for this statement is still lacking. The ping pong balls represent the protein subunits of the adeno virus. Note that an icosahedron can be formed by arranging spherically-shaped subunits in the way shown.

Figure R-21 is a model of an icosahedron made up of ping pong balls. The *Tipula paludosa* virus has this general shape, but has many more subunits than the model contains. They are fitted together with a geometrical nicety such that a twenty-faced surface of equilateral triangles is obtained. If we are going to speculate on how viruses are made, we have to discover a mechanism which will explain how nucleic acid of the infecting virus can influence the cellular machinery in such a way that separate protein subunits (of molecular weight of about 15,000) can be made and then assembled with such a high degree of geometric nicety, rather than as a blob.

Figure R-22 shows the tobacco mosaic virus, which is rod-shaped, or more properly, tube-shaped. The length of this virus is about 0.3 micron and its diameter is 0.015 micron. It contains 95% protein and 5% ribonucleic acid. We know more about the structure of this virus, surely, than any other virus. Because it is rod-shaped, there is some hope that we might be able to understand how this virus is put together and obtain some evidence about its substructure. In the electron microscope, an ordinary picture like Fig. R-22

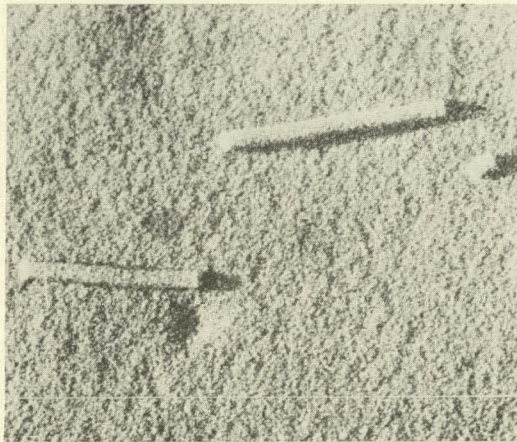


Fig. R-22. An electron micrograph of tobacco mosaic virus, showing the characteristic rod shape. No periodic array of surface structure has been found on this virus by electron microscopy, although such an array is believed to exist from x-ray analysis. X300,000.

gives no evidence of substructure whatsoever. But we can do a little trick, as Fig. R-23 shows. If we treat the virus particles with detergent, the protein is dissolved but not the nucleic acid. In Fig. R-23 one can see part of a virus particle, out of which extends a thread. From enzymatic studies we know that this thread is the RNA of tobacco mosaic virus. This helps us to answer these questions: "If a virus is to be made as a helix, with 5% RNA, how is the RNA to be put in? Is it to be inserted as multiple strands, say 10 or 12 strands of RNA, or is it to be put in as a single strand?" The answer seems to be that a single long strand is all that is necessary. By measurements made of a few hundred of these strands and knowing the amount of RNA that must have been in the virus to start with, we find that one long strand of RNA, such as is seen in the illustration, contains enough material to account for the 5% RNA content of the virus.

However, most of the evidence concerning the substructure of tobacco mosaic virus has come from chemical and x-ray studies. By chemical methods it is easy to separate the protein and the nucleic acid. This can be done with concentrated salt solutions, with acetic acid, with phenol, and in other ways. When this is done one obtains a protein subunit with a molecular weight of 18,000, which is made up of about 150 amino acids. The sequence of these amino acids is now being examined in our laboratory, and we hope to have an answer in a year or two. This study has been going for two years but its completion will take a long time.

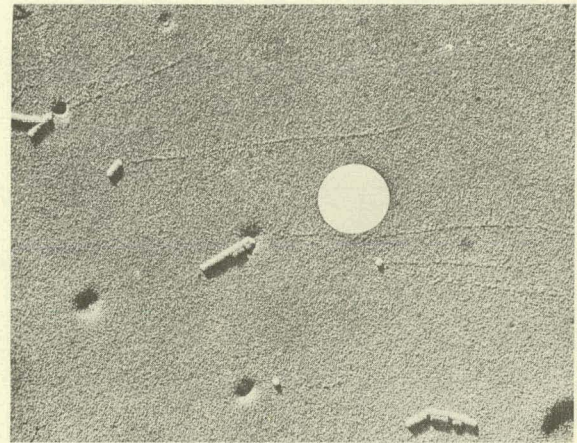


Fig. R-23. Tobacco mosaic virus which has been partially disintegrated by treatment with a detergent. The fine strands extending from the ends of the particles are the fibrils of RNA known to exist within the virus (the white sphere shown in the picture is a reference object of polystyrene latex). X120,000.

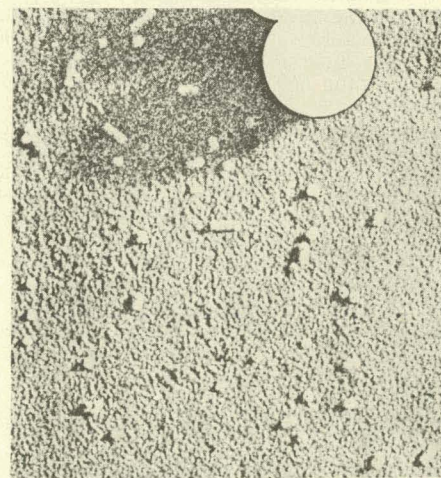


Fig. R-24. An electron micrograph of the protein fraction of tobacco mosaic virus which has been freed of RNA and which has been slightly reaggregated by reducing the pH of the solution. Continued treatment at low pH causes the protein fraction to reaggregate to rods which are indistinguishable from tobacco mosaic virus except that they are not uniform in length. X200,000.

Figure R-24 is an electron micrograph of the protein fraction which has been slightly re-aggregated. By treating the protein with just a little bit of acid, once can cause it to reaggregate in a form shown here; the small discs actually represent just a cross section of the virus particle itself. By treating the protein with acid for a longer time, one can cause it to reaggregate back to a shape that is very much like the complete native virus.

The protein can be purified by itself and the nucleic acid can be purified by itself, so that one already had an idea from chemical analysis that there were protein subunits of molecular weight 18,000. The x-ray evidence of crystals of DNA supports this concept. Figure R-25 shows the combined evidence of electron microscopy and x-ray analysis in quite a bit of detail. The rod (really a tube, because it has a hole in the middle) is a helix--a gigantic helix compared with the size of an atom. Figure R-25 shows the protein subunits of molecular weight 18,000, which are structural subunits as well and show up in the x-ray structure. The nucleic acid is neither in the center nor on the outside, but imbedded in the protein. This is shown schematically, but we now know that we should show the nucleic acid as a single fiber that goes round and round the helix, following the pitch of the protein portion. When we disintegrate the virus chemically we break up these subunits, one from the other. These are the ones with the molecular weight of 18,000.

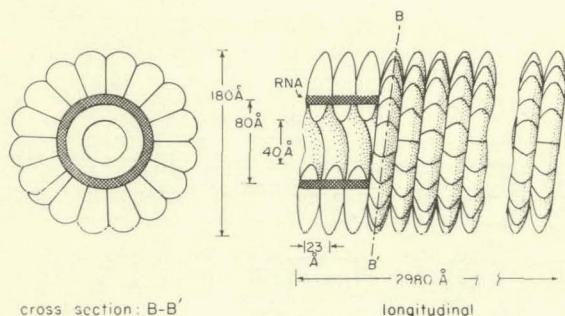


Fig. R-25. A diagram showing the arrangement of the protein subunits of tobacco mosaic virus, and showing the localization of the RNA. Note that the protein subunits are arrayed in the form of a helix. The helix has a hollow tube along its center, as shown here. It is now known that the RNA, shown schematically here some 40 angstrom units from the axis of the virus rod, follows the pitch of the helix. The information leading to this schematic representation of the structure of tobacco mosaic virus has come from electron microscopy and from x-ray analysis.

An interesting property of the tobacco mosaic virus is that we can easily put one together (or at least back together) in the laboratory. We do not know with certainty how this is done in nature, but we do have some idea. The laboratory procedure is first to take the virus apart chemically into its protein fraction and its nucleic acid fraction. The protein fraction consists of the subunits of molecular weight 18,000. The nucleic acid fraction consists of very thin fibers, which are the single helix-type polymer. The next step is to put these back together as a virus particle. To do this, one takes the nucleic acid fraction and the protein fraction

both in solution, and lowers the pH to about 5.5. The two are then mixed together in a test tube. After a few minutes we get what is shown in Fig. R-26. The result may not be surprising now, but it certainly was when this was first done, for we once again have a virus particle! As far as we can tell, antigenically, chemically, x-ray crystallographically, and even biologically, as regards their infectivity, these particles are identical with the natural particles with which we started. In other words, from such a reconstituted mix, we can get back about 80% of the original infectivity. At first this sounds like a mystery, but it is largely an exercise in polymerization. What happens is that the helix just grows round and round, with the protein units added one at a time to the RNA helix. The helix stops growing in length when the end of the nucleic acid fiber is reached.



Fig. R-26. An electron micrograph of tobacco mosaic virus which has been reconstituted from its constituent parts: protein and RNA. The reconstituted material appears like the native virus and has a fair uniformity of length. Such rods are almost as infective as the native virus before it was disintegrated and reconstituted. X55,000.

Now the question arises, "Is the entire reconstituted virus particle required in order to be infective, or is one of the components infective by itself? We have found that the protein is not infective. But Fig. R-27 illustrates an experiment which has led us to believe that the nucleic acid is infective. This is a "mixed-breed" type of experiment. After the nucleic acid and the protein have been separated, instead of putting them back together in the normal way, we put them back together by mixing. That is, we start with a virus of one strain, which we will call "A" and a virus of another strain, which we will call "B." We then combine the protein of one parent (B) and the nucleic acid of the other parent (A) and infect plants with the reconstituted virus. We find that the progeny that are grown in the plant, after infection, have the strain characteristic of the parent that furnished the nucleic acid. In other words, the experiment shows that the progeny always partake of the same genetic characteristics as the parent that furnished the nucleic acid. In fact, we now know that without exception the nucleic acid by itself is infectious. For example, it is possible to get about 5% of the infectivity level of native tobacco mosaic virus by infecting with its nucleic acid alone.

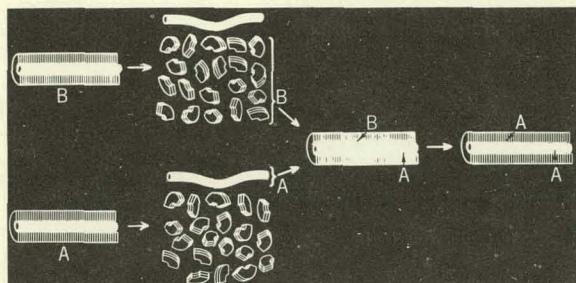


Fig. R-27. A schematic representation of a reconstitution experiment in which nucleic acid from one strain of tobacco mosaic virus has been reconstituted with the protein of another strain. In this case the nucleic acid from the strain called "A" has been combined with the protein from strain called "B." Following infection the progeny virus has all the characteristics of strain "A" as shown in the last particle on the right. This is a clear indication that the nucleic acid determines the genetic characteristics of the progeny, and was the experiment which led to the later discovery that the RNA of tobacco mosaic virus is by itself infective.

Thus, we have learned that the only part of a virus that is infectious is the core of the virus, i. e., just its nucleic acid. All viruses seem to have this core, with nucleic acid on the inside. Unfortunately, we do not yet have any evidence about the sequence of the nucleotides in the nucleic acid. To obtain this information is a much more difficult task than can be accomplished at the present. But we do know that this material, be it DNA or RNA, is infectious. Of course, with bacterial virus it was already known that only the DNA goes in, so that we could infer that in the natural world of bacterial viruses DNA is infectious. But it is fascinating to find in the artificial world of what man can do, that the RNA of viruses is infectious.

Suppose now that we take the juice of an infected plant from which the tobacco mosaic virus particles have all been removed. When we extract the nucleic acid that remains in the juice, we find that it is infectious. This indicates that more infectious nucleic acid is made than is finally wrapped up in the complete particle. Why then must protein be present in the complete particle? An answer may be found by reflecting on why you want a shopping bag when you buy a lot of small packages at a store. It is probable that the protein serves primarily as a protective cover for the nucleic acid, as something to wrap it in.

In just the last few months it has been found that the ordinary specificity rules of virus infection break down. Previously, we believed that a given type of virus will infect only one host species or family. For example, I have worked in tobacco mosaic virus for years, but I have no infection from the tobacco mosaic virus because I am not a tobacco plant. Likewise, poliovirus has been known to infect best only the cells of primates, that is, monkeys, humans, and so forth and to infect less well the cells of other mammals. Certainly, poliovirus will not infect the cells of a frog. This selectivity is probably because the protein will not allow the virus to attach to the cells of a frog. But, unless the virus attaches to the cell, the nucleic acid cannot go in and cause infection. However, it has recently been discovered that if the protein is stripped off and only the nucleic acid is placed upon the cells, we can infect tissue-cultured frog cells with polio "virus." The frog cells will then produce true poliovirus, and the particles that come out look just like the original poliovirus that furnished the nucleic acid. The virus particles that are now produced by the infected cells (the complete virus particles) no longer recognize the frog cells. That is, once the protein coat is around the nucleic acid, the virus cannot infect the frog cells. It now appears likely that within a few years we shall be able to infect plant cells with animal viruses, or infect animal cells with plant viruses, if we use only the nucleic acids. We virologists are then going to have to watch our step, because we would hate to come down with a case of tobacco mosaic virus!

Bibliography

1. F. M. Burnet, The Structure of the Influenza Virus, Scientific American, 37 ff. (Feb. 1957).
2. F. M. Burnet, Viruses, Scientific American, 43 ff. (May 1951).
3. F. H. C. Crick, Nucleic Acids, Scientific American, 188 ff. (Sept. 1957).
4. H. Fraenkel-Conrad, Rebuilding a Virus, Scientific American, 42 ff. (June 1956).
5. G. S. Stent, The Multiplication of Bacterial Viruses, Scientific American, 36 ff (May 1953).
6. The Viruses, F. M. Burnet and W. M. Stanley, Eds., Vol. I (Academic Press, New York, 1959).

S. THE BIOLOGY OF VIRUS

Gunther S. Stent

Viruses are of the greatest practical importance. Not only do they cause some of the worst diseases that afflict man, but they attack plants and animals as well. A very large effort of virologists all over the world is now being directed toward controlling virus diseases. Of fundamental value in this investigation is the study of how viruses reproduce. That is the subject of this lecture.

In the free state, virus particles manifest no metabolic phenomena at all. However, when a virus comes in contact with a living cell, a remarkable metamorphosis takes place: the virus comes to life. Processes then ensue inside the infected cell which lead to reproduction of the infective particle.

A given virus cannot infect every type of living cell. Those viruses that attack cells of animals are called animal viruses. Also there are plant viruses and insect viruses. However, this lecture will be devoted to those viruses that infect bacteria, in particular Bacterium coli. This bacterium, which is found in the human intestinal tract, can be infected by a very large range of viruses.

It may come as a surprise that even the smallest of all organisms—bacteria—are themselves the prey of virus disease. Forty years ago, when bacterial viruses were discovered, there was great excitement, for it was felt that at last a means was found for controlling bacterial diseases. The line of thought was obvious: to cure a patient infected with diphtheria, one would simply administer a virus specific for the diphtheria bacterium. The virus would seek out the diphtherium bacterium and destroy it, thereby curing the patient. In fact, it was thought that an epidemic such as cholera could be averted by this same method. One would merely isolate a virus that infects cholera bacilli, disseminate this virus among the people, and then encourage everyone to be as unhygienic as possible so as to facilitate the spread of this anti-cholera virus. In this way, the cholera bacteria would all be quickly eradicated.

Unfortunately, this hope was not realized. Therapy by bacterial viruses turned out to be quite fruitless. Today antibiotics are widely successful, much more so than was ever thought possible by even the most enthusiastic supporters of bacterial-virus therapy. In fact, bacterial viruses are of virtually no practical consequence nowadays. However, as objects of fundamental biological study, they are of the utmost importance.

Figure S-1 shows that bacterial viruses are small tadpole-shaped objects having a head and tail.

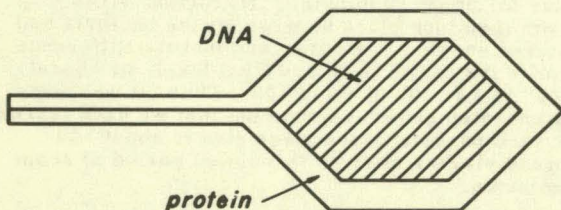


Fig. S-1. A bacterial virus of the T2 strain.

The head is filled with DNA (desoxyribonucleic acid), a substance which will be discussed in more detail below. The outer coat of the head is made of protein. Thus, the main chemical components of virus are protein and DNA, which occur in nearly equal amounts.

Figure S-2 shows a cell of the bacterium coli surrounded by virus particles of the type that can infect it. The relative sizes of the virus particles and the coli cell are evident. The length of the coli cell is about $1/10,000$ inch, which is about the limit of optical visibility. To observe the virus particles, one must use an electron microscope.

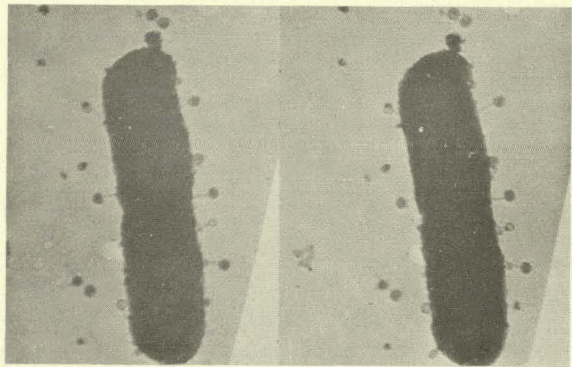


Fig. S-2. The bacterium coli, which infects the human intestinal tract. Virus of the T5 strain are seen adsorbed on the cells.

The reproduction cycle of the bacterial virus is represented in Fig. S-3. When virus particles are introduced into a population of bacteria suspended in some medium, the particles diffuse toward a bacterial cell. When a virus particle succeeds in hitting a bacterium, it fixes itself by means of its tail to the surface of the cell.

After attachment takes place the tail contracts, pushing a kind of hypodermic needle inside the bacterial cell. The nucleic acid (DNA) of the virus is then injected into the interior of the cell, as indicated by the dotted line in Fig. S-3a. The protein coat of the virus never enters the cell.

Once infection has occurred, the reproduction drama of virus growth ensues. Since only the DNA enters, it must be the agent responsible for virus reproduction inside the cell. Also, DNA must represent the chemical substance which carries the hereditary continuity from one virus generation to the next.

If we examine the cell 10 minutes after infection has occurred (Fig. S-3b), we find that the membrane containing the protein is still attached outside the cell. Inside the cell we find some hollow particles that contain protein, but no nucleic acid. Because they contain no nucleic acid, these virus precursors are incapable of producing infection. The synthesis of the viral nucleic acid is, however, already taking place, as can be demonstrated by direct chemical tests.

The nucleic acid will soon combine with the protein to form infective virus particles.

Twenty minutes after infection (Fig. S-3c), we find the first infective virus particles inside the cell. These contain both protein and nucleic acid. There are still some virus precursors present, whose growth process is not yet completed.

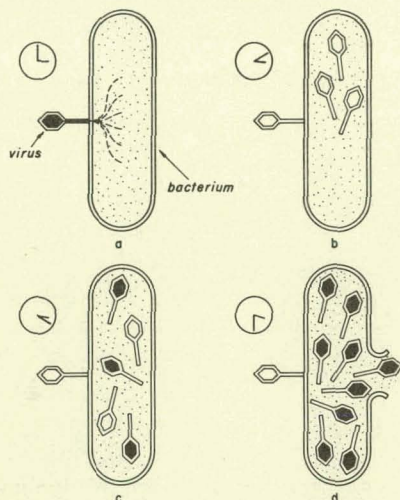


Fig. S-3. Reproductive cycle of bacterial virus: (a) at the moment of infection; (b) 10 minutes later; (c) 20 minutes after infection; (d) 30 minutes after infection.

After 30 minutes (Fig. S-3d), a hundred or so infective virus particles have accumulated inside the cell. The cell then bursts open, liberating the progeny into the medium.

What we have just described illustrates the central problem of biology, namely, to know what kind of structure manages to reproduce itself. We now turn to consider the techniques and methods by which we discover the information just discussed.

One of the reasons why bacterial viruses are so suitable for fundamental biological studies, and why they can be used so easily, is that an accurate assay is easily made of the number of virus particles present in a given sample. Suppose that we want to know the number of bacterial viruses in a certain solution. The method is as follows. A nutrient agar is poured onto a petri plate. This agar is then inoculated with about a million bacteria of a type susceptible to the virus we wish to assay. Next, a dilution of the virus suspension is mixed with the bacteria, so that about 100 virus particles are placed on the plate. This mixture of bacteria and virus is incubated overnight at 37°C. During incubation the bacteria multiply, each bacterium giving rise to a colony of descendants. By morning the surface of the nutrient agar is covered with a thick film of bacterial cells. The previous day, the plate was completely clear, but now the surface is turbid with a thick lawn of bacterial growth. However, each place where a virus happened to come to rest there will be a hole. What has happened is that each virus infected a bacterial cell. Half an hour later each cell burst open, liberating several hundred progeny particles. These progeny infected neighboring cells and in turn multiplied on them. After another half hour, several hundred more progeny were released from each cell. So as the bacteri-

al lawn grew, a tiny virus colony also grew at each site where an original virus came to rest, creating a hole. We call these holes "plaques." To determine the number of virus particles deposited on the plate, one has only to count the number of plaques and take into account the dilution factor. People who work with bacterial viruses spend at least half their time counting these plaques. Even so, this is incomparably easier than making assays of plant or animal viruses. For example, one might need perhaps 200 mice to make an assay similar to that done with a single agar dish.

The method used to measure the specific infectivity of a virus preparation is as follows: One first determines the number of infective virus particles by the plaque method. Next, he counts the number of particles by means of an electron microscope. For bacterial viruses, it turns out that the ratio of infective particles to physically visible particles is one. In the case of plant and animal viruses this ratio may be one in a thousand, or in some cases only one in a million. That is, in the case of plant and animal viruses only a very small percentage of the particles seen in an electron microscope ever do anything biologically interesting. Thus, in working with plant or animal viruses there is always the uncertainty that the particle causing a biological effect does not represent the bulk of the preparation. However, when dealing with bacterial viruses, we know that every particle is infective. This is one of the main advantages of working with bacterial viruses.

The plaque assay method is also employed to follow the kinetics of virus multiplication. For this purpose we use a bacterial culture containing, say, 100,000,000 cells per milliliter. Next, we introduce an equal number of viruses, which immediately begin to diffuse toward the bacterial cells. The virus particles fix themselves on the surface of these cells, and the infection is underway. At one-minute intervals (measured from the moment that the viruses were introduced), aliquots of this growing virus-infected culture are removed. The aliquots are then suitably diluted and plated on agar for the plaque assay in order to see how many infective centers exist.

It turns out that in such an experiment the number of viruses remains the same for about 25 minutes. The number then suddenly begins to rise, continuing to increase for perhaps 10 minutes, and again levels off. This is represented by the solid curve in Fig. S-4. What has happened in this "one-step growth experiment" is that during the first 25 minutes, or latent period, the viruses were multiplying inside the bacteria. At that point the bacteria began to break open and liberate the virus particles they harbor inside. This liberating process continued for about 10 minutes. No further virus growth then took place because all the bacteria had ruptured and died. We thus see that the difference between the initial level and final level, or "burst size," is a factor of about 100. Thus, it was from this one-step growth experiment that we have learned that each parental virus gives rise to about 100 progeny viruses after an incubation period of about 30 minutes.

Closer to the problem that we want to investigate is the question: "When do the viruses actually appear inside the cell?" The one-step growth experiment does not answer this—it merely tells when the viruses are liberated into the medium. To answer

this question, we must break open the bacteria artificially. This can be done by pressure, certain chemicals, or ultrasonic sound. After we have broken the bacteria we plate aliquots of the infective culture on agar to count how many infective viruses are present. This tells us the intracellular accumulation of viruses after a given interval rather than the time of their spontaneous liberation.

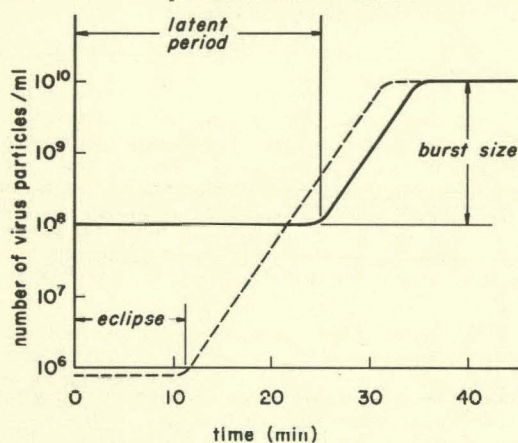


Fig. S-4. The kinetics of bacterial virus growth: solid curve represents the one-step growth experiment; dashed curve indicates the intracellular growth experiment.

The first person to perform this intracellular growth experiment was greatly surprised at the result. He naturally expected that immediately after infection there would be the same number of viruses as were introduced. But to his amazement he found none, not even the virus particles that were put in. In fact, for the first 12 minutes he found essentially no infectivity in bacterial cells that were artificially ruptured. After 12 minutes, there was a sudden increase in infectivity, which rose for about 20 minutes, and then leveled off at the same final value as in the case of the one-step growth experiment. This is represented by the dashed curve in Fig. S-4.

What is the reason for this lack of infectivity during the first 12 minutes, or period of "eclipse?" It will be recalled that when infection occurs, the nucleic acid is injected but the protein remains outside the bacterial cell. The nucleic acid is not infective unless it is inside a protein coat, which it uses as a syringe. However, by the time 12 minutes have elapsed, a certain amount of viral protein has been synthesized inside the cell. This combines with nucleic acid to form an infective progeny. When this occurs, the eclipse is over. From this point on, more and more viral nucleic acid combines with protein to make additional virus particles, until the final level is reached. Thus, in this intracellular growth experiment the same final level is attained as in the one-step growth experiment.

What is the nature of the nucleic acid, which plays such an important role in carrying the hereditary information of the virus particle? Figure S-5 indicates the nature of a chain that makes up nucleic acid or DNA. It is a polymer of very high molecular weight, built of repeating units composed of molecules of the sugar, deoxyribose, linked through phosphate ester bonds. Thus, along the main chain of this polymer we have a succession of sugar, phosphate, sugar, phosphate, sugar, etc., continuing up to a molecular weight of 10 million. This is an extremely

long molecule, much longer than a protein molecule. To the side of each sugar is attached one of four purine or pyrimidine bases.

D.N.A.

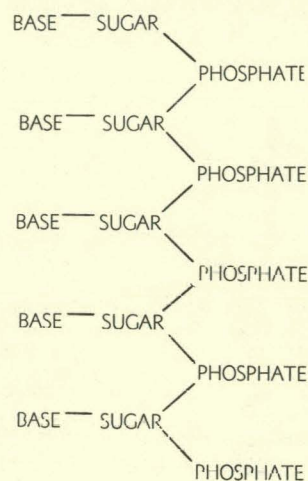


Fig. S-5. A chain making up desoxyribonucleic acid, often referred to as DNA or nucleic acid.

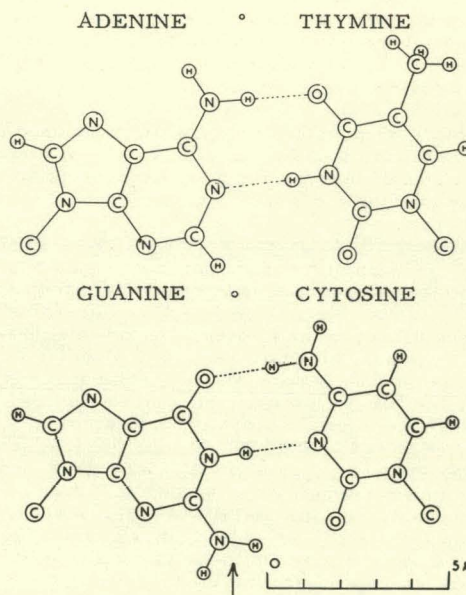


Fig. S-6. The four bases which are attached to each sugar molecule contained in a DNA chain. The bases are adenine, thymine, guanine, and cytosine.

Figure S-6 shows the nature of these bases: one is called adenine, one thymine, one guanine, and one cytosine. These are complex heterocyclic, organic molecules which form the real working part of DNA.

Desoxyribonucleic acid consists of a double chain (see Fig. R-5 of Section R), each like that indicated in Fig. S-5. These two chains are helically intertwined and are held together by purine and pyrimidine molecules that jut out sideways. The hereditary information resides in the particular ar-

rangement of these purine and pyrimidine residues along the chain.

Another reason why it is so convenient to work with bacterial viruses is that they can be easily studied with radioactive tracers. One of the tracers used extensively is phosphorous 32. This isotope emits beta rays, which are among the easiest to observe with a radiation detector.

A common procedure is to add P^{32} (perhaps in the form of phosphoric acid) to a culture in which the bacteria are growing. The bacteria then assimilate this radiophosphorous into their phosphorylated constituents. If these bacteria become infected with viruses, the viruses then assimilate the materials making up the bacteria and become, therefore, themselves radioactively labeled. After purification of such viruses by centrifugation we obtain a preparation in which the only radioactivity resides in the bacterial viruses themselves. In particular, in the case of the P^{32} label, the radioactivity is contained in the phosphate links which hold together the sugars in the DNA. The protein contains no phosphorous, nor does the virus have any lipids or other constituents which contain phosphorous. Therefore, in a bacterial virus that has been labeled with radiophosphorous, the radioactivity resides exclusively in the DNA. We can follow the fate of the DNA by simply following the radioactivity which the virus contains.

A second radioisotope used extensively in this work is sulfur 35, which is also a beta emitter. The same procedure as outlined above is followed to obtain a preparation of bacterial viruses in which all the radioactivity is in the viruses themselves. The difference is that there is no sulfur in DNA. All the sulfur is in the protein, i. e., in its two amino acids, methionine and cysteine. In this way, we can obtain two virus preparations that are identical, except that one is labeled with radiosulfur and the other with radiophosphorous. We now have a method for following the fates of the protein and the DNA separately.

Suppose that we infect a population of bacteria with a virus labeled with radiosulfur. We know that the virus attaches itself to a cell by its tail. After a few minutes we place this culture in a Waring blender. This sets up such tremendous shear forces that the protein coating is torn away from the cell. We discover that this has actually occurred by using a radiation counter. We then use the plaque assay test and find that in spite of this rough treatment, the cell is infected. From these facts we conclude first, that the protein has not entered the cell (for otherwise we could not have stripped it off), and secondly, that the protein is not necessary to produce infection. To discover the role of DNA in producing infection, we repeat the experiment, this time using radiophosphorous rather than radiosulfur. In this case the Waring blender does not strip off the radioactivity, indicating that the DNA is inside the cell.

Finally, we turn to consider briefly some genetic studies on bacterial viruses, which show that the hereditary apparatus of these tiny organisms is in many respects very similar to that of higher forms. The type of plaque which an ordinary wild virus forms when plated on an agar dish has a characteristic morphology, namely, a clear center with a sort of turbid halo around it. In the many thousands of plaques that we observe daily, we occasionally find one that looks different. It is bigger and does not have the turbid halo, but instead is clear around its edges. We refer to this rare type of plaque as an r plaque.

If we carefully isolate the contents of such a virus colony and allow the particles to reproduce, we find that the offspring all make the same type of aberrant plaque. That is, the alterations that have suddenly arisen perpetuate themselves in a stable way, producing a new genetic type, i. e., a mutant.

There also exists another plaque-type mutant. This is designated by letter m, standing for minute, since it makes a very tiny plaque.

It may come as a surprise to learn that bacterial viruses exhibit a kind of sexuality. This can be demonstrated by infecting a bacterium with both the r-type and the m-type virus. Since more than one virus particle can grow in a single bacterium, both of these can multiply inside a given cell. Among the hundred or so progeny coming from a cell, some will be of the r type and some the m type. In addition to those that are exactly like their parents, there will be some which give rise to the normal plaque, i. e., are neither r nor m, but resemble the original wild type.

Where do these normal wild-type viruses come from? This can be answered in terms of conventional genetic ideas. The genetic structure of a virus, i. e., its "chromosome," can be represented as a single straight line. Each gene (i. e., a locus which controls a different activity of the virus) can be assigned different locations on this linear chromosomal map. Just as in chromosomes, a form of crossover can occur in a "cross" between two genetically different viruses. In Fig. S-7 are indicated the chromosomes of the two parent viruses of such a cross. It is apparent that if a crossover occurs between the point m and r, then a normally nonmutant chromosome arises (as well as a chromosome bearing both m and r mutant loci).

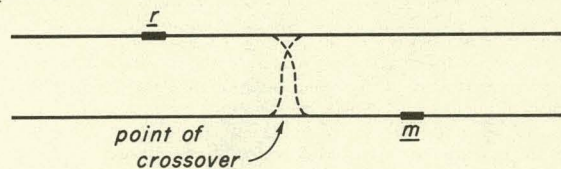


Fig. S-7. Genetic crossover between chromosomes of two different mutants.

Bibliography

1. M. H. Adams, *Bacteriophages* (Interscience Publishers, New York, 1959).
2. F. H. C. Crick, The Structure of the Hereditary Material, *Scientific American*, 54 ff. (Oct. 1954).
3. M. Delbruck and M. Delbruck, Bacterial Viruses and Sex, *Scientific American*, 46 ff. (Nov. 1948).
4. C. A. Knight and D. Fraser, The Mutation of Viruses, *Scientific American*, 74 ff. (July 1955).
5. S. E. Luria, The T2 Mystery, *Scientific American*, 92 ff. (Apr. 1955).
6. A. Lwoff, The Life Cycle of a Virus, *Scientific American*, 34 ff. (Mar. 1954).
7. J. L. Melnick, Viruses Within Living Cells, *Scientific American*, 38 ff. (Dec. 1953).
8. G. S. Stent, The Multiplication of Bacterial Viruses, *Scientific American*, 36 ff. (May 1953).
9. *The Viruses*, F. M. Burnet and W. M. Stanley, Eds., Vol. II (Academic Press, New York, 1959).

T. LIPOPROTEINS AND HEART DISEASE

Alex Nichols

The relationship of lipoproteins to heart disease has been bandied about in newspapers and magazines to a considerable extent. There is always the problem of interpreting these popular accounts. In this lecture we will try to give you some basis for evaluating this information.

It is important first to know something of the heart, its arteries, and its musculature. Figure T-1 shows that the chamber walls are very strong muscular cells. This drawing also reveals the venous and arterial systems, which carry blood to and from the heart with each pumping motion. Also evident are some small arteries that crown the heart. These are the coronary arteries, which will be our major topic in this lecture.

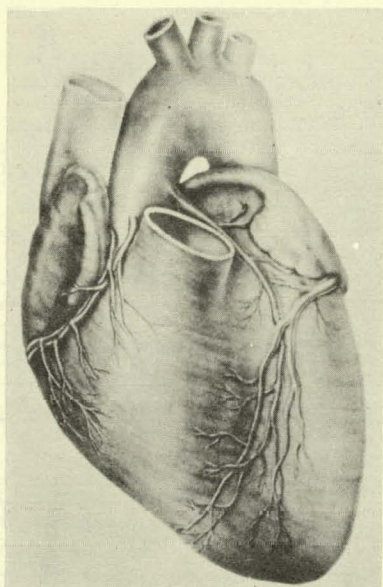


Fig. T-1. The human heart.

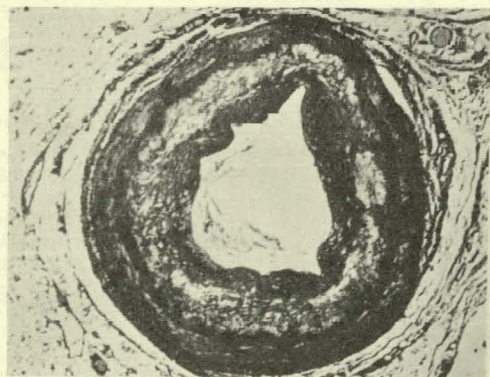


Fig. T-2. Cross section of an artery of a person who died of coronary disease, showing how the effective passageway for blood has narrowed.

The heart continually pumps blood to the various tissues of the body, supplying to them the nutrients essential for life. However, the nourishment for the heart itself comes through the small coronary arteries. If these become obstructed, the heart tissue suffers starvation and suffocation.

Examination of the hearts of persons who died from heart attacks has shown a significant narrowing of the passage through the coronary arteries (Fig. T-2).

In some cases these arteries were completely plugged. Chemical analysis of the blocking material revealed that it is made up of cholesterol, triglycerides, phospholipids, calcium, and decayed tissue. As this material begins to collect, partially obstructing the coronary artery, there is an ever greater tendency for a blood clot to form. If this happens, a portion of the heart will not receive nutrients. This means that the affected tissue dies, causing a heart attack. This is very painful and often fatal. On the other hand, if there is only a partial blockage, the condition is called angina pectoris. With this, there is pain and also the possibility that the blood supply will eventually become completely blocked, causing a heart attack.

A corollary finding to this was that the coronary system has a fantastic recuperative power.

When a blockage occurs, collateral circulation develops in time. Small arteries form around the block, allowing the affected muscle to recuperate. The circulation will not be as good as it was before the disease started, but it is sufficient to enable the person to get around. This is mainly the reason for not overworking after a heart attack.

It makes a great difference where the narrowing occurs. A blockage in the upper region of the heart is the most deadly, because the blood supply is cut off from a larger area.

So far we have been discussing studies on individuals with heart disease. The question now arises, "What about presumably normal individuals?" An interesting investigation was made in Korea on soldiers killed in battle. Surprisingly enough, the coronary arteries of many of these healthy young men were narrowed to some extent. From this and other studies it was apparent that heart disease starts rather early in life. Its symptoms may not appear until the person has reached middle age. Then without any warning, a clot will form causing pain or even death.

Unfortunately, at present we cannot tell with certainty if a specified young person is a candidate for a heart attack in later life. An x-ray machine cannot detect this narrowing, nor can an electrocardiogram (EKG). An EKG will indicate only that actual damage to the heart has occurred. Yet we would like to find out early who is more susceptible and establish preventive measures for these individuals. How does one go about doing this?

Our approach has been a statistical one. It has involved studying a representative population of thousands of individuals. All persons with diagnosed heart disease were excluded from the study, because we wanted to learn how a susceptible individual differs from the others before he has any symptoms. This population was then observed over a period of four years or so, since it takes that long for an appreciable number to become coronary victims. During this period certain variables were recorded for each person: his height, weight, age, blood pressure, etc. Also, detailed analyses were made of his blood. For instance, the amounts of blood lipoproteins (particles that carry blood fats) and cholesterol (a blood fat) were determined.

At the conclusion of the test period all the data were analyzed. We then looked for significant differences in the above variables between the people who developed heart trouble and those who did not. Unfortunately, during this study we were unable to record every conceivable variable. But among the variables that were recorded, three were found to be significant:

- high concentrations of blood lipoproteins and blood fats,
- high blood pressure, and
- overweight.

A person with any one of these would have a greater than average risk of having a heart attack; a person with all three would have the greatest risk.

We also found that of these three factors, the first two are about twice as significant as the last one in indicating the risk. However, because of the statistical nature of the study it would be impossible to predict that a given individual will or will not have a heart attack. This can be illustrated by Fig. T-3, which plots blood pressures for both a coronary and a noncoronary group.

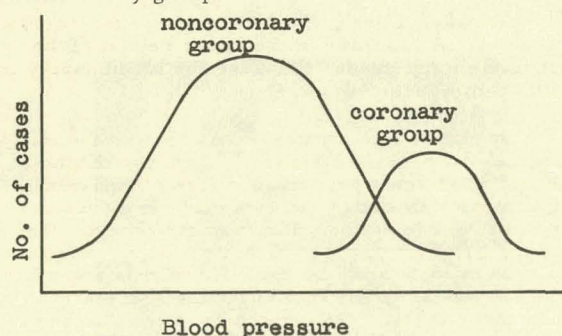


Fig. T-3. Graph showing how, on the average, coronary victims have higher blood pressure than noncoronaries. (Not to scale.)

It is apparent that the mean blood pressure for the coronary group is significantly higher than for the noncoronary group. But there is some overlap of the curves, indicating that not every person with high blood pressure will have a heart attack. It is merely more likely that he will than if his blood pressure were normal.

Suppose that we were to draw a blood sample from a person. If we let the blood stand about two or three hours, a clot would form (Fig. T-4).

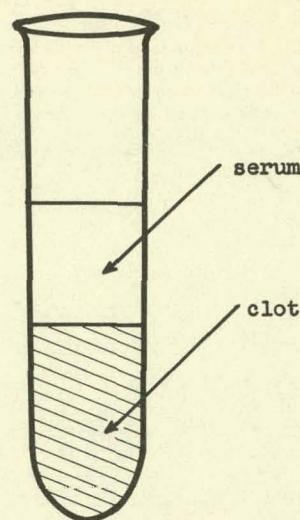
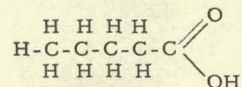


Fig. T-4. After standing 2 to 3 hours in a test tube, blood separates into two distinct bands.

This clot, which settles to the bottom, contains the red blood cells. Above this we would see a clear, yellowish liquid known as the serum. The volume relationships are about 50-50. It is the serum that we are principally interested in, because it contains the lipids. The lipids are extracted from serum by the classical technique of shaking it with organic solvents. After doing this we find that serum contains cholesterol esters, phospholipids, and triglycerides.

These materials are chemical compounds containing long-chain fatty acids. These chains are composed of carbon atoms surrounded by hydrogen atoms:



These chains, which vary in length, are the principal compounds that are present in the fats in the body. In other words, they are present in the cholesterol esters, in the phospholipids, and the triglycerides. These chains are linked to certain backbones. For example, there is a glycerol backbone consisting of three carbon atoms; linked to each carbon atom is a fatty-acid compound. The resulting combination is known as a triglyceride, a trilipid compound, or a tri-fatty-acid compound. Thus, from a strict extraction we find that lipids are present in serum, just as they are present in other tissues of the body.

From a population study similar to that described above we found that these lipids are definitely elevated in the coronary population. Likewise, they are elevated in individuals who subsequently have coronary attacks, not in every case, but on the average.

The next question to be asked was concerned with how these compounds are dispersed in an aqueous solvent. It is known that if one mixes water and vegetable oil, a two-phase system results. The oil floats to the top and the water goes to the bottom. If this were the case in a human we would all become senile rather early, because the fats would quickly end up in our brains. Therefore, something in our body fats contributes to a one-phase system in which

this oil or fat is dispersed throughout the aqueous phase. A common example of this is in mayonnaise, where the fat is uniformly dispersed to give a one-phase system. We then turned to the problem of finding out what properties of fat particle in serum allow it to go to a one-phase system. We found that the major stabilizing factor for fats in the blood is protein. Thus, proteins give stability to blood fats, and it is the combination of fat with protein which gives the serum lipoproteins. However, we still do not know the details of the linkages and bondings. There are probably some polar bonds; also there are bondings of a nonpolar nature which keep a polar-type compound like protein linked to highly nonpolar materials like triglycerides or cholesterol esters.

Materials like proteins and lipoproteins are transported in the bloodstream in the form of large molecule groupings. That is, molecules of cholesterol do not travel alone. Salt ions cruise around by themselves, but fats travel in the form of large molecule groupings.

In order to study protein particles, we apply a whole battery of biophysical techniques such as the ultracentrifuge, diffusion apparatus, electrophoresis, paper chromatography, and electron microscopy. These specialized techniques are necessary because proteins are small—colloidal size. For example, using the ultracentrifuge we can produce a force 150,000 times as strong as gravity. After about 24 hours of this type of spinning, serum separates into a white layer on top, a clear solution next, followed by a bright yellow-orange layer, with a brownish layer underneath that (Fig. T-5). The white layer contains the large lipoproteins, because their density is only 0.8 to 0.9 g/cc, while that of the clear solution is about 1.0073 g/cc. The clear material contains no proteins and no extraneous material other than the salts that are contained in blood. The proteins, of density 1.3 g/cc, are in the bottom layers. By adding salts, one can control the density of the clear material in the intermediate layer and in this way float various particles from the lower layers to the top. When we raised the density of the clear liquid

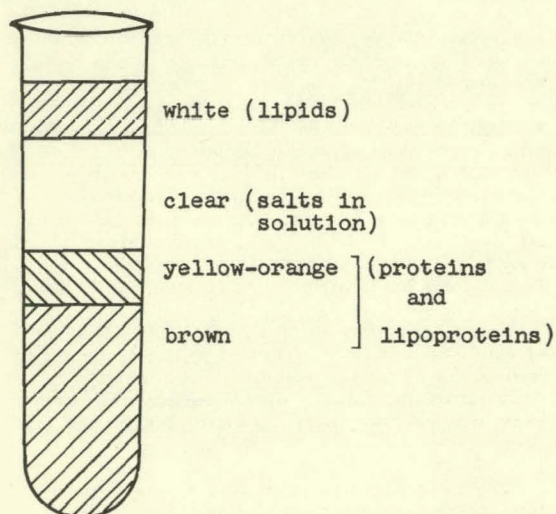


Fig. T-5. Distribution of human serum after being spun 24 hours in an ultracentrifuge.

to 1.06 g/cc, the orange zone floated to the top. This orange layer was analyzed and was found to contain the intermediate size lipoproteins. We again raised the density of the clear liquid, this time to 1.2 g/cc,

and the yellow layer floated up. It was found to contain a group of small lipoproteins. After this process has been repeated all that is left at the bottom are the proteins, the albumins, and the globulins that are contained in the blood. Thus, by this technique we are able to separate the lipoproteins according to their densities and sizes, and thereby characterize them. In fact, the analytic ultracentrifuge will perform this fractionation and then automatically give a graph showing the concentration of each fraction (Fig. T-6). The position of each peak indicates the density, and the area under the peak represents the concentration of molecules present in the sample. Three broad peaks occur in the density range 0.8 to 1.2, corresponding to the concentration of lipoproteins. However, the exact size and shape of these peaks varies from individual to individual.

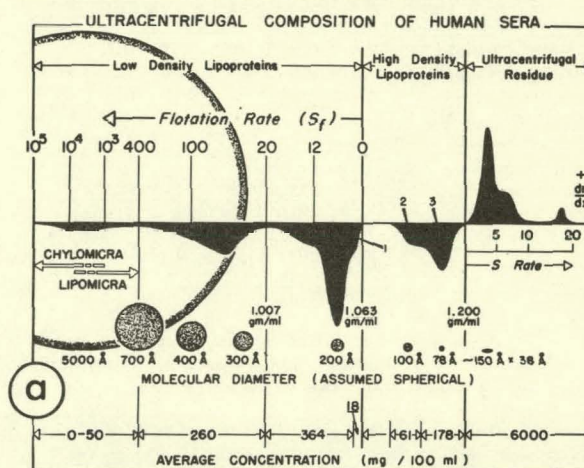
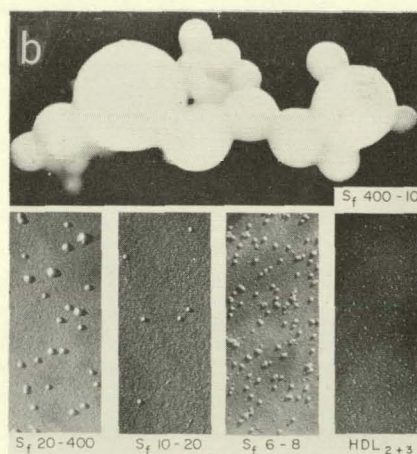


Fig. T-6. (a) Ultracentrifugal composition of human sera; (b) Electron micrographs of specific lipoprotein fractions. The S_f designations refer to rate of migration in the ultracentrifuge.



We ran a study of some 50,000 individuals to find out how the concentrations of these compounds are related to the eventual occurrence of coronary disease. We found that the large lipoproteins and intermediate-size lipoproteins have a direct correlation to the occurrence of heart attacks. This focused our attention on studying these two groups for their chemical structure and for their various physical properties.

The composition of the large lipoproteins is principally triglyceride, but some phospholipid, cholesterol, and protein are present. Triglyceride is the compound present in vegetable oil, coconut oil, cottonseed oil, corn oil, and safflower-seed oil. The principal component of the intermediate-size lipoproteins is the cholesterol esters--the steroid-type compound in the adrenal glands, i. e., the sex hormones. The small lipoproteins are made up mainly of phospholipids--the compounds that are prevalent in eggs. However, all the other lipids are present in a certain proportion, as is protein. The differences in densities between these three groups arise from the fact that there is more lipid and less protein in the light large group, and conversely, more protein and less lipid in the smaller dense group.

We found that the greater the concentration of large and intermediate-size lipoproteins (which include cholesterol), the greater is the probability of an eventual coronary attack. Interestingly enough, males are generally higher in the concentration of these two groups than are females until about the age of 50 or 60. Then the females begin to catch up. This would indicate that males under 50 to 60 have a higher incidence of coronary disease than do females, and this is actually the case (Fig. T-7). Above that age the incidence is about the same for males and females.

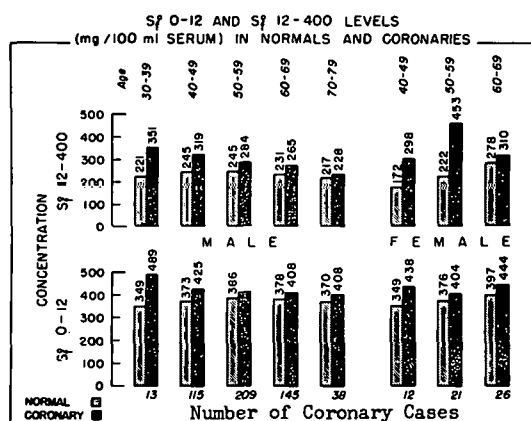


Fig. T-7. Incidence of coronary disease in males and females.

There seems to be little, if any, relation between the small lipoproteins and coronary disease.

How does diet affect these concentrations? We have run studies on samples of people, some of whom had diagnosed heart trouble. For example, we put overweight people (in whom the concentration of lipoproteins was high) on a low-calorie diet, one low in saturated fats and carbohydrates. The concentration of the large and intermediate-size lipoproteins was then observed to drop. Other studies showed

that if you are a person in whom the intermediate-size lipoprotein concentration is high, it will do no harm to reduce the amount of saturated fats in your diet. But if your lipoprotein level is normal, there probably is no reason to try to lower it. The foods that are high in saturated fats include ice cream, butter, lard, beef fat, etc. Those that contain unsaturated fat are vegetable oil, corn oil, peanut oil, cottonseed oil, etc. On the other hand, if you have high serum concentrations of the large lipoproteins, then it appears that lowering the intake of carbohydrates is beneficial. We found diets high in sugars and starches and low in fat associated with elevated concentrations of the large lipoprotein molecules. These are definite dietary approaches to controlling lipoprotein concentrations directed towards lowering the risk from coronary heart disease.

Lipids deposit not only in the arteries but also under the skin of the hands, eyelids, etc. Individuals who have these deposits in their externals will invariably have them in their arteries. These people also have a high incidence of heart disease.

One of the most compelling reasons for believing that proper diet is essential to avoiding heart disease is the following: Physicians have found that people with lipid deposits on the eyelids, elbows, hands, etc. will respond to dietary treatment. That is, if the lipid concentration in the bloodstream is lowered, in time these external deposits melt away.

Thus, there is hope that through dietary control, many people who would otherwise become victims of coronary disease will continue to live healthy lives. However, a great deal of study remains to be done.

Bibliography

1. O. DeLalla and J. W. Gofman, Ultracentrifugal Analysis of Serum Lipoproteins, in *Methods of Biochemical Analysis*, D. Glick, Ed., Vol. I (Interscience, New York, 1954).
2. J. W. Gofman, Diet in the Prevention and Treatment of Myocardial Infarction, *American Journal of Cardiology* 1, 271-283 (1958).
3. J. W. Gofman et al., Evaluation of Serum Lipoprotein and Cholesterol Measurements as Predictors of Clinical Complications of Atherosclerosis, *Circulation* 14, 691 ff. (1956).
4. Gofman, Nichols, and Dobbin, *Dietary Prevention and Treatment of Heart Disease* (Putman, New York, 1958).
5. Nichols, Dobbin, and Gofman, The Influence of Dietary Factors Upon Human Serum Lipoprotein Concentrations, *Geriatrics* 12, 7 ff. (1957).

U. ORIGIN AND EVOLUTION OF THE SOLAR SYSTEM

Otto Struve

When inquiring into the origin of the solar system, one recognizes immediately that its most important body is the sun itself. The sun contains 99.9% of the mass of the whole system. All the planets and their satellites, the asteroids, the meteors, and the comets taken together weigh only about 1/1000 as much as the sun. Therefore, if we want to learn the origin of the solar system, we must first seek to understand how the sun came to be what it is.

In our galaxy, the Milky Way, there are about 200 billion stars. In addition to these stars, the Milky Way contains a very large amount of gas and dust, in fact an amount roughly equal in weight to all the stars. Since the mass of the average star is 2×10^{33} grams and there are about 2×10^{11} stars, the total mass is about 4×10^{44} grams. Some of the gas and dust can be seen as a luminescent cloud in the vicinity of a very bright star. These clouds consist principally of hydrogen and helium, with all the other elements of the periodic table present in trace amounts.

It is reasonable to suppose that the stars (including our sun) originated as a result of a contraction of this gas. The force responsible for this contraction is gravitation. Opposing this contraction is the tidal force of the Milky Way as a whole. Figure U-1 shows a gas cloud at a certain distance a from the center of the galaxy. The solid arrows represent the force producing contraction, and the dashed arrows represent the tidal force that tends to pull the gas cloud apart. If the average density ρ of the gas cloud is great enough, the cloud will condense. Otherwise it will be disrupted. The condition for condensation is

$$\rho > \frac{4M}{a^3} \quad (1)$$

where M is the total mass of all the stars in the galaxy and a is the distance from the center of the galaxy to the center of the gas cloud.

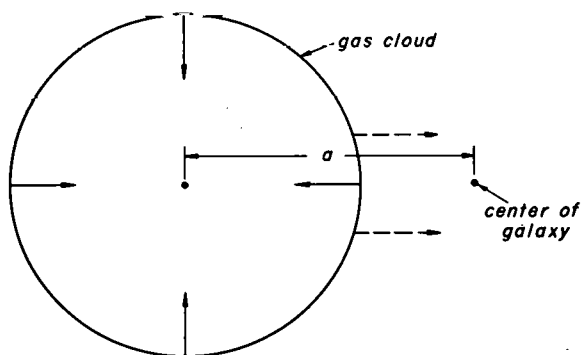


Fig. U-1. Diagram of a gas cloud, showing the conflicting forces.

As an example, consider the Crab nebula, a famous gas cloud. It is approximately 25,000 light years from the center of the galaxy, whose mass is 8×10^{44} grams. Putting these values into Eq. (1) (and supplying the necessary conversion factors), we

find that in order for this cloud to condense, its density ρ must be

$$\rho > 10^{-21} \text{ g/cm}^3 \quad (2)$$

This is a very low density. However, the average density of the nebula is equivalent to about one hydrogen atom per cubic centimeter. Since one hydrogen atom weighs only about 10^{-24} g, the density of the nebula is a factor of 100 too low for contraction.

There are, however, some gas clouds that have a density sufficient to contract. Astronomers have searched for objects that could have been formed as a result of such contraction. These are called "globules." A globule is a gas cloud that has contracted until it is nearly spherical in shape. It is not yet a star and is not luminous, but it is sufficiently stable gravitationally to resist the tidal forces of the Milky Way and it will continue to contract. When this process of contraction is completed, the globule will be a star. Our own sun undoubtedly came into being in this manner.

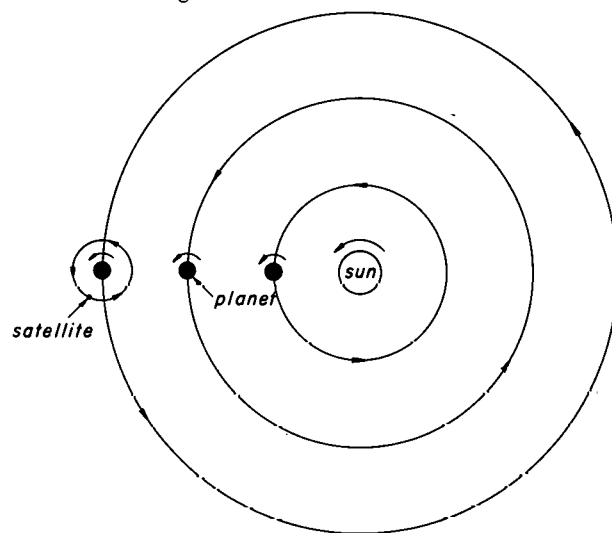


Fig. U-2. Simplified drawing of the solar system, showing the motions of the various bodies.

There are certain regularities in the motions of the planets and other bodies in our solar system that cannot be the result of chance. For example, all the planets revolve about the sun in approximately the same plane. (See Fig. U-2) Had the sun somehow captured the planets out of interstellar space, there would be no reason for the planetary orbits to be in the same plane, nor for the planets to travel in the same direction. Furthermore, many planets have natural satellites. The earth has one, Mars two, and Jupiter and Saturn many. With very few exceptions these satellites also travel around their respective planets in this same plane and direction. Also, each planet, as well as the sun itself, rotates in the same direction relative to the plane.

These facts suggest that the origin of the solar system must be related to the rotation of the primordial gas cloud out of which the sun and planets condensed. It is believed that the sun condensed from part of a gas cloud, and that the planets then condensed from the remainder of it. For the planets to form, the density relationship of Eq. (1) would have to apply. That is, the density ρ of the solar cloud must have been greater than four times the mass of the tide-raising body (the sun), divided by the cube of the distance between the solar cloud and the sun. The mass of the sun is 2×10^{33} g, and the distance a can be taken as 5 astronomical units, the radius of the orbit of Jupiter.* Thus, for planets to form from the solar cloud, its density must have exceeded 10^{-9} g/cm³.

The mass of the planets is known. If their mass were distributed evenly throughout the solar cloud, the density of the solar cloud would have been 10^{-11} g/cm³. Since this value is 100 times too low for condensation, we know that either the planets could not have condensed out of the solar cloud, or else we have overlooked some important factor. If we were philosophers we might say, "Therefore, the planets do not exist. There is no earth and no

people are on the earth." This conclusion, of course, is nonsensical. Somehow the planets must have condensed out of the solar cloud.

This problem was solved by Kuiper, an astronomer at the University of Chicago. He showed that only 1% of the original cloud went into the planets. The remaining 99%, mainly hydrogen, remained gaseous and was driven out of the solar system by radiation pressure from the sun. This conclusion also explains a commonly known fact, namely, that the planets are far less abundant in hydrogen than the sun. So far, this explanation is consistent with all the principles of astronomy and physics.

Bibliography

1. O. Struve, The Evolution of Stars, Scientific American, 34 ff. (Mar. 1953).
2. O. Struve et al., Elementary Astronomy (Oxford University Press, New York, 1959).
3. H. C. Urey, The Planets, Their Origin and Development (Yale University Press, New Haven, Conn., 1952); see also the review of this by O. Struve in Scientific American, 68-70 (Aug. 1952).

* One astronomical unit is defined as the distance between the sun and the earth.

V. THE ROLE OF SPACE SATELLITES IN GATHERING ASTRONOMICAL DATA

Otto Struve

The date of October 4, 1957, when the first Russian sputnik was launched, will be remembered in the history of astronomy forever and will play as great a role in that field as did the discovery of America by Columbus in the field of geography. It was the greatest single advance ever made in the field of astronomy. This development will enable us to gather a very large amount of information that cannot be obtained in any other way.

One of the greatest difficulties in astronomy is that the atmosphere of the earth is opaque to all radiations whose wave length is shorter than 3,000 angstrom units. Therefore, our knowledge of radiation from outer space is limited. The extreme ultraviolet light coming from stars and nebulae is unobservable because of this opacity of the atmosphere of the earth. The atom of hydrogen, for example, produces all kinds of radiations. When these radiations are observed with a spectroscopic, distinct lines can be seen that are characteristic of the element. The Balmer series of hydrogen is well known; also there is a series of lines in the infrared, and another in the extreme ultraviolet. These spectrum lines of hydrogen in the extreme ultraviolet can be observed in a physical laboratory because we can construct a spectrograph and evacuate it. Air is so opaque that even the small amount that is contained in a laboratory spectrograph would render these radiations invisible, even to photographic film. However, to observe the radiations from hydrogen that come from stars and nebulae, it is necessary to go outside the atmosphere of the earth.

This has been done, first, with high-flying rockets. When a rocket is 100 or more miles above the surface of the earth, the amount of air remaining above that level is no longer opaque, so that these radiations of hydrogen can be observed. The radiations of hydrogen are but one example of the many radiations that can be observed only in this way. Dr. Herbert Friedman of the Naval Research Laboratory recently gave an address in which he described observations of sunlight by means of rockets. He stated that they have recorded radiations down to wave lengths of much less than 1 angstrom unit, that is, radiations of the kind that we would call gamma rays. There are many important results that we want to obtain with regard to the source and the origin of these radiations from the sun.

A rocket remains at a high altitude for only a very short time. Therefore, there is usually not enough time to get observations of stars and nebulae. Sunlight is so strong that even during the few seconds or minutes that the rocket is high enough, spectrographic observations can be made. This is not true of starlight. In order to observe starlight of extremely short wave lengths, it is necessary to provide a satellite which will be in orbit for a fairly long time.

Such a satellite will contain a spectrograph and telescope, as well as other instruments. The satellite must be guided from the ground by means

of radio control, so that the spectrograph or telescope is always pointed at the right place. This engineering problem has, in principle, been solved. It is now possible to guide the satellite accurately while it is in orbit so that it will not turn around its axis or tumble but always have its telescope pointed in the direction that the astronomer wishes. This is only one of a great many problems that have already been attacked.

Of greatest interest at present is the hope that we may soon learn more about the possibility of life elsewhere in our solar system or on planets that may exist in connection with other stars. In our solar system there are only two planets other than the earth that conceivably could support life. These are Venus and Mars. The more distant planets--Jupiter, Saturn, Uranus, Neptune, and Pluto--are probably too cold to support life. Mercury, on the other hand, is very much too hot. But in the case of Venus, the surface temperature is close to the boiling point of water. We calculated this from our knowledge of the amount of heat that comes to the earth from the sun, relating this to the relative distances between the earth and sun, and Venus and the sun. However, we are not entirely certain of the properties of the atmosphere of Venus, and therefore are not sure that the actual surface temperature of Venus is what the calculations show. If it is above the boiling point of water, then there would be no life on Venus and probably there has never been any life on Venus. But it is possible that these calculations are not quite accurate enough and that conditions are still favorable to the possibility of life on that planet.

In the case of Mars, the results are more definite. In the first place, the atmosphere of Mars is quite thin, enabling us to see the surface of Mars. Because the atmosphere is thin, the calculations of the surface temperature of Mars are much more reliable than those of Venus. These results indicate that the average surface temperature of Mars is well below that of the earth; near the equator of Mars the temperature rises during the day appreciably above the freezing point of water, while at night the temperature is very low. This means that there is a marginal possibility for the existence of life on the planet Mars. Whether or not life exists there is unknown. There is some hope that satellite observations may give us an indication as to whether or not the kind of organic molecules exist on Mars that conceivably might be needed to form living organisms. To determine this, the infrared part of the spectrum would be observed. This portion is filtered out of the radiations from outer space by the water vapor in our atmosphere.

We can say definitely that no life can exist at distances closer to the sun than the orbit of Venus or greater than the orbit of Mars. Between these two planets there is a zone which my colleague, Dr. Huang of the astronomy department, has called the habitable zone of the solar system. Life could exist within this zone but cannot exist inside nor outside of it. Now, do other stars have planets in a habitable

zone? No star has ever been observed with a planet. This is not an indication that stars have no planets, but is readily explained in terms of the inadequacy of our telescopes. Even the nearest star to earth, which is the star called Alpha Centauri, is about 4 light years away from earth. Suppose that this star has a planet the size of Jupiter and that this planet is at the same distance from that star as Jupiter is from the sun. Then that planet would not be visible even with the largest telescopes now in existence. In fact, there is very little hope that such a planet could ever be observed from the earth either by looking through a telescope or by making a photograph, because it would be too faint.

There are, however, indirect reasons that make it appear very probable that the process of planet formation in connection with the formation of stars is a common phenomenon. This was not thought to be true a few years ago when the origin of the solar system was not explained the way I have tried to sketch in my previous lecture (Section U). Earlier, it was believed that the sun was first formed as a single object and that there were no planets associated with it. Then some time later another star happened to pass close to the sun; because of the gravitational forces between these two stars, material was ejected by the sun. This material then formed a gaseous mass that later condensed into planets.

However, collisions between stars are exceedingly rare. Moreover, it is absolutely certain that the sun has never had any collision with another star. It is also certain that the sun has in the past 5 billion years never had another star come sufficiently close to it to cause these prominences to erupt and form a cloud of gas, out of which planets could later condense. It was realized at the time this earlier theory was proposed that this was a rare phenomenon, but astronomers said, "Well, this seems to be the only way that we can account for the existence of planets." Since they knew that it was a rare phenomenon, they went on to say, "Well, the existence of planets must be an extremely rare phenomenon in the Milky Way." Thus, James Jeans and others even thought that perhaps the solar system with its planets is unique in the Milky Way, that there is no other star in the Milky Way that has ever undergone this kind of a process and has planets.

Other stars must have originated in the same general way as the sun; it is therefore quite probable that they also have planets. We now know a great deal about the properties of stars in the Milky Way, and there are many billions of stars that resemble the sun in every way that is subject to observation. There are billions of stars that have approximately the same mass as the sun, the same surface temperature, the same luminosity, and the same size. From these facts it would be entirely unreasonable to suppose that the sun differs from all other stars only by possessing a family of planets. It is much more reasonable to believe that because these billions of stars resemble the sun in every respect that we can detect, they must also resemble the sun with regard to planets, which we cannot now detect because our telescopes are not sufficiently powerful. But there is a further reason that makes this argument even more compelling, in fact so compelling that at present no astronomer has any doubt about the existence of many, many stars with planets, and the argument is essentially the following:

Let us suppose that there was a big cloud of gas, that this cloud had some slow rotation, and that

out of this cloud a star was formed, but no planets were associated with it. It is well known that if a large body having a slow rotation contracts, the small body that results from the contraction will rotate very fast. This phenomenon is known as the conservation of angular momentum. We know that gas clouds do have rotation, because we have observed them in the Orion nebula and elsewhere. If the nebula then condenses into a single star, the star must spin rapidly around its axis. Stars such as this have been observed. Some have rotational velocities of the order of 500 kilometers per second at the equator, and their period of rotation is of the order of a few hours. But the sun is not that kind of a star. The sun has a very slow rotation, turning around its axis once in a month. The question arises, "Why has the sun such a slow rotation?" It is unreasonable to suppose that the cloud out of which the sun was formed had no rotation, because if it had no rotation there would be no symmetry to the system of planets. Thus, the cloud must have had rotation to begin with. How could the sun have been formed with the extremely slow rotation that we now observe it to have, when the angular momentum of the original cloud must have been large? The answer is that the sun has planets. These planets have absorbed a large part of the angular momentum in their orbital motions.

We can test this hypothesis by making a computation of the following kind: Assume that the angular momentum of every planet: Mercury, Venus, Earth, Mars, Jupiter, and so on, was added to that of the sun. How fast would the sun then spin around its axis? It turns out that the sun would be a very rapidly spinning star of the kind that we observe in the Milky Way.

All the stars that resemble the sun in its spectrum, mass, size, and luminosity are stars of slow rotation just like the sun. Therefore, it is quite certain that all of these stars have deposited some of their angular momentum in other bodies, which presumably are planets. That is, the stars that resemble the sun and have slow rotation must be stars that possess planets. If this argument is true, and I am sure that it is, we must conclude that out of the 200 billion stars in the Milky Way, more than 50% possess planets. The other stars are rapidly-spinning stars and probably have no planets.

What is the possibility that life exists on any of these planets of other stars? It is in this connection that Dr. Huang has made some very interesting calculations. In the case of the sun, it will be recalled, the habitable zone lies between the orbits of Venus and Mars. The earth is about in the middle of this zone. If we consider another star that is very much hotter than the sun, this zone will be farther out; therefore, the volume of this zone will be much larger than the volume of the habitable zone of the sun. This would lead us to suspect that a very hot star has a better chance of possessing planets that are capable of supporting life. On the other hand, if we consider a star that has a lower temperature than the sun, the habitable zone will be quite small, so small in fact that the probability of a planet being in that zone is almost zero.

However, before reaching definite conclusions, we must consider yet another phenomenon, namely, the length of time that a star exists with a given temperature and luminosity. Calculations show that the process of condensation by which stars are formed is quite rapid. After a star has condensed

and become the kind of object that we observe it to be, the internal temperature of that star is sufficient to produce nuclear reactions. These nuclear reactions continue for very long time, as long as there is enough hydrogen to be converted into helium. The sun, for example, is already 5 billion years old, and it can continue shining at its present rate for perhaps ten times that interval. The period that it has already been shining was ample for permitting living organisms to be produced by chance encounters of atoms and molecules. One can only make a rough guess as to how long there has been life on the earth, but it may be on the order of a billion years. The very hot stars produce energy at such a tremendous rate by converting hydrogen into helium that they will not shine for billions of years, but burn out in a million years or less. Even though such a star has a very large habitable zone, the star does not shine long enough for life to develop. Therefore, we must exclude these hot stars. Also, we must exclude the stars that have a temperature much lower than that of the sun, because the probability of finding a planet within their very narrow habitable zone is extremely small. This leaves only the stars that resemble the sun as those which possess planets having conditions favorable to the development of life.

Within a distance of about 17 light years from the sun there are 42 stars. Of these 42 stars there are three that satisfy the conditions necessary to have planets on which life could develop. These three are the sun, a star designated tau in the constellation Cetus, and epsilon in the constellation Eridanus.

Of course, it is impossible at present to send a space vehicle to a planet 17 light years away. It is, however, possible to think of other methods of detecting the presence of intelligent life. Every radioastronomer in the world today is looking at his tracings of radioemission to be sure that he will not miss any radio signals from one of these stars that cannot be explained in terms of natural causes. This may sound far-fetched, but at the observatory where I am going, the National Radioastronomy Observatory in West Virginia, the astronomers have formed a cooperative which they call the "co-op of the little green men." The purpose of this group is to look for some sign of intelligence in these signals.

One of the most interesting problems is that of producing a satellite not around the earth but around the planet Venus. There are a good many reasons for trying to do this. One is that Venus has no natural satellites. In order to determine the mass and shape of a planet, as well as the concentration of matter inside it, we require observations of some satellite, be it natural or man-made. We have a great deal of information about the nature of Mars,

Jupiter, Saturn, and the earth from a study of their natural moons. But because Venus has no satellites we know very little about it.

Assuming that an artificial satellite could be made to orbit Venus, how would we then observe it? A satellite the size of the Russian Sputnik, or even a hundred times larger, would not reflect enough light to be visible by a telescope on earth. But if it were provided with a radio transmitter, these signals could be observed from the earth with radio telescopes. Radio signals have already been detected that were sent from the earth to Venus and reflected back, so there is no great problem about making a radio transmitter for the satellite of sufficient power to be observed from the earth. Unfortunately, radiotelescopes have a very low resolving power, i. e., they give a blurred image. One cannot tell where within that spot the signal comes from. But there is another method that would enable us to observe the motion of a satellite around Venus. That is to determine its Doppler effect. If the satellite were going away from us at the time of observation, then the radio signal would be of a lower frequency. If it were approaching us, the frequency would be increased. This change in frequency due to the Doppler effect can be measured with a very high degree of precision. By plotting this frequency change as a function of time, we would obtain a graph. The shape of this curve would indicate the physical features of Venus--its mass, shape, and concentration of mass within the planet.

Bibliography

1. L. C. Eiseley, Is Man Alone in Space?, Scientific American, 80 ff. (July 1953).
2. H. I. Ewen, Radio Waves from Interstellar Hydrogen, Scientific American, 42 ff. (Dec. 1953).
3. J. D. Kraus, Radio Telescopes, Scientific American, 36 ff. (Mar. 1955).
4. G. Reber, Radio Astronomy, Scientific American, 34 ff. (Sept. 1949).
5. J. A. Van Allen, The Artificial Satellite as a Research Instrument, Scientific American, 41 ff. (Nov. 1956).
6. J. P. Wild, Radio Waves from the Sun, Scientific American, 40 ff. (June 1955).

W. RADIATION AND LIFE IN SPACE

Cornelius A. Tobias

We cannot imagine life on this earth without radiation. Biosynthesis of some essential organic molecules proceeds by photosynthesis, thus making visible radiation essential to life. The more penetrating natural radiations from the radioactivity of the earth and from cosmic radiation have been present since the beginnings of the solar system and appear to have a definite role in causing genetic variation. Thus radiation is a factor in evolution. Studying the Universe, we must perceive at once that conditions on the surface of the earth are very special indeed to allow physical, chemical, and radiation environment so gentle and stable that in the span of perhaps less than two billion years a great manifoldness of living organisms has arisen.

As far as men can reach with the tools of astronomy, the Universe consists of highly dense matter in the stars which are in a state of thermonuclear interaction at many millions of degrees temperature. The space between the stars is a very high vacuum except where filled by cosmic gas or dust. The entire spectrum of known (and perhaps unknown) electromagnetic and corpuscular radiations is present in space. At present it is believed that most of matter and radiation is within the galaxies, and if this assumption is true, then we know that the energy content of the Universe is mainly contained in the mass of the stars and to a very small extent in the cosmic radiation.

We receive a good deal of radiation from the sun; part of the solar spectrum (visible light) is quite essential to life, being the main energy source of photosynthesis. The visible infrared and ultraviolet spectrum also contribute to the surface and air temperature and keep these in narrow range. The sun appears to be a sphere of 5800° Kelvin surface temperature, having a spectrum of emission of electromagnetic waves near the theoretical "black-body" emission. Most of the ultraviolet and x-ray components are absorbed in the upper atmosphere, and the visible and infrared rays deposit an average of 2 calories/sec-cm² on the earth's surface.¹ Deviations from the ideal "black-body" spectrum due to the higher temperatures of the solar interior, and the violent solar surface phenomena have been demonstrated in measurements from high altitude balloons and rockets,² and a summary of present-day knowledge is presented in Figure W-1. The most striking deviations are the emission lines from hydrogen in the solar atmosphere, the strongest of which is the Lyman alpha line in the far ultraviolet and the x-ray emission, coming in part from highly ionized atoms in the solar corona. During solar flares, x-rays up to several million volts in energy reach the upper atmosphere, and the x-ray intensity at some wavelengths can be as much as 100,000 times the black-body radiation.

The ultraviolet rays of the sun are in part responsible for dissociation of air molecules and the appearance of radicals of O, N, H, and their products³ above 50 kilometers altitude. Lower, in the stratosphere, ozone gas is present which in turn absorbs near ultraviolet radiations, thus protecting

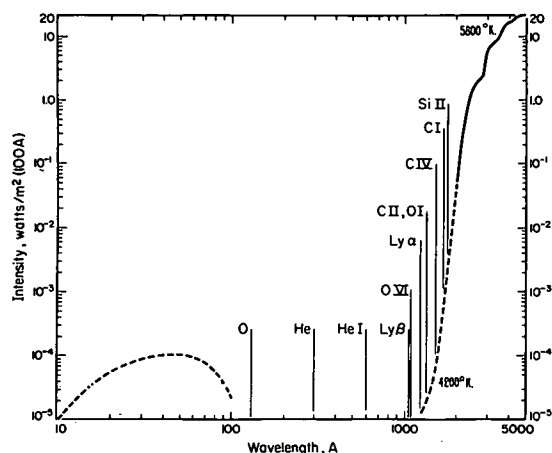


Fig. W-1. Intensity of vertically incident solar radiation in the spectral region below 3000 Å.²

life below. On the surface of the moon and on the planet Mars the ultraviolet radiation could seriously influence the existence of living cells. It is in the region below 3000 Å° where nucleic acids and protein absorb strongest and where lethal and sterilizing effects of ultraviolet are greatest.⁴ Thus, we can postulate that on Mars or the moon living cells would have to be resistant to ultraviolet rays, possess protective coating, or live in places sheltered from the direct rays. On the other hand, the surfaces of Venus and Jupiter appear to be well shielded from ultraviolet rays by their atmospheres.

Space ships can be shielded from ultraviolet rays and low energy x-rays; the more penetrating components of primary cosmic rays and their secondaries present a health problem. Actually the surface of the earth is quite well shielded from the more penetrating components of intragalactic radiation. The atmosphere absorbs most of the cosmic rays, and the earth's magnetic field deflects away most of the charged particles impinging on the surface. It appears that less than 2% of the primary cosmic ray intensity in space reaches us at ground level. These protective features are present to varying degrees on the other planets.

Primary Cosmic Rays

These are radiations which have their origin in extraterrestrial sources. Our present pictures of the primary cosmic radiation is that of rapidly moving atomic nuclei, stripped of their electrons and converging incessantly on the earth homogeneously from each direction in space.⁵ Protons are most abundant, and the frequency of heavier nuclei diminishes with increasing atomic number. In addition to the positively charged atomic nuclei electrons, fast neutrons, gamma rays, and antiparticles have been postulated as part of the primary rays, and we know that neutrinos are also present. Observation of the latter classes of primary particles has not as yet been successful because of their low abundance

and the fact that such measurements need to be done away from the disturbing influence of the earth's atmosphere and magnetic field, at distances of more than 50,000 miles.

Near the earth a mixture of primary and secondary cosmic radiations is observed. Part of the primary flux of particles is reflected and another part captured by the earth's magnetic field, thus causing a redistribution of particles. In the atmosphere, by inelastic collisions the primary particles break up into a number of different kinds of secondaries.

Table W-I lists the frequency of various atomic nuclei in the primary cosmic radiation at rocket altitude. The heavy-ion data were obtained by Yagoda⁶ in high-altitude rocket flights using photographic emulsions to record the tracks of the particles.

From the isotropic distribution of the primary rays, it seems clear that the majority of them must originate outside of the solar system, and in order to explain their presence one must propose a satisfactory injection and acceleration mechanism, as well as one for elimination of some of the particles. We know several types of stellar events that involve emission of ionized matter into space. Fermi suggested that collisions of particles with ionized magnetic clouds can cause their acceleration. From general considerations of energy density in space, it is now believed that most of the cosmic rays within our galaxy originate here and are contained within it⁸ by virtue of the magnetic fields at the edge of the galaxy. The particles may have a mean life of perhaps 10^6 years before they escape into intergalactic space. The strongest sources of cosmic rays in our galaxy are believed to be the radio stars, many of which are located near the center of the galaxy.⁹

Table W-I

Flux and distribution of charge at 41° N in heavy primary cosmic rays (6)	
Nucleus	Abundance %
Carbon, nitrogen, oxygen	58
Fluorine, neon	9.1
Sodium to silicon	20
Silicon to calcium	low
Calcium and iron	11
Total heavy flux $4.98 \pm 0.65 \text{ m}^{-2} \text{ sec}^{-1} \text{ sterad}^{-1}$	

The level of cosmic radiation is governed by the equilibrium between the generating process and the escape process, and locally on the presence of magnetic activity resulting from cosmic clouds. The above-outlined principles predict an energy distribution for the primaries in agreement with the experimentally found distribution at high energies per nucleon (above several Bev/nucleon). If $N(E)$ be the number of particles in energy range dE , then

$$\frac{dN}{dE} = -E^a,$$

where $a = -1.8$. This relationship is followed up to the highest primary events: $E \sim 10^{16}$ ev.

Radiation Belts

With the first successful American satellite, Explorer I, regions of high-intensity radiation around the equator were discovered, now bearing the name "Van Allen radiation belts" after the leader of the research group which made the original observations.¹⁰

Although much remains to be learned, we already have some detailed knowledge of the nature and origin of these rays, best summarized in Van Allen's own reports.¹¹⁻¹⁴

In Geiger-Mueller counters up to 600 kilometers, the counting rate is low, and the radiation is chiefly due to the penetrating component of cosmic rays, being about 15 millirem per day. It is believed that one could send a manned satellite to this region without undue radiation risk to the crew.¹⁵

Above 600 kilometers the radiation intensity doubles for each 100-kilometer altitude until the peak of the inner zone is reached at about 3400 kilometers. From here the radiation intensity decreases again to an altitude of 8000 kilometers, and then goes through a second peak at about 18000 kilometers. At 17 earth radii, the cosmic ray counting rate is down to about 2 per second. Actually some data are available up to 658,300 kilometers,¹⁶ which is well beyond the moon.

The moon has no appreciable influence on the intensity; and according to Van Allen the omnidirectional cosmic ray intensity is $1.8 \pm 0.3/\text{cm}^2 \text{ sec}$ (March 3-6, 1959), whereas Vernov et al. report a value of $2.3 \pm 0.1/\text{cm}^2 \text{ sec}$ on January 2, 1959. Rossi earlier assumed that the energy flux of cosmic radiation in space is

$$3.5 \times 10^{-3} \text{ erg cm}^{-2} \text{ sec}^{-1} \text{ ster}^{-1}$$

and that cosmic ray energy density in space is

$$1.4 \times 10^{-12} \text{ erg cm}^{-3} \sim 1 \text{ ev cm}^{-3}.$$

These figures correspond to about

$$7.2 \text{ primary particles cm}^{-2} \text{ sec}^{-1}$$

and to a dose level of about 25 millirad per day.* The measurements are about three times lower than the last figure. Because variations take place in the intensity, more data seem necessary to know the average level.

Figure W-2 presents our present idea of the distribution of radiation in the two radiation belts. There have not been enough rocket flights to know the exact distribution, but the actual distribution is at least as complex as the one shown.

Even before cosmic rays were discovered the existence of radiation ring currents around the earth has been postulated by Poincare, Stoermer, and others to account for minor variations in the earth's magnetic field. For some years both experimental¹⁷ and theoretical evidence was available to the effect that a magnetic field can trap charged particles injected at the proper angle and energy. The particles, once trapped, spiral around the magnetic lines of force. The shape of the earth's magnetic field is

* 1 rad = 100 ergs g^{-1} ; 24-hour day

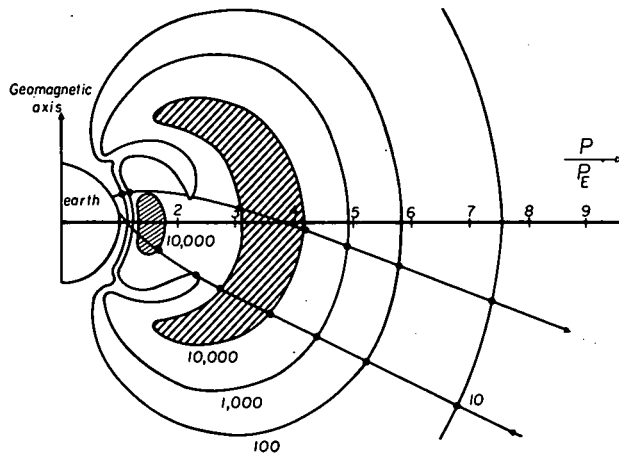


Fig. W-2. A plot in a geomagnetic meridian plane of the intensity structure of the radiation belts around the earth. The numbers associated with the several contours of constant intensity are the true counting rates in counts per second of a Geiger Mueller tube in Pioneer III or in satellite 1958E.¹²

such that some of the particles are reflected when they get near 40° latitude north or south. It is possible that particles moving in the earth's magnetic field will be accelerated also, owing to gradients in the magnetic field.

The radiation intensity in the Van Allen band is high because particles are trapped in a finite region of space. It is believed that at least in the outer zones renewal of the electrons and some of the positive particles happens when the earth collides with an ionized magnetic cloud originating perhaps from the sun. The outer belt with its low magnetic field appears to be more efficient in retaining electrons than positive nucleons, and very few if any of the latter are present.

There is direct experimental proof based on tracks of particles on photographic emulsions¹⁸ that the inner belt contains, in addition to electrons, protons up to 700 Mev kinetic energy. According to suggestions and theories by Christofilos,¹⁹ Singer,²⁰ and others, these protons are the result of radioactive decay of neutrons which are produced by inelastic collisions of primary cosmic rays in the upper atmosphere. The neutrons decay into protons, electrons, and neutrinos with a 12-minute half life.

Loss of particles from both radiation belts can happen along the magnetic lines of force near the north and south poles, particularly during magnetic storms. It is believed that the particles thus lost from the radiation belt are at least in part responsible for the auroral phenomena.

The storing capacity of the radiation belt for charged particles has been directly demonstrated in the "Argus" experiments originated by Christofilos²¹ when small atom bombs were exploded between the two natural radiation belts. As observed in satellites and rockets,²² artificial radiation bands resulted. By such means it is possible to cause regions around the earth with considerable radiation intensities and persistence — a potential man-made hazard to space travel.

Particles are lost from the lower radiation belt also by scattering with the atoms of the earth's atmosphere. This phenomenon explains the low level of radiation below 600 km altitude where air pressure begins to rise.

Table W-II summarizes the present status of our knowledge with respect to intensity levels in the radiation belts.¹⁴ Without shielding, the surface dose in the first 0.01 mm of skin would exceed 40,000 rad per second at the peak of the outer zone. With a shield of 1 g/cm^2 of light material, the inner zone would give rise to about 10 rad/hour, the outer zone on the order of 50 rad/hour. Because of the presence of high-energy protons in the inner zone, shielding there seems impractical. In the outer belt most of the electrons are stopped by 1 g/cm^2 absorber and further shielding is chiefly a matter of absorbing the secondary x-rays. Their half thickness in lead is about 0.7 g/cm^2 so that 4 g/cm^2 reduces the radiation to one rad per hour.

Table W-II

Van Allen Radiation Belts (August, 1959) ^a	
Inner Zone	
Altitude 3,600 km above surface	
Energy	Intensity
Electrons > 20 kev	$2 \times 10^9/\text{cm}^2 \text{ sec sterad}$
> 600 kev	$1 \times 10^7/\text{cm}^2 \text{ sec sterad}$
Protons > 40 Mev	$2 \times 10^4/\text{cm}^2 \text{ sec}$
Outer Zone	
Energy	Intensity per $\text{cm}^2 \text{ sterad}$
Electrons > 20 kev	10^{11}
> 200 kev	10^8
> 2.5 Mev	10^6
Protons > 60 Mev	$< 10^2$

^aJ. Van Allen and L. A. Frank.

Radiation measurements to 658,300 kilometers with Pioneer IV SUI 59-18, August 1959.

It would then appear that the radiation belts limit the desirable space in which manned space flight should be encountered to certain regions:

- Flight below 600 km altitude seems reasonably safe.
- A space rocket may safely leave the earth in a narrow cone near the magnetic poles.
- Shielding of the order of 10 g/cm^2 thickness removes most of the hazard from the outer radiation belt.

- d. Shielding of the protons of the inner radiation belt seems impractical.
- e. Beyond about 15 earth radii, the effects of the radiation belts may be crossed.
- f. The radiation belts may be crossed by a rapidly moving rocket at the expense of receiving a few rem dose.

Many points remain to be explored. We do not know for example if particles heavier than protons are present. Some temporal variation in the belts has already been observed, but much more knowledge is needed. Dosage during magnetic storms at high latitudes should also be explored.

The Heavy Primary Component and its Variations

With increasing knowledge of the radiation belt, attention is being focused again on the primary heavy particles. These particles have greater biological importance than their low relative abundance would indicate. They are important because their linear energy transfer (LET) or ionization density is much greater than that of protons or alpha rays.²³ A heavy nucleus of charge Z will have $LET \cdot Z^2$ times greater than a proton moving with the same velocity. For example, an iron nucleus ($Z = 26$) when completely stripped of electrons has linear energy transfer 676 times a comparable proton. Heavy primaries which represent only about 1% of the positive primary flux actually contribute between 1/3 and 1/2 of the total dose in unshielded situations. It is also necessary to know the "relative biological effectiveness" of the heavy ions.

Our knowledge of the low-energy end of the primary heavy-nucleon spectrum has been very limited. Owing to the geomagnetic cutoff, only nucleons higher than about 0.7 Bev per nucleon reach the top of the atmosphere in the temperate zones. Over the magnetic poles lower-energy particles could come in, and their absence led to an early conclusion to the effect that primary rays usually would not include many particles of low energy. Study of the cosmic ray variations has, however, shown that there are many low-energy particles generated by the sun. Such particles arrive with preference in the northern temperate zones. In order to completely assess the possible radiation hazard, we must know the charge spectrum and energy distribution of heavy primaries to levels lower than 100 Mev per nucleon.

Cosmic Ray Variations

Many important solar flares have been recorded during the past 15 years.²⁴ During an important solar flare in 1956,²⁵ there was an increase of as much as 30 fold in neutron intensity near ground level²⁶ within 15 minutes, reflecting similar increases in the primary particles (at 54°N , 71°W), whereas mesons increased only 58%.²⁷ The increase in the primary low-energy component, up to 4 Bev/nucleon, was much greater than that in the high-energy part of the spectrum. One of the most interesting aspects of cosmic ray increases after solar flares is that the rays strike the earth in very uneven distribution. Firor calculated the probable zones of impact of cosmic rays on the earth's magnetic field; some zones receive more dose than others. Geomagnetic latitudes between

25° and 60° exhibit an increase, while the equatorial North American continent falls in the heaviest irradiated zone. Existing data on the geographical distribution of the flare-type increases bear out the theoretical predictions, and thus we are now quite certain that the particles arriving associated with the solar flares do originate in the sun and are accelerated in a reasonably direct manner. Extrapolation of available data led Schaefer²⁸ to the speculation that during solar flares the energy (E) distribution of heavy primaries might vary as E^{-6} or E^{-7} . If this were true, we could expect during flares 10,000 times the usual dose in a space ship with 1 g/cm^2 shielding. Taking into account the duration of the flare, a person might receive as much as 25 rad, and in unusually large flares possibly 10 times that amount.

During the International Geophysical Year about one flare a month has been reported. In order to know the flare radiation spectrum in detail it is necessary to send rockets and balloons to high altitude during various phases of each flare. Since we do not know of advance signs of an impending flare, constant optical observation of the sun is used. The first successful interception of a solar flare was in July, 1959. In this event it was reported²⁹ that the low-energy component of primary cosmic radiation transiently increased by a very large factor. Measuring near the top of the atmosphere, Brown³⁰ found for the energy interval 100 Mev $< E < 400 \text{ Mev}$ that the number of protons increased like $E^{-4.5}$. It also seems reasonable that should one attempt to fly closer to the sun than we are at present, the cosmic rays of solar origin should increase in intensity at least as fast as the inverse-square-distance law would predict. Now a high-primary-radiation intensity in prolonged flight or in residence on one of the planets may make serious limitations on life span and evolutionary processes. Thus, solar cosmic rays should be considered in the definition and limitations of Strughold's Helioecosphere.³¹ Contributions of solar-flare doses of cosmic rays to the average dose level at ground level are not significant. The quantitative observations have been too recent, however, to predict what variation in the size of solar flares we may expect in times to come.

Flares are also accompanied by gamma rays probably secondary to nuclear interactions. Their energies of 3 to 5 Mev have been reported.³⁰

Magnetic Storms

Although high-energy components of a flare can reach the earth in minutes or hours, most of the ion plasma expelled from the sun adds to the solar corona, and some parts of it move in ionized "magnetic" clouds which reach the earth in about a day. Collision of these clouds with the earth's magnetic field appears to cause a variety of events; one of these appears to be the discharge of low-energy electrons and protons from the radiation belt near the polar zones with accompanying phenomena of aurora, redistribution of the Heavyside layers of electrons, radio blackout, and worldwide meteorological changes. At the same time the magnetic clouds appear to act as effective shields for high-energy galactic cosmic rays, so that at ground level the over-all cosmic ray intensity usually decreases for 1 to 3 days.

In addition to the solar flares, other types of solar events also influence cosmic ray intensity as reviewed by the author.³² Among these are variations with the 20-year solar cycle and a rather large variation with 11-year cycle corresponding to the frequency of sunspots.

The Biological Problem

If we are to insure radiation safety for man in rockets and satellites, there are important tasks remaining for the radiation physicist and biologist. Some of these will be listed briefly:

Dosimetry

In order to understand and measure the magnitude of biological hazard, we must be able to measure the all-essential factors. These include the measurement of the time rate of "absorbed dose," that is, the energy transmitted to each unit mass of tissue per unit time. We should also know the nature, energy, and spatial distribution of the radiations along with nuclear interactions they may cause in the atoms of tissue. The radiobiologist wishes to know the microscopic distribution of frequency of ionizing tracks with different ion density, that is, the distribution of "linear energy transfer." Complete understanding of these factors requires development of special dosimeters. These should accompany biological test object on satellite flights. Their presence seems essential for some purposes in view of the very large spatial and temporal variations in cosmic rays.

Shielding

It may be stated that man or most living organisms could not exist for long periods of time in extraterrestrial space without some radiation shielding. We already know that a reasonable shield (1 g/cm²) will stop a good deal of radiation, and with increasing knowledge, refined shields can be designed, even though some parts of space (e.g. the inner radiation belt) still remain hazardous.

Biological Effects of Heavy Ions

We do have actual and potential means to study effects of heavy cosmic ray primaries on accelerated beams at ground level. The effects of proton and alpha-particle beams have been under study for some 10 years using cyclotrons³³ and two linear accelerators which are available for study of heavy-ion effects. At the Berkeley HILAC machine, carbon, oxygen, neon, and argon beams are available with 10 Mev energy per nucleon. During the past two years the effects of these beams on some unicellular organisms,³⁴ phage, and enzymes have been studied, and a beginning was made toward studies on animal tissues, particularly skin and brain. The range of these particles is very limited, however, (500 microns in tissue for carbon) so that one should try to accelerate heavy nuclei up to levels of 1 Bev per nucleon in order to observe their effects on the whole animal.

When compared to effects of widely used radiations, heavy ions impress with their ability to inflict great local damage, distributed in a statistical manner to locations near heavy-ion tracks. It is believed that the greatest effect of such tracks is on cells essential to the whole organ-

ism, e.g. certain essential parts of the central nervous system (hypothalamus, brain stem, optic nerve) or on cells of the developing embryo. When the influence of radiation upon extraterrestrial life is being considered, many challenging problems appear. Can a single organism or a population adapt themselves to a high radiation level? The theoretical answer suggests that the highest level is the one where complete reproduction is possible for many generations before the dose received becomes lethal to the parent or, by genetic changes, to the offspring. For some lower organisms, for example yeast cells,³⁵ we have some idea about the magnitude of these levels. For mammals our knowledge is mostly by inference.

When man visits other planets of the solar system he may be confronted with primordial conditions on some planets, strange forms of life on others, and perhaps he will find planets where ancient life has become extinct. It seems to be of import to ascertain the role of penetrating and solar electromagnetic radiation to the generation of biological materials and to the mutation and evolution of the species.^{36, 37} Great strides are being made in these fields; still we must admit that we do not know the relative role of radiation as compared to chemical environment and temperature in evolution. The interplay of these factors during the last two billion years of our own history has been of decisive influence in bringing the living systems to their present state.

The fundamental problems just mentioned can be attacked vigorously in our present laboratories. No matter how far we can carry such investigations, it seems nevertheless certain that space travel will bring further surprises and much new knowledge.

References

1. Kopal and Zdenek, *Physics of the Sun*, Brit. J. App. Phys. 7, 119 (1956).
2. K. Watanabe, in *Advances in Geophysics*, E. Landsberg, Ed., Vol. V (Academic Press, New York, 1958) p. 153.
3. Joseph Kaplan, *The Chemosphere*. (Paper delivered at Second International Symposium on the Physics and Medicine of the Atmosphere and Space, November, 1958). In press.
4. A. Hollaender, Ed. *Radiation Biology*, Vol. II (McGraw-Hill, New York 1955).
5. P. Freier et al., Evidence for Heavy Nuclei in the Primary Cosmic Radiation, *Phys. Rev. Series II* 74, 213 (1948).
6. Herman Yagoda, Observations on Heavy Primary Cosmic Ray Nuclei Above the Atmosphere, *Geophysical Research Papers No. 60*, ASTIA Document No. AD-152585, 1958.
7. E. Fermi, On the Origin of Cosmic Radiation, *Phys. Rev.* 75, 1169 (1949).
8. F. DeHoffmann and E. Teller, Magneto Hydrodynamic Shocks, *Phys. Rev.* 80, 692 (1951).

9. Albrecht Unsold, Über den Ursprung der Radiofrequenzstrahlung und der Ultrastrahlung in der Milchstrasse, *Zeitschrift für Astrophysik* 26, 176 (1949).
10. J. A. Van Allen et al., Observations of High Intensity Radiation by Satellites 1958 Alpha and Gamma, *Jet Propulsion* 28, 588 (1958).
11. J. A. Van Allen et al., Radiation Observations with Satellite 1958E, *J. of Geophys. Res.* 64, 271 (March 1959).
12. J. A. Van Allen and L. A. Frank, Radiation Around the Earth to a Radial Distance of 107,400 KM, *Nature* 183, 430 (1959).
13. J. A. Van Allen and L. A. Frank, Radiation Measurements to 658,300 KM with Pioneer IV, State Univ. of Iowa Report SUI 59-18, August, 1959.
14. J. A. Van Allen, On the Radiation Hazards of Space Flight, State Univ. of Iowa Report SUI 59-7, May, 1959.
15. W. H. Langham, Implications of Space Radiations in Manned Space, Institute of Aeronautical Sciences Report 59-52.
16. S. N. Vernov et al., Study of Terrestrial Corpuscular Radiation and Cosmic Rays During Flight of the Cosmic Rocket, *Doklady Akademii Nauk SSSR* 125, 304 (1959).
17. J. W. Chamberlain, *Advances in Geophysics* (Academic Press, New York, 1958). 4:110.
18. S. C. Freden and S. White, Protons in the Earth's Magnetic Field, UCRL-5581T, May, 1959.
19. N. C. Christofilos, Quoted on p. 13 of J. Van Allen, The Geomagnetically-Trapped Corpuscular Radiation, State Univ. of Iowa Report SUI 59-16, 1959.
20. S. F. Singer, Radiation Belt and Trapped Cosmic Ray Albedo, *Phys. Rev. Letters* 1, 171 (1958).
21. N. C. Christofilos, The Argus Experiment, UCRL 5548, 1959.
22. J. A. Van Allen et al., Satellite Observations of Electrons Artificially Injected into the Geomagnetic Field, State Univ. of Iowa Report SUI 59-10, May, 1959.
23. C. A. Tobias, Radiation Hazards in High Altitude Aviation, *J. Av. Med.* 23, 345 (1952).
24. H. Elliott, *Progress in Cosmic Ray Physics*, J. G. Wilson, Ed. (Amsterdam, 1952), Chap. VII.
25. R. B. Brode and A. Goodwin, Jr., Extraordinary Increase of Cosmic Radiation of February 23, 1956, *Phys. Rev.* 103, 377 (1956).
26. J. A. Lockwood et al., Cosmic Ray Neutron Intensity Increase with Solar Flare of February 23, 1956, *Phys. Rev.* 103, 247 (1954).
27. J. Firer, Cosmic Ray Intensity Time Variations and their Origin, IV, *Phys. Rev.* 94, 1017 (1954).
28. Herman J. Schaefer, Solar Influences on the Ester Atmospheric Radiation Field and their Radiobiological Implications. (Paper delivered at Second International Symposium on the Physics and Medicine of the Atmosphere and Space, November, 1958). In press.
29. J. R. Winckler et al., *Phys. Rev. Letters*. In press.
30. R. R. Brown and D'Arcy, Observations of Solar Flare Radiation at High Latitude during the Period July 10-17, 1959, *Phys. Rev. Letters*. In press.
31. H. Strughold, The Ecosphere of the Sun, *J. Avia. Med.* 26, 323 (1955).
32. C. A. Tobias, Radiation Hazards in Space Flight, UCRL 8815, January, 1958.
33. C. A. Tobias and Tor Brustad, Radiobiological Studies with Accelerated Heavy Ions. (Paper delivered at Second International Symposium on the Physics and Medicine of the Atmosphere and Space, November, 1958). In press.
34. J. L. Born et al., Biological and Medical Studies With High-Energy Particle Accelerators, in *Progress in Nuclear Energy*, J. C. Bugher et al., Eds., Series VII, Vol. II (Pergamon, New York, 1959), pp. 189-206.
35. C. A. Tobias et al., The Action of Penetrating Radiations on Yeast Cells. (Proceedings of the Second Geneva Atoms for Peace Conference, United Nations.) In press.
36. S. L. Miller and H. C. Urey, Organic Compound Synthesis on the Primitive Earth, *Science* 130, 245 (July 31, 1959).
37. M. Calvin, Chemical Evolution of the Origin of Life, *American Scientist* 44, 248 (July 1956).