FIRST INTERNATIONAL E. COLI GENOME MEETING

CONF-9209255--Absts.

DE93 002367

SEPTEMBER 10-14, 1992
MEMORIAL UNION
UNIVERSITY OF WISCONSIN

The organizing committee extends particular thanks to the National Research Foundation, the National Center for Human Genome Research, and the Department of Energy for grant support which permitted this meeting to occur. We also wish to acknowledge the following corporations for generous support: Applied Biosystems, Clontech Laboratories, DNASTAR, Dupont, Fotodyne, Genetics Computer Group, New England Biolabs, Novagen, and Promega. Special thanks go to Professor Debby Joseph of the University of Wisconsin Computer Science Department for her invaluable help.

Bill Reznikoff
Ross Overbeek
Monica Riley
John Roth
Kenn Rudd

GENERAL INFORMATION

THURSDAY EVENING RECEPTION

There will be a buffet dinner reception on Thursday evening from 6:30-8:30 PM and all symposium participants are invited to attend. The reception will be held at the Tripp Commons in the Memorial Union. Registrants can register and pick up their symposium materials in the Tripp Commons anytime between 5 and 10 PM Thursday evening.

SUNDAY EVENING BANQUET

The Sunday evening banquet at the Tripps Commons of the Memorial Union will begin at 7:30 PM. There will be cocktails starting at 6:45 PM.

RESTAURANTS NEAR THE MEMORIAL UNION

There are numerous restaurants near the Memorial Union; many of them on State Street (one block south of the Memorial Union). The restaurants include fast food eateries, ethnic restaurants (Greek, Chinese, Turkish, Afghan, Mexican and Italian), and local delis. The symposium representative can give you more information during registration.

MESSAGES

Representatives from the Memorial Union Conference Office will post any messages for participants on a bulletin board next to the symposium entrance at the Great Hall or the Union Theater. Anyone needing to reach a symposium participant during the day can leave a message at 608-262-2755.

POSTERS

There will be two poster session; 2:00-5:00 PM on both Friday and Saturday afternoons. These poster sessions will occur in the room which immediately adjoins the Great Hall. Those attendees who will be presenting posters will be assigned to the appropriate poster session with their registration on Thursday. Posters can be set up anytime before 1:30 PM on the appropriate day using the abstract number to identify the correct poster board. All posters must be removed by 6:00 PM each day.

WORKSHOPS

There will be special workshops on Friday, Saturday and Sunday afternoons. Participants wishing to display software packages during the Friday and Saturday workshops should indicate this during the Thursday registration procedure.

■ ■ ■ ■ ■

FIRST INTERNATIONAL
E. COLI GENOME MEETING
SEPTEMBER 10-14, 1992
MEMORIAL UNION
UNIVERSITY OF WISCONSIN

PROGRAM

Thursday, September 10 - (Tripp Commons)

4:00-10:00 PM	Registration
6:30-8:30	Welcoming Reception Buffet Dinner

Friday, September 11 - (Great Hall)

7:45-8:30 AM	Continental Breakfast
8:30-8:45 AM	Welcome (Bill Reznikoff)

Large Scale Sequencing Efforts - (Great Hall)

8:45-9:15 AM	K. Mizobuchi - Systematic sequencing of the E. coli genome: analysis of the 0-2.4 and 4.2-6.2 min regions
9:15-9:45	F. Blattner - The E. coli genome project
9:45-10:15	K. Isono - Development of a transposon- and PCR-based sequencing strategy and its application to the analysis of the terC region of E. coli K-12
10:15-10:45	Coffee break
10:45-11:15	G. Church - Automated multiplex sequencing of bacterial genes and proteins
11:15-11:45	P. Gillevet - Multiplex genomic walking: Integration of the wet lab and computer lab into a single prototyping environment
2:00-5:00 PM	Workshop-Tools for Sequence Analysis (Old Madison Room) Poster Session and Industrial Exhibits (adjoining Great Hall)

Molecular Evolution - (Great Hall)

7:00-7:30 PM	D. Dykhuizen - Gene Variability, clonality and recombination with E. coli
7:30-8:00	C. Hill - The Rhs elements of E. coli: Complex composites of homologous and unique sequences
8:00-8:30	R. Milkman - Molecular evolution of the E. coli chromosome
8:30-9:00	Coffee break
9:00-9:30	H. Ochman - (To be announced)
9:30-10:00	G. Olsen - Molecular phylogeny: Tracing the origins of organisms and their genes
10:00-10:30	E. Ohtsubo - Involvement of IS and transposable elements in rearrangements of the E. coli genome

Saturday,September 12 - (Union Theater)

7:45-8:30 AM Continental breakfast

Databases - (Union Theater)

8:30-9:00 AM M. Berlyn - CGSC database for genotypes, genes, products, maps: Description and demonstration
9:00-9:30 A. Danchin - Colibri: A functional database for the Escherichia coli genome
9:30-10:00 M. Kroger - ECD: Collection and management of E. coli DNA sequence data and a proposal for an integrated database for E.coli
10:00-10:30 Coffee break
10:30-11:00 F. Neidhardt - The gene-protein database of Escherichia coli
11:00-11:30 R. Overbeek - An integrated genomic database (IGD)
11:30-12:00 K. Rudd - Informatics analysis of the E. coli genome
12:00-12:30 M. Riley - Database of all currently identified E. coli gene products
2:00-5:00 PM Workshop - Using Databases (Old Madison Room)
Poster session and Industrial Exhibits (adjoining Great Hall)

Molecular & Genetic Approaches to Chromosome Structure -(Union Theater)

7:00-7:30 PM D. Berg - Transposon facilitated sequencing of cloned DNAs
7:30-8:00 M. Casadaban - Plaque forming Mu elements for gene fusing and cloning
8:00-8:30 J. Miller - DNA Rearrangements at short homologies
8:30-9:00 Coffee break
9:00-9:30 J. Roth - Genetic approaches to genome stability in Salmonella
9:30-10:00 G. Stormo - Computer methods for identifying interesting sequences

Sunday, September 13 - (Great Hall)

7:45-8:30 AM Continental breakfast

Global Regulatory Networks - (Great Hall)

8:30-9:00 AM T. Ikeda - Is nitrogen limitation sensed via the glutamine pool?
9:00-9:30 E. Lin - The ARC signal transduction system for global respiratory control of gene expression in Escherichia coli
9:30-10:00 B. Wanner - Using sequence "tags" to identify new Pi-regulated genes and genes for Pi- independent control of the pho regulon
10:00-10:30 Coffee break
10:30-11:00 J. Wild - Regulation and function of E. coli heat shock proteins
11:00-11:30 M. Weickert - The galactose regulon of E. coli
2:00-5:00 PM Workshop - Sequencing Technology (Old Madison Room)
6:45-9:00 Banquet - (Tripp Commons)

Monday,September 14 - (Great Hall)

7:45-8:30 AM Continental breakfast

Chromosomal Structure - (Great Hall)

8:30-9:00 AM K. Drlica - Control of DNA supercoiling
9:00-9:30 N.P. Higgins - HNS protein constrains DNA supercoils in vitro and modulates global transcription in vivo
9:30-10:00 D. Pettijohn - Bacterial chromosome structure and targeting of DNA-binding proteins
10:00-10:30 Coffee break
10:30-11:00 R. Wells - Supercoiling stabilizes Z-DNA in E. coli
11:00-11:30 J. Lupski - Conservation and distribution of short, interspersed repetitive DNA sequences in eubacterial genomes

Wrap Up

11:30-1:00 Open discussion - led by Bill Reznikoff

■ ■ ■ ■ ■

Abstracts of Talks

T 1

SYSTEMATIC SEQUENCING OF THE ESCHERICHIA COLI GENOME: ANALYSES OF THE 0-2.4 AND 4.2-6.2 MIN REGIONS

K. Mizobuchi[1*], B. Fukuda[2], H. Inokuchi[3], K. Isono[4], K. Makino[5], T. Miki[6], T. Mizuno[7], H. Mori[8], T. Nagata[1], A. Nakata[4], Y. Yamamoto[9], and T. Yura[8], [1]Dept. of Biophys. and Biochem., Faculty of Sci., Tokyo Uno., [2]Dept. of Biochem., Kanazawa Uei. School of Medicine, [3]Dept. of Biophys., Kyoto Uni., [4]Dept. of Biol., Kobe Uni., [5]Inst. for Microbial Diseases, Osaka Uni., [6]Dept. of Bacteriol., Kyushu Uni., [7]Dept. of Agricul. Chem., Nagoya Uni., [8]Inst. for Virus Res., Kyoto Uni., and [9]Dept. of Genetics, Hyogo Medical College.

Using the Kohara lambda clone library, we determined two contiguous 111,402- and 86,114-bp sequences corresponding to the 0-2.4 and 4.2-6.2 min regions of the E. coli chromosome, respectively. When the DNA sequences were analyzed, a number of novel genes were found whose deduced protein sequences were significantly similar to known proteins from various organisms. Those include, for example, a cluster of 4 genes responsible for the symbiotic nitrogen fixation or electron transport (fixA, B, C, X) in Azorhizobium caulinodans. We also found a new 120-bp repetitive sequence distributed among enteric bacteria. In addition, several IS elements and genes such as leuA, B, C, D and rrnH that had been mapped but notsequenced were identified. There are about 90 and 75 genes in these two regions with little spacing.

T 2

THE *E. COLI* GENOME PROJECT

D. L. Daniels, V. Burland, G. Plunkett III, and Frederick R. Blattner*.
Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA.

The genome of E. coli is being sequenced at Wisconsin by a combination of random and directed strategies. A set of overlapping lambda clones from wild type strain MG1655 spanning the 1 megabase region from 80 to 100 minutes are being sequenced initially. Synopsis of this analysis will be presented for the 200 kbp region from 82 to 86 minutes.

Each lambda clone is shotgunned into M13 and sequenced at random to a redundancy of 4 to 6. Then, using site specific recombination sites in our Janus M13 vector to invert the inserts, selected clones are sequenced on the opposite strand. We are using a highly automated adaptation of the radioactive Sanger sequencing method with a film scanner. Since the ambiguity rate within assembled sequence from our production line at this point is about 1 in 200 bases, and insertion-deletion errors are common, considerable effort must be expended to analyse open reading frames, codon usage statistics and protein database homologies to produce a final edited and annoted sequence suitable for release to GenBank. Protein coding regions constitute about 85% of this genome. About 2000 new proteins are expected to be discovered. Given the extensive body of data on E. coli, the achievement of the complete genomic sequence is sure to suggest further biological experiments. Such logical extensions will require a level of accuracy that may be difficult to justify in the case of larger genomes.

T 3

DEVELOPMENT OF A TRANSPOSON- AND PCR-BASED SEQUENCING STRATEGY AND ITS APPLICATION TO THE ANALYSIS OF THE *TERC* REGION OF *E. COLI* K-12

Kasai, H.[1], Isono, S.[2], Kurnit, D.[3], Berg, D. E.[4], and Isono, K.[1, 2] * [1] Postgrad. Sch. and [2] Dept. Biol., Fac. Sci., Kobe Univ., Rokkodai, Kobe 657, Japan; [3] Howard Hughes Med. Inst., Univ. Michigan Med. Cent., Ann Arbor, Michigan 48109-0650,U.S.A.; [4] Dept. Mol. Microbiol. and Genet.,Washington Univ. Med. Sch., St. Louis, MO 63130, U.S.A.

We have developed a strategy for efficient sequence analysis of the genome of *E. coli* K-12, by using a modified version of mini-*Tn*5 transposon of Phadnis et al. (1989) and the ordered λ phage clones of Kohara et al. (1987). Transposon was serially introduced into each of the desired clones. Transposon-containing clones were selected by blue plaque formation on a $dnaB_{amber}$ $lacZ_{amber}$ *E. coli* strain and their insertion points were determined by "analytical PCR". The segments between the transposon inserts and λ phage arms were amplified by "preparative PCR" using a biotinylated primer. Amplified DNA fragments were coupled to paramagnetic beads through their biotin tail, collected and purified magnetically. The biotin-tagged strands were then used as templates for fluorescence-based automatic nucleotide sequencing. The strategy was applied to the analysis of the *E. coli terC* region, using a total of 28 Kohara clones which cover approximately a 150 kb-long stretch. The current status of this project will be reported.

T 4

AUTOMATED MULTIPLEX SEQUENCING OF BACTERIAL GENES AND PROTEINS

George M. Church*, Nathan D. Lakey, Hamid Ghazizadeh, Laura Jaehn, Peter Richterich@, Keith Robison, Andy Link. Harvard Medical School and Howard Hughes Medical Institute, Boston, MA 02115. @Collaborative Research Inc., Waltham, MA.

We have recently integrated direct transfer electrophoresis, automated multiplex hybridizations and automated film reading to sequence scattered E.coli genomic DNA and three cosmids. Sequence patterns for two cosmids were detected using chemiluminescence with oligonucleotide probes directly conjugated to alkaline phosphatase. Primers for the directed walking and dideoxy sequence confirmation steps were synthesized with 15 base tags for detection. For the cosmids, 20 gels resulted in 9216 sequences on film. Film data were automatically read and assembled using the programs REPLICA and GTAC. Another program automatically finds and graphically annotates ORFs including matches to database sequences. Among many other features in the cosmids, this turned up new potential operon(s) covering 13 kbp of long ORFs without homologs in any databases. A database of protein abundances, amino-termini, pIs, MWs, and sub-cellular-fractions has been used in several sequencing and model building projects.

T 5

MULTIPLEX GENOMIC WALKING: INTEGRATION OF THE WET LAB AND COMPUTER LAB INTO A SINGLE PROTOTYPING ENVIRONMENT

P.M. Gillevet*, S. Smith, A. Ally, F. Barton, R. Clark, M. Dolan, E. Hsu, L. Marquez, M.S. Purzycki, J. Sartell, B. Richter, J. Williams, C. Wang, and W. Gilbert Harvard Genome Laboratory, Harvard University, Cambridge, MA 02138 USA

We are presently sequencing the entire genome of Mycoplasma capricolum, one of the smallest of free living organisms by a Multiplex Genomic Walking strategy. This technique involves the repetitive hybridization of sequencing membranes with oligonucleotide probes to acquire sequence data in discrete steps along the genome. The technique allows one to walk a genome in a directed manner eliminating the problems associated with random shotgun assembly. Furthermore, the repetitive stripping and hybridization process is relatively simple to reproduce and has the potential to be easily automated. The Genetic Data Environment (GDE), a X Windows based Graphic User Interface has allowed the seamless integration of a core multiple sequence editor with pre-existing external sequence analysis programs and internally developed programs into a single prototypic environment. This system has facilitated linkage of the Harvard Genome Lab's internal database and automated data control systems into one Graphic User Interface which can handle the archiving and analysis of both random fluorescent sequencing data and genomic walking data from the Mycoplasma project. Finally, it has facilitated the integration of the Genomic sequence data into a PROLOG database environment for the Comparative analysis of Mycoplasma capricolum and other organisms.

T 6

GENE VARIABILITY, CLONALITY AND RECOMBINATION WITHIN ESCHERICHIA COLI.

Daniel E. Dykhuizen*. Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, N.Y. 11794.

Repeated DNA sequencing of the same genes from different strains of clonally reproducing organisms can be used to determine rates of mixis (genetic exchange), and to infer the various evolutionary pressures on the species. Evidence for mixis can be obtained from comparing gene trees. In clonally reproducing organisms like E. coli, trees constructed from the different genes should not be significantly different, given that there is no mixis. While, if the trees are significantly different, this is evidence for mixis. What is meant by significant in this context is a difficult problem since it includes not only problems of statistical sampling but also the model of the evolutionary process used, which must include assumptions about the processes of mutation and selection. Sequence data on 5 genes will be presented and related to recombination, clonal frame and rates of evolution.

THE *Rhs* ELEMENTS OF *E. coli*: COMPLEX COMPOSITES OF HOMOLOGOUS AND UNIQUE SEQUENCES Charles W. Hill. Department of Biological Chemistry,Penn State University College of Medicine,Hershey,PA,17033,USA

The five *Rhs* elements collectively comprise nearly one per cent of the *E. coli* K-12 genome. They are classified as accessory genetic elements because they are not present in all wild *E. coli*. They share a 3.7 kb homology, called the *Rhs* core, but their total size is much larger, the largest, *RhsC*, being 9.6 kb. The cores fall into two subfamilies based upon sequence similarity, with the *RhsA-B-C* core subfamily diverging about 22% from the *RhsD-E* subfamily. Each core maintains a single ORF, the start codon coinciding with the first base of the homology. The respective core ORFs then extend between 139 and 177 codons beyond the ends of the homologies into unique sequence. Thus they appear capable of producing large proteins (>156 kD), with nearly identical 141 kD amino-terminal portions and a variable carboxy-terminal. Three of the K-12 *Rhs* elements share another large homology, the 1.3 kb H-repeat which contains a 1134 bp ORF. The H-repeat structure differs from that of the core in that the homology begins 144 bp before the initiation codon. The 1% sequence divergence between the H-repeats of *RhsB* and *RhsE* contrasts with the 22% divergence between their cores, suggesting that these composite elements have been assembled from components with very different evolutionary histories. The *Rhs* elements differ from other accessory genetic elements in their natural distribution. This is seen from the fact that the *Rhs* profiles of strains in the ECOR collection correlate closely with the phylogeny established through enzyme electrotyping. Two examples have been found in which the core extensions found in a particular element in K-12 are associated with a different element in one of the ECOR strains.

MOLECULAR EVOLUTION OF THE *E. COLI* CHROMOSOME

*Roger Milkman**. Department of Biological Sciences, The University of Iowa, Iowa City, IA 52242-1324 USA

Comparative sequencing of some 37 wild strains over an 8-kb stretch reveals a mosaic pattern of distinct *sequence types*, in which various segments up to 1 kb or more are either embedded in an extensive *clonal frame* or are contiguous. This pattern of small segments, evidently recombinant replacements, contrasts with the much larger size of entrant DNA fragments associated with conjugation and transduction, the known chromosomal recombination mechanisms of *E. coli*. Possible explanations are: 1. direct incorporation of small entrant molecules; 2. reduction of originally large replacements by repeated overlapping events; and 3. incorporation of a discontinuous set of small fragments from a single large entrant molecule. Direct and indirect means of deciding among these are discussed. The nature of recombination, qualitative and quantitative, in *E. coli* is critical to the reconstruction of its microevolutionary history from the observed patterns and estimated nucleotide susbtitution rate. -- A case is also made for the treatment of homology as a quantitative variable (in terms of recency of common ancestry).

MOLECULAR PHYLOGENY: TRACING THE ORIGINS OF ORGANISMS AND THEIR GENES

Gary J. Olsen.* Department of Microbiology,
University of Illinois, Urbana, IL 60801, USA

The most common use of molecular phylogeny is the comparison of broadly distributed genes (such as those for ribosomal RNAs, *c*-type cytochromes, or translation elongation factors) to study the relationships of organisms. This has been immensely successful in helping to infer the origins of and relationships among extant life-forms. In addition, sequence-based phylogenetic identification of microorganisms has led to the development of new medical diagnostics and new approaches to microbial ecology.

As a greater variety of genes and proteins are sequenced, molecular phylogeny plays another important role: it can be used to examine the relationships of the individual pieces of the genome to one-another. In this manner, molecular phylogeny addresses the questions "What is the origin of a given gene?" and, more broadly, "How did early organisms with a much simpler set of genes and functions evolve into the complex organisms and genomes that we observe today?"

I will explore some of the conceptual and practical issues in treating sequence and genome analysis as problems of phylogenetic inference.

INVOLVEMENT OF IS AND TRANSPOSABLE ELEMENTS IN REARRANGEMENTS OF THE E. COLI GENOME

Eiichi Ohtsubo* and Masaaki Umeda, Institute of Applied Microbiology, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113, Japan

The E. coli chromosome contains many kinds of insertion sequences (IS), such as IS1, IS2, IS3, IS5 and IS30 in multiple copies, and a transposable element $\gamma\delta$. The chromosomal locations and orientations of most IS elements in the E. coli K-12 chromosome have been recently determined, and the nucleotide sequences of the flanking regions of some elements have been also determined. The map positions of IS elements could be correlated well with the positions where various illegitimate recombinational events have occurred within the chromosome or between the chromosome and other genomes such as sex-factor plasmids. These events include Hfr formation between the F plasmid and chromosome, formation of F' plasmids from Hfr cells and of several kinds of λ transducing phages, formation of tandem duplications and inversions, and deletions and transpositions of chromosomal DNA segments. IS elements and $\gamma\delta$ mediate these events either by transpositional recombination or by homologous recombination at any pair of copies of an IS element.

T 12

CGSC DATABASE FOR GENOTYPES, GENES, PRODUCTS, MAPS: DESCRIPTION & DEMONSTRATION

Mary B.Berlyn*, Dept. of Biology & School of For.- Envir.St.,Yale Univ., New Haven, CT 06511 and Stanley Letovsky, Letovsky Associates, 286 W.Rock Ave., New Haven, CT 06515

The CGSC Database is a relational representation of the repository of genetic information developed at the *E. coli* Genetic Stock Center by its curator, Dr. Barbara Bachmann. The database follows the original strain descriptions by describing strain genotypes for 4500 strains of *E. coli* K-12 in terms of allelic and structural mutations, mating type, and plasmids and documenting them with references and information on originators and derivation of strains. It also describes genes, map location, gene function and products. Mutation subtypes include basepair substitutions, inversions, insertions, transpositions, and deletions of chromosomal segments, the special insertions of or within the F plasmid in Hfr and F' strains, and others. Each mutation is described as a change in an interval on the chromosome defined as the entity-type "Site" (which includes but is not restricted to genes). Sites have endpoints that are assigned coordinates on a map. References and source-persons can be linked to any of the major database objects. Uncertainties in genotype or map location are explicitly represented. The database is used to query for strains with a given combination of alleles, mutation types or properties; for isogenic pairs of strains; for deletions or plasmids spanning specific genes or regions; for derivations; for map positions or other site properties; and for gene product information. Graphical enhancements produce regional maps of genes and insertions and prepare high-quality formatted reports. We have developed and incorporated into the database software mapping utilities intended to automate and enhance the analysis and integration of diverse mapping data. Integrated maps of regions, the conflicts and uncertainties associated with constructing them, and the query capabilities of the database will be demonstrated.

T 13

COLIBRI : A FUNCTIONAL DATABASE FOR THE ESCHERICHIA COLI GENOME

C. Médigue[1], A. Viari[2], A. Hénaut[3] & A. Danchin[1]*. [1]Institut Pasteur, Paris, France. [2]Institut Curie, Paris, France.[3]Centre de Génétique Moléculaire, CNRS, Gif sur Yvette, France

Several data libraries have been created to organise the many data obtained worldwide on the *E. coli* genome. Because this now amounts to 40% of the whole genome sequence it becomes necessary to organise the data in such a way that appropriate procedures can associate knowledge produced by experiments about each gene to its position on the chromosome and in relation to other relevant genes, for example. In addition global properties of genes, affected by the introduction of new entries, should be present as appropriate description fields. A database, implemented on Macintoshes using the Database Management System 4th dimension, will be described. It is constructed around a core constituted by all known contigs of *E. coli* sequences, and links data collected in general libraries (unmodified) to data associated with evolving knowledge (with modifiable fields). Biologically significant results obtained through the coupling of appropriate procedures (learning, or Factorial Correspondance Analysis) will be presented. The database is available through a 4th Dimension runtime, and through FTP on Internet.

ECD : Collection and Management of *E.coli* DNA sequence data and a proposal for an integrated database for *E. coli*

Manfred Kröger, Ralf Wahl, Gabriel Schachtel and Peter Rice
Institut für Mikrobiologie und Molekularbiologie, Justus-Liebig-University Giessen, and EMBL, Heidelberg, Germany

We have collected the DNA sequence data of *E.coli* K12 over the past years [1]. The actual status at July,7th is 40.6 % (1914 826 bp) plus 2% from other strains. We provide a quaterly update of our ECD database together with the CD-ROM version of the EMBL database. The DNA sequences may be assembled as contigs or may be read as individual EMBL files. Crossreferences allow the use of GenBank files. References for sequenced mutants as well as recent mapping data are incorporated. The genetic map data of B.Bachmann and the 2D-protein index data of VanBogelen and Neidhardt are completely integrated. We work on integration of Kohara map data. The database may be used for assembly of most actual contigs and statistical purposes [2]. It may used for searches on map data, on genetic names, on references, on protein data or on database entry names or accession numbers, respectively. We like to propose the building of an international integrated *E.coli* database for all data - including X-ray data - concerning structural and physiological questions and crossreferences to other bacteria.

[1] M.Kröger et al., Nucleic Acids Res. **20** (Suppl.) 2119-2144 (1992).
[2] R.Merkl et al., Nucleic Acids Res. **20**, 1657-1662 (1992).

THE GENE-PROTEIN DATABASE OF *ESCHERICHIA COLI*

F. C. Neidhardt*, R. A. VanBogelen and R. L. Clark.
Department of Microbiology and Immunology,
University of Michigan, Ann Arbor, MI 48109-0620, USA

The Gene-Protein Database of *E. coli* is both an index relating genes to their protein products as they appear on two-dimensional polyacrylamide gels, and a catalog of information about the function, regulation and genetics of individual proteins collated from the literature or generated by two-dimensional gel analysis. When work began on the database 15 years ago, individual protein spots were usually identified one-at-a-time as known proteins or the product of known genes using several chemical, genetic and physiological procedures to accomplish the identifications. This work continues, supplemented by improvements in sequencing small amounts of proteins obtained from gels. Increasingly, however, progress is coming from systematic *expression analysis* of genomic segments and from the translation of open reading frames, using procedures made possible by the availability of ordered libraries of chromosome segments, significant improvement in methods for selectively expressing cloned segments in vivo, and the accumulation of extensive DNA nucleotide sequence information. Hardcopy editions of the database are being published annually (*Electrophoresis*, reprints available); the database is also available electronically and on CD-ROM from the National Center for Biotechnology Information. The recently completed 5th edition of the database contains information on 663 protein spots (including 383 identified by gene or protein name), representing at least one-sixth the total protein-encoding information of the *E. coli* genome.

T 16

AN INTEGRATED GENOMIC DATABASE(IGD)
G. Michaels, R. Overbeek, M. Price, M. Raju

We will present an integrated database that includes EcoSeq, the Swiss Protein Data Bank, Swiss Protein Data Bank, the Swiss Enzyme Data Bank, the EcoSeq Database, the ProSite Dictionary of Protein Sites and Patterns, the compilation of compounds distributed by Peter Karp, an encoding of the more common metabolic pathways, sequence data relating to Salmonella, and several genetic maps of bacterial genomes.

T 17

INFORMATICS ANALYSIS OF THE *E. COLI* GENOME
Kenneth E. Rudd. The National Center for Biotechnology Information,The National Library of Medicine,The National Institutes of Health,Bethesda,MD,20894,USA

The current status of the EcoMap, EcoSeq and EcoGene databases will be described, including GeneScape, a relational database of the *E. coli* genomic map, and BigSeq, a single DNA sequence representing the entire *E. coli* chromosome, including 40% determined DNA sequence and 60% Ns. Results of various studies will be presented including an information content analysis of ribosome binding sites, DNA sequence gap and contig distributions, both novel and established repeat element mapping, identification of putative, unidentified genes in published DNA sequences, and restriction site distributions. Applications of the methods to other genomes will be discussed. Proposals for coordinating the completion of the *E. coli* genomic DNA sequence will be presented.

T 18

DATABASE OF ALL CURRENTLY IDENTIFIED *E. COLI* GENE PRODUCTS

Monica Riley, Marine Biological Laboratory, Woods Hole, MA 02540

The gene products of over 1700 genes of *E. coli* have been characterized well enough to allow one to assess the present state of knowledge in relation to the ultimate goal of complete knowledge of all of the genes, the gene products and the regulatory circuits that are necessary to make a free-living cell. A database has been assembled of *E. coli* genes with characterized gene products, derived both from the list compiled by B. Bachmann (Microbiol. Rev. 54, 130, 1990) and from lists of more recently defined genes assembled and kindly shared by Kenn Rudd (National Library of Medicine, NIH, Bethesda). The gene products have been categorized by cellular function. They have been grouped into five major categories, Intermediary Metabolism, Biosynthesis of Small Molecules, Metabolism of Macromolecules, Cellular Processes and Other Functions. These groups are further subdivided into altogether about 60 subgroups. Although classification was necessarily arbitrary, perhaps even idiosyncratic, especially for multifunctional gene products, still the classification provides an overall picture of how many genes are doing what, and some sense of how far we have come and how far we have yet to go. The data will be displayed with an invitation to scientists to correct errors and enter new information.

T 19

TRANSPOSON FACILITATED SEQUENCING of CLONED DNAs

Douglas E. Berg[1]* and Claire M. Berg[2]. [1] Washington University Medical School, St. Louis, MO, 63110, and [2] University of Connecticut, Storrs, CT 06269

Transposable elements are valuable as sequencing tools because they can serve as mobile binding sites for "universal" PCR or sequencing primers, and their movement never obscures the linkage among different segments of the target DNA. We developed four transposon-based strategies that collectively should allow much more efficient DNA sequencing than is possible with shotgun subcloning or primer walking methods. **1.** Large DNA segments can be cloned in pJANUS, a unique γδ-based "deletion-factory" cosmid vector, and then analyzed using transposition to generate deletions that extend in each direction from a γδ end and give access to each DNA strand. These deletions are selected using genes for sensitivity to sucrose and streptomycin that flank γδ in pJANUS, and are recoverable because a replication origin was placed within γδ. **2.** Pre-existing plasmid clones are analyzed using a mini-derivative of γδ, whose insertion is selected by plasmid transfer in bacterial mating. **3.** DNAs cloned in λ vectors can be accessed using Tn*5supF*, whose insertion is selected by plaque formation on *dnaB*-amber *E. coli*, and direct and crossover PCR to map inserts and generate sequencing templates. **4.** λ clones can be converted to plasmid clones using a Tn*5supF* derivative containing a replication origin and resistance marker, cleavage at λ-target DNA junctions, and intramolecular ligation.

1. Wang, Blakesley, D. Berg & C. Berg, in prep. 2. C. Berg, Vartak, Wang, Xu, Liu, MacNeil, Gewain, Wiater & D. Berg, Gene 113:9-16, 1992. 3. Krishnan, Kersulyte, Brikun, C. Berg and D. Berg, NAR 22:6177-6182, 1991; 4. Krishnan, Kersulyte & D. Berg, in prep.

T 20

PLAQUE FORMING MU ELEMENTS FOR GENE FUSING AND CLONING J. Chen, X. She, and M. J. Casadaban*. Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, IL, 60637,USA.

The Mu transposon bacteriophage can be a very useful tool for studies of the genome of E. coli and related enteric bacteria. Here we describe two new types of Mu elements for gene fusing and DNA cloning. These elements have all the genes necessary for phage growth and plaque formation without a complementing helper phage. The gene-fusing elements have the *lacZ* or *aph* (Km/Neo) reporter genes in the non-essential region near the right end of the Mu genome. The full size of these elements allow them to transpose and form gene fusions ten times more frequently than defective Mu d elements and a hundred times more frequently than mini-Mu elements. The cloning elements have the pACYC184 replicon and the Cm^R gene with a DNA segment from the Mu left end which permits occasional recombination to form mini-Mu plasmid elements with host DNA insertions. The resulting clones have stable DNA structures since they are deleted for the Mu transposition genes. An ampicillin resistance gene was incorporated in all these elements as an independent selectable marker. Derivatives of these phage which can infect a larger number of bacteria of the family *Enterobacteriaceae* were made by incorporating a segment of a related tail gene from the P1 bacteriophage. This allows the Mu phage to bind to lower, more conserved segments of the *Enterobacteriaceae* surface lipopolysaccharide. The non-defective and broad-host range features of these phage facilitate their introduction into *Enterobacteriaceae* with different restriction systems and with weak or partially blocked receptors for Mu.

T 21

DNA REARRANGEMENTS AT SHORT HOMOLOGIES
J. H. Miller*, M. A. Schofield, R. Agbunag, L. Amii, and C. Cruz.
Department of Microbiology and Molecular Genetics, University of California, Los Angeles, CA, 90024, USA

We have constructed a system which allows us to monitor inversions at short inverted repeats. We have used this system to evaluate the role of recombination in generating rearrangements at short homologies. We report the effects of different repair deficient backgrounds on these rearrangements, and also describe mutants with elevated rates of inversions.

T 22

Genetic Approaches to Genome Stability in Salmonella

John Roth*, Department of Biology, University of Utah, SLC UT 84112

The bacterial chromosome structure (gene order) can be altered in major ways by homologous recombination between repeated sequence, by activities of transposible elements and by acquisition of foreign sequences from other organisms (horizontal transmission). Recombination between direct repeats on sister chromosomes can generate duplications and deletions. This rearrangement is common and appears to be mediated by both the recF and the recBC pathways of recombination. Inverse repeats in the same chromosome can in principle recombine to yield an inversion; this occurs only by the recBC pathway. Inverse sequences at some at some chromosomal sites interact recombinationally but do not generate inversions. We suspect that the failure of these sites to generate inversions is due to the lack of nearby sites that might allow the recBCD enzyme to access the recombining sequences, possible double strand break sites. The known recombination systems seem to favor duplications and disfavor inversions. The insertion sequence IS200 is the only known transposable element in the chromosome of genetic strains of Salmonella. It transposes infrequently but appears to catalyse cointegration but resolves cointegrates poorly; known inserts no terminal repeats and generates little or no duplication of target material. Insertional mutations are stable to reversion. The six chromosomal copies of IS200 provide direct order homologous sequences; homologous recombination between these sequences is a major source of duplications in Salmonella. The cobalamin synthetic genes represent almost 1% of the Salmonella genome. Most of these genes are in a single operon (26 genes). The 57% GC content of these genes and their absence from other enterics suggest that this pathway was inherited horizontal rather than by descent.

T 23

COMPUTER METHODS FOR IDENTIFYING INTERESTING SEQUENCES

Gary D. Stormo. Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309-0347

We have been exploring several methods for identifying interesting regions of biological sequences. If the problem is simply to examine a sequence and note unusual patterns, SEQUENCE LANDSCAPES is an effective method for identifying long repeats, rare sequences, patterns with unusual frequencies and other potentially interesting characteristics. If the problem is to find a sequence pattern associated with a particular function, methods that find maximally similar local alignments are needed. For example, the binding sites of the CRP protein can be discovered by running the CONSENSUS program on the upstream regions from a set of genes that are CRP-regulated. A related method based on the Expectation Maximization algorithm has been used to identify the consensus pattern of *E. coli* promoters, including their variable spacing. These methods can also be used to find protein motifs associated with particular functions, and can be used in conjunction with biochemistry experiments to rapidly determine the specificity of any DNA-binding protein.

IS NITROGEN LIMITATION SENSED VIA THE GLUTAMINE POOL?

T.P. Ikeda*, A.E. Shauger, S.G. Kustu. Department of Plant Pathology, University of California at Berkeley, 147 Hilgard Hall, Berkeley, CA 94720, USA.

We are determining what metabolites set the growth rate of *Salmonella typhimurium* when environmental nitrogen is limiting. Since glutamate and glutamine are the primary carriers of fixed nitrogen in the cell, we have monitored their pool sizes as potential signals of nitrogen limitation. To measure pools, we performed a rapid, "no-harvest" extraction in which cell suspensions were pipetted directly into methanol. After lyophilization of extracts, amino acids were derivatized with o-phthaldaldehyde, separated by reversed phase HPLC, and quantitated by fluoresence. Compared to cells grown on minimal media with ammonia (the optimal nitrogen source), cultures grown on the poorer nitrogen sources proline and arginine exhibited decreased growth rates and very low glutamine pools (< 5 % the pool of cells grown on ammonia) but had normal or nearly normal glutamate pools. A mutant capable of faster growth on arginine had a higher glutamine pool. To determine whether decreases in glutamine pool size could account for slow growth, we established a "calibration curve" for growth rate as a function of pool size by using a variety of *glnA* mutant strains in which glutamine synthetase is defective to different degrees. A *glnA* mutant strain that had the same growth rate on ammonia as wild-type did on proline (230 and 280 min, respectively) also had a comparably low glutamine pool. The same was true for a second *glnA* mutant and the arginine fast-grower. Hence it is our working hypothesis that *S. typhumurium* senses external nitrogen limitation as a decrease in the internal glutamine pool and that the low glutamine pool is responsible for slow growth.

THE ARC SIGNAL TRANSDUCTION SYSTEM FOR GLOBAL RESPIRATORY CONTROL OF GENE EXPRESSION IN *Escherichia coli*

E. C. C. Lin* and S. Iuchi. Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA 02115

In *E. coli* expression of a group of operons involved in aerobic metabolism is regulated by a two-component signal transduction system. The *arcB* gene (min 69.5) specifies the membrane sensor protein and *arcA* (min 0) the cytoplasmic regulator protein. ArcB is a large protein belonging to a subclass of sensors that have both a transmitter domain (on the N-terminal side) and a receiver domain (on the C-terminal side). This receiver domain has large segments homologous with the ArcA protein. Mutations in either *arcA* or *arcB* can anaerobically derepress several dehydrogenases of the flavoprotein class, the cytochrome *o* complex, and members of the citric acid cycle, the glyoxylate shunt, and the pathway for fatty acid degradation. Upon stimulation by a reducing compound in the metabolic network, ArcB undergoes autophosphorylation at a conserved His residue. The phosphoryl group on the His residue is then transferred to a conserved Asp in the receiver domain and ArcB becomes catalytically active as a kinase for ArcA. Thus the phosphorylation state of the receiver domain kinetically controls the intermolecular kinase activity of ArcB. The phosphorylated form of ArcA can act either as a pleiotropic repressor or activator. Like many globally controlled operons, the targets of the ArcA-P are scattered throughout the chromosome.

T 26

USING SEQUENCE "TAGS" TO IDENTIFY NEW PI-REGULATED GENES AND GENES FOR PI-INDEPENDENT CONTROL OF THE PHO REGULON

B. L. Wanner*, P. M. Steed, W. W. Metcalf, and M. R. Wilmes-Riesenberg. Department of Biological Sciences, Purdue University, West Lafayette, IN 49707.

Transposon mutagenesis was used to isolate mutants with *lacZ* transcriptional fusions to *P*i-starvation-*i*nducible (*psi*) genes, in the late 1970s. Many *psi* genes were identified over the years by classical techniques; many more were recently analyzed by sequencing the transposon-chromosomal junctions and identified by searching DNA databases with the junction sequences, which act as gene "tags." Results showed that the PHO regulon encodes thirty-one *psi* genes arranged in eight separate genes and operons, and that many other *psi* genes are also members of other regulatory systems. Transposon mutagenesis was also used to isolate mutants altered in Pi-independent control of the PHO regulon. As with the *psi* genes, many were identified by classical techniques; others were identified by searching databases with tags cloned from these mutants. In this way, we unexpectedly identified genes in central metabolism that affect PHO regulon control. Further studies showed that Pi-independent control of the PHO regulon is coupled to pathways in central metabolism in which Pi is a substrate. This form of "cross regulation" may be of fundamental importance in biology.

T 27

REGULATION AND FUNCTION OF *E. coli* HEAT SHOCK PROTEINS

J. Wild, D. Straus, A. Kamath-Loeb, P. Rossmeissl and C. A. Gross. Department of Bacteriology, University of Wisconsin, Madison, WI,53706, USA

A universal response of all organisms to temperature upshift and other stresses is an enhancement of the synthesis of a set of evolutionary conserved proteins called heat shock proteins (hsps). In *E. coli*, approximately 20 hsps are induced upon heat shock. Their synthesis is controlled by σ^{32}, an alternate σ factor of RNA polymerase. Induction of hsp synthesis occurs as a result of a rapid increase in the intracellular concentration of σ^{32}. Changes in σ^{32} stability and /or synthesis account for variations in σ^{32} level.

Most of the *E. coli* hsps function as proteases or molecular chaperones. The vital role of these chaperones is the maintenance of cellular proteins in a correctly folded state. This is achived either by preventing the misfolding, denaturation and aggregation of polypeptides or by promoting proper folding, oligomerization and assembly. A specific function of chaperones in protein export is maintaining the precursor form of secreted proteins in the export-competent state. We show that DnaK and DnaJ hsps participate in the export of alkaline phosphatase and play a critical role in strains lacking the secretion specific chaperone SecB by substituting for its function.

T 28

THE GALACTOSE REGULON OF *ESCHERICHIA COLI*

Michael J. Weickert* and Sankar Adhya, National Cancer Institute, National Institutes of Health, Bethesda, MD, 20892, USA.

The *E. coli* genes and operons encoding proteins for transport and metabolism of D-galactose, and regulation of these genes, are scattered around the chromosome, yet are coordinately regulated at the level of transcription by three proteins. Negative regulation of transcription is mediated by two isorepressors, GalR and GalS, which are members of the GalR-LacI family of regulators. GalS negatively regulates the *mgl* operon (high affinity galactose transport) and its own transcription, while GalR regulates the *gal* operon (galactose utilization). GalS also modulates *gal* operon expression in a manner epistatic to GalR, and GalR, in a reciprocal manner, affects *galS* expression but is epistatic to GalS. The cyclic AMP receptor protein (CRP) is a positive activator of transcription of *mgl*, *galS*, and P1 promoter of the *gal* operon.

The *gal*, *galS*, *mgl*, and *galP* (low affinity galactose transport) promoters contain almost identical DNA control elements including similarly placed CRP binding sites and upstream operators (O_E). With the exception of *mgl*, each gene or operon also has an internal operator (O_I). Although the *galR* promoter also contains a CRP binding site and two operators, but the position of these elements relative to the transcription start site is unlike that of the other promoters of the *gal* regulon. The *galR* promoter is insensitive to these three regulators. The effect of isorepressors on *galP* is currently being examined, although previous studies indicated that *galP* is negatively regulated by GalR and is relatively insensitive to catabolite repression.

T 29

CONTROL OF DNA SUPERCOILING

K. Drlica*, J. Rouviere-Yaniv, A. BenSaid, M. Sioud, and M. Malik. Public Health Research Institute, 455 First Avenue, New York, NY 10016 and Institut de Biologie Physico-chimique, 13, rue Pierre et Marie Curie, 75005, Paris, France

DNA supercoiling in bacterial cells is controlled by the concentrations and activities of two DNA topoisomerases, DNA gyrase and DNA topoisomerase I. Although these topoisomerases tend to keep supercoiling constant, environmental shifts such as changes in external salt concentration or oxygen tension do cause changes in transient and steady-state supercoiling. During these shifts both supercoiling and the ratio of [ATP] to [ADP] change in concert. Others have shown that *in vitro* gyrase either adds or removes negative supercoils according to the ratio of [ATP] to [ADP]. Thus supercoiling may be sensitive to cellular energetics. Supercoiling is also affected by the absence of the bacterial histone-like protein HU, and possible explanations will be discussed. Also to be discussed is the demonstration that hammerhead ribozymes are very effective in bacterial cells. These RNA molecules, which can be designed to cleave specifically any RNA target, may be very useful replacements for conditional lethal mutations in genetic studies.

T 30

HNS PROTEIN CONSTRAINS DNA SUPERCOILS IN VITRO AND MODULATES GLOBAL TRANSCRIPTION IN VIVO.

N. Patrick Higgins* and Victoria McGovern. Department of Biochemistry, University of Alabama at Birmingham, Birmingham, AL 35294

The *E. coli* chromosome is organized in a negatively supercoiled form with two discrete supercoil phases. Half of bacterial supercoils are in an interwound and energetically stressed conformation that is lost when the chromosome is nicked with gamma radiation. The remaining half of the supercoils are constrained, the restraining factors presumably being bound proteins. Finding which proteins constrain half of a bacteria's negative supercoils has been difficult. The two most abundant chromosome associated proteins are HU and HNS. HU has the ability to constrain negative supercoils and we find that in K-glutamate buffer HNS constrains supercoils as well.

While HU and HNS both constrain negative supercoils with equal efficiency, their mechanisms are antagonistic. HU disrupts HNS-derived supercoils and vice versa. In addition to this biochemical difference, these two proteins induce dramatically different cell physiology when over-expressed to various extents. This leads to a competitive model wherein HNS binding silences transcription while HU binding facilitates transcription in active chromosome domains.

T 31

BACTERIAL CHROMOSOME STRUCTURE AND TARGETING OF DNA-BINDING PROTEINS. D.E. Pettijohn*, V.L. Shellman, O. Pfenninger, and V. Vissa. Department of Biochem., Biophys.& Genetics, Univ. of Colorado Health Sciences Center, Denver, CO 80262

This research examines the diffusion and sorting of specific proteins, as they interact with DNA packaged in bacterial nucleoids. The central concern is elaborating the effects on protein targeting of the high DNA packaging density. One approach studies the movements of DNA-binding proteins (examples HU or HNS) and non-DNA-binding proteins (examples insulin or albumin) when a protein, labelled with a fluorescent chromophore, is introduced into permeabilized, viable, E. coli cells. Fluorescence microscopy/image processing allows an estimation of relative diffusion rates within the nucleoid. Complimentary studies use: a)photo-bleaching methods to study diffusion of FITC-labelled proteins in the presence of purified DNA b) gel retardation methods to study protein binding-dissociation at high DNA concentrations. Results suggest that the chromosome structure actually facilitates the targeting rates and the process is highly dependent on DNA strand exchange mechanisms.

T 32

SUPERCOILING STABILIZES Z-DNA IN *E. COLI*

Robert D. Wells, Institute of Biosciences and Technology,
Texas A&M University, Houston, Texas, 77030, U.S.A.

Non B-DNA structures are stabilized by negative supercoiling *in vitro* and *in vivo*. Conformations investigated to date include left handed Z-DNA, intramolecular triplexes, nodule DNA, and cruciforms.

The B to Z structural transition of varying lengths (74 to 14 bp) of (CG) tracts was used as a superhelicity probe to examine the local topological changes induced by transcription at defined genetic loci *in vivo* in *E. coli*. The local topology reporter sequences indicate that under steady state transcription, the region upstream from the promoter experiences an increase in negative supercoiling, whereas the region downstream from the terminator displays a decrease in negative supercoiling. Thus, this direct *in vivo* evidence demonstrates that the translocation of an RNA polymerase elongation complex along the double helical DNA generates positive supercoils in front of it and negative supercoils behind it.

Hence, the *E. coli* chromosome undergoes dynamic changes during biological processes such as transcription.

T 33

CONSERVATION AND DISTRIBUTION OF SHORT, INTERSPERSED REPETITIVE DNA SEQUENCES IN EUBACTERIAL GENOMES

J. Versalovic[1], T. Koeuth[1], and J.R. Lupski*[1,2], [1]Institute for Molecular Genetics and [2] Department of Pediatrics, Baylor College of Medicine, Houston, TX, 77030, USA

Dispersed repetitive DNA sequences have been described in eubacteria. To assess the distribution and evolutionary conservation of prokaryotic repetitive elements, consensus oligonucleotides matching repeated DNA elements were used in slot blot hybridization and PCR amplification (rep-PCR) experiments with genomic DNA from diverse eubacterial species. Both REP and ERIC probes hybridized preferentially to genomic DNA from Gram-negative enteric bacteria and related species. These probes also annealed to multiple clones on filters containing ordered sets of lambda phages and cosmids representing the entire genome of *E. coli* K-12 strain W3110. Four of seven *E. coli* loci known to contain ERIC sequences were identified by hybridization to filters containing the Kohara lambda phage miniset library. At least seven new loci containing ERIC elements were identified by hybridization. These hybridization results demonstrate that REP and ERIC sequences are dispersed throughout the *E. coli* genome. Oligonucleotides matching Repetitive Extragenic Palindromic (REP) and Enterobacterial Repetitive Intergenic Consensus (ERIC sequences) were synthesized and tested as opposing PCR primers in the amplification of eubacterial genomic DNA. REP and ERIC consensus oligonucleotide primers produced clearly resolvable bands by agarose gel electrophoresis following PCR amplification. These band patterns provided unambiguous DNA fingerprints of different eubacterial species and strains. Rapid methods using whole cells directly from colonies or cultures will enable the rapid distinction of strains for genomic analysis. Differences in REP- and ERIC-PCR-based fingerprint patterns demonstrate that REP- and ERIC-like sequences are conserved and distributed differently in bacterial genomes. REP-PCR DNA fingerprints distinguished closely related *E. coli* K-12 laboratory strains, HB101 and W3110, by the presence of a ~400 bp band present in W3110. This prominent band may reflect a different arrangement of REP sequences in the genomes of W3110 and HB101.

Abstracts of Posters

DNA SEQUENCE DIVERGENCE OF IS*711*: A THEORY ON THE PROGRESSION OF TRANSPOSITION EVENTS

Besty J. Bricker and Shirley M. Halling
National Animal Disease Center, ARS, USDA, Ames, IA 50010 USA

IS*711* is an insertion sequence recently discovered in *Brucella ovis*, a pathogen of the ovine reproductive system. *Brucella* is a genus of gram negative bacteria containing six closely related species. The insertion sequence is present in more than one copy in *B. ovis*, is 842 base pairs long, and has a 20 base pair imperfect inverted repeat. The element was not observed in a survey of other gram negative bacteria including *E. coli*, *Rhizobium* and *Agrobacterium*. Because the element has been found in *Brucella* world-wide, the introduction of IS*711* into the *Brucella* genome is presumed to be an ancient event. IS*711* hybridized to genomic DNAs from all *Brucella* species. Copy numbers ranged from six copies per genome to greater than twenty copies per genome. The hybridization pattern of six of the IS copies was essentially the same in all the *Brucella* species. Transposition to these six common sites is either an ancient event preceeding the evolution of species or a more recent event involving transposition to "hot spots" for insertion. DNA sequence determination of IS*711* copies from several of the common loci supports the second hypothesis. DNA sequence polymorphisms were species-specific rather than locus-specific. The six common sites of insertion appear to have arisen by transposition to preferred target sites.

USING DNA MARKOV CHAIN MODELS FOR PREDICTING *E.COLI* GENES

M. Borodovsky*, J. McIninch, School of Biology,
Georgia Institute of Technology, Atlanta, GA 30332-0230, USA

The automation of the present large scale sequencing of the *E.Coli* genome makes acute the problem of fast and accurate prediction of protein coding regions in newly sequenced DNA. The new Bayes' type gene predicting method is based on homogeneous and non-homogeneous Markov models of non-coding and coding regions of DNA sequence. The values of the model parameters were obtained from the training sets of coding and non-coding *E.coli* DNA according to the maximum likelihood principle. It has been shown that predictive accuracy of the algorithm steadily improves when Markov chain models of increasing orders (from first to fifth) are utilized.

A number of existing gene predicting algorithms generate false signals when they meet the "shadow" of the true coding region which is actually located on the complementary strand. Thus, one needs to spend more time and efforts for filtering preliminary predictions. A special version of the method using additional non-homogeneous model of the coding region shadow was designed. This version performs simultaneous analysis of sense and anti-sense strands of DNA and does not produce false signals. The functional meanings of isolated DNA fragments of 96 bp length is predicted with a 14.2% false negative rate (coding as non-coding) and a 21.7% false positive rate (non-coding as coding). It has been also shown that the DNA sequence ambiguities (up to 5%) do not significantly affect the predictive accuracy.

The first version of the program (GenMark) was used for the analysis of the 91,400 bp *E.Coli* DNA sequenced at the University of Wisconsin. Now the second program version can be used through the E-mail server at the Georgia Institute of Technology Computer Center. The program generates the list of predicted genes. The graphical output indicating coding regions in six frames is available in the form of a PostScript-file as well. An IBM PC version of the program has been developed.

SHOTGUN SEQUENCING WITH THE JANUS VECTOR.

V. Burland*, D. L. Daniels, G. Plunkett III and F. R. Blattner.
Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA.

Janus is an M13-derived cloning vector that enables single-stranded template DNA of either strand to be prepared from phage culture supernatants by standard protocols. To make the opposite strand available for packaging into the phage, an inducible site-specific recombination system inverts the cloned insert during the growth of the phage culture. This is particularly useful in large projects since only one DNA preparation system must be set up on a large scale; no subcloning or double-strand sequencing are necessary.

After generating a library of clones in the Janus vector, the sequencing strategy is as follows: first, sequences are collected from random clones until most of them will align into contigs; then individual clones are identified for inversion. These are chosen so that the new data from the opposite end of the insert (and from the second strand), will link contigs together and add depth to specific areas of poor data. The system has the advantages of both directed and random strategies with only the cloning work required for random sequencing.

EXPRESSION MAPPING THE *ESCHERICHIA COLI* CHROMOSOME

R.L.Clark*, J.Bogan, and F.C.Neidhardt. Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, 48109, USA

The mapping of protein spots on 2-D gels to their corresponding genes by matching gel based and sequence based molecular weights and isoelectric points is not generally possible due to the deviation between gel based and sequence based values and the high density of spots on standard gels. The spot density can be reduced and all of the proteins of *E. coli* can be mapped by uniquely expressing small chromosome fragments (the Kohara miniset inserts) from a T7 promoter plasmid in a cell background containing an integrated T7 RNA polymerase if rifampicin is added to block synthesis from *E. coli* promoters. Ten miniset inserts were cloned and expressed as indicated. A total of 70 spots were mapped to the chromosome on the basis of their expression from a particular Kohara insert. Seventeen of these spots were known sequenced proteins not previously identified on 2-D gels and 12 were previously identified proteins. The remaining 41 spots were from inserts containing unsequenced regions and cannot be unambiquously identified on 2-D gels until the sequence is available.

THE NUCLEOTIDE DISTRIBUTION AND FRACTAL DIMENSION OF GENOMIC SEQUENCES

D.J.Cork* and B.G.Nguyen. Departments of Biology and Mathematics, Illinois Institute of Technology, Chicago,Il, 60616, USA.

A three dimensional algorithm depicts graphically the nucleotide distributions within genomic sequences. A fractal dimension of the genomic sequence can be calculated from this information. Starting at the origin of an x-y-z coordinate grid, each nucleotide is plotted one unit away from the prior nucleotide in the direction assigned to the nucleotide according to the following rule: +i = +(1,0) for a = adenine, -i = -(1,0) for t = thymine, +j = +(0,1) for g = guanine, and -j = -(0,1) for c = cytosine. The z axis shows the nucleotide position number (N) starting from the 5'end of the genomic sequence at coordinate(0,0,0). Selected genomic sequences have been examined from viruses,bacteriophages and bacteria. Those possessing a nearly linear distribution of nucleotides have fractal dimensions closer to one. The Silicon Graphics Iris Workstation has been used by the Graphics Center of IIT (Dr. Roberge, Dr. Grace, and Mr.Wu) to graphically display the structures. It is suggested that genomic sequences can be classified according to their fractal dimension.

A HIGH-THROUGHPUT ROBOTIC SYSTEM FOR SEQUENCING REACTIONS.

Donna L. Daniels* and Frederick R. Blattner.
Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA.

As part of the *E. coli* genome sequencing project at the University of Wisconsin, we have developed an automated system to perform the numerous pipetting and incubation steps involved in dideoxy sequencing. The advantages of the automated method over manual methods include savings of labor, elimination of pipetting errors, and the achievement of high throughput rates. We have modified the Gilson model 212B liquid handler to include heating and cooling of samples in microtiter dishes and dispensing of microliter volumes. The robot was programmed to perform all steps of the sequencing protocol, starting with 24 DNA templates and reagents and yielding 96 reactions ready to load on a gel. The procedure takes 45 minutes. Currently a single human operator runs three robots simultaneously — dispensing reagents and DNA templates to the robots; initiating the robot runs; and sealing, labeling and storing the resulting reactions.

COMPARISON OF *murB* SEQUENCES FROM *S. TYPHIMURIUM* AND *E. COLI*

P. M. Dombrosky* and K. D. Young. Department of Microbiology and Immunology, University of North Dakota School of Medicine, Grand Forks, ND, 58202, USA

A library of temperature sensitive mutants of *S. typhimurium* have been isolated by M. Schmid (Princeton), among which are mutants that lyse after a shift from 30° to 44°C. A genomic library of *S. typhimurium* in pBR328 was used to screen several of these lysis^{ts} mutants for complementing genes. One plasmid, pMBS187, rescued three strains (SE5318, SE5403, and SE5045) from lysis at high temperature. Randomly primed probes from this plasmid hybridized to λ534 of the Kohara *E. coli* genomic library, corresponding to ~89 minutes on the *E. coli* chromosome. A 3.7 kb DNA subclone from pMBS187 contained the complementing gene. Sequences from this DNA fragment were highly similar to the *E. coli* gene, *murB*, that encodes the enzyme UDP-*N*-acetylenolpyruvoylglucosamine reductase, which mediates one of the cytoplasmic steps in the synthesis of peptidoglycan precursors. Though a high degree of similarity exists between the *S. typhimurium* and *E. coli* genes, the temperature sensitivity could not be complemented by the cloned *E. coli* *murB* gene (supplied by M. Pucci and T. Dougherty, Bristol-Myers Squibb).

RAPID DNA SEQUENCING WITH ULTRATHIN HORIZONTAL ELECTROPHORESIS GELS,

R. W. Dunst*, B. P. Walsh, and M. F. Roethle. Biotechnology Center, FOTODYNE Incorporated, New Berlin, WI,53151,USA.

The electrophoretic separation of DNA sequence mixtures has been facilitated by the use of horizontal ultrathin gels. These gels dissipate heat more efficiently than traditional vertical sequencing gels and thus may be run at substantially higher power settings for greater speed. We describe experiments performed to optimize parameters for manual DNA sequencing using the GeneSprinter™ Sequencing System. This system employs active thermal regulation through circulating water to a horizontal sequencing cell and is driven by a 10kV power supply. Voltage/current/power profiles were adjusted to yield high resolution sequencing results in short electrophoresis run times. The effects of gel thickness, gel matrix and concentration, andbuffer composition were investigated for their effects on system performance. Over 250 bases of readable sequence data were obtained from single load electrophoretic runs of less than 20 minutes. The results of multiple loading runs will also be presented.

AUTOMATED FRAGMENT ASSEMBLY IN THE WISCONSIN SEQUENCE ANALYSIS PACKAGE

I. Edelman, B. Butler, and J. Devereux. Genetics Computer Group, Inc.
Madison, Wisconsin, 53711, USA

GelMerge is the newest program in the GCG Fragment Assembly System, designed to aid researchers in assembling and managing the fragments of a sequencing project. **GelMerge** rapidly and automatically assembles sequence fragments into *contigs*. **GelMerge** first finds the two fragments in the project with the longest overlap and then aligns them. This process of overlap determination and alignment continues until all of the overlaps among fragments in the project have been found and aligned. New sequences can be entered into the Fragment Assembly Project at any time. When **GelMerge** is run again, overlaps between these new fragment sequences and existing contigs will be found and aligned into increasingly larger contigs.

When two overlapping contigs are assembled into a single, larger contig, all of the fragment sequence information in both of the original contigs is used to create the final alignment. The alignment is *not* created simply from the consensus sequence of the two original contigs. **GelMerge** can recognize and remove vector sequences from individual fragments prior to finding overlaps. As your project progresses, you may, at any time, choose to start over and reassemble the entire project from either original or edited fragment sequences.

Aligning 300 fragments with an average length of 400 bases (119,994 bases, total) into a single 21,319 base contig takes 5.75 minutes of CPU time on a DEC VAXStation 4000 (Model 60).

BALANCE EQUATIONS FOR DIFFERENCES OF INVERTED DOUBLETS IN E.COLI GENOME AND PHYLOGENOMIC COMPARISONS

A. A. Filyukov*, D. J. Cork, Computer Science* and Biology Departments, Illinois Institute of Technology, Chicago, Illinois, 60616, USA

Doublet frequencies in evolutionary distinct groups (published by Nussinov) obey a balance type equations that may be used for the aims of phylogenomic comparisons of known regions of *E.Coli* genome with other groups of genomes. According to these equations, the classification of different types of genomes can be done by establishing the three denominations of the preferenced dinucleotide circulation balances. The specified genomes can be described by assignment to each type of dinucleotide circulation balance a special rate value that can be calculated from the dinucleotide frequencies data. The classification data of *E.Coli* Genome are compared with types and rates of classification for other bacteria, bacteriophages, RNA viruses, and nonvertebrate organisms. From the classification data obtained, it is suggested that bacteriophages have some similar structure of dinucleotide patterns with the *E.Coli* genome.

A TIMING COMPARISON OF PREPASS ALGORITHMS FOR FRAGMENT ASSEMBLY W.L. Istvanick*, J. Meidanis, G. Rabin, F. Blattner, D. Joseph.
Departments of Computer Science and Genetics, University of Wisconsin, Madison, WI, 53706, USA

The goal of the Prepass stage of fragment assembly is to estimate the amount of overlap between pairs of fragments. One method for computing such an estimate is to calculate the number of common sub-fragments of a fixed length. A window of fixed length is moved along each fragment one base at a time, and the sub-fragment present in each window is stored in a data structure. The resulting data structures are compared to yield the number of common windows between two fragments.

Methods based on four different data structures have been implemented for storing this information: Sorted Dictionaries, Hash Tables with Separate Chains, Hash Tables with back prediction, and Suffix Trees. These different implementations were compared in a timing study on four separate computational platforms (DEC Workstation, SUN Workstation, MacIntosh, IBM Compatible Personal Computer).

EcoCyc: Electronic Encyclopedia of *E. coli* Genes and Metabolism
P. D. Karp*, Artificial Intelligence Center, SRI International, 333 Ravenswood Ave. EJ229, Menlo Park, CA 94025, and M. Riley, Marine Biological Laboratory, Woods Hole, MA 02543.

The EcoCyc project is compiling a database (DB) describing the genes and intermediary metabolism of *E. coli*. The DB will describe each reaction of central *E. coli* metabolism, and the enzyme that carries out each reaction, including the cofactors, activators, inhibitors, and subunit structure of the enzyme. Initially the project will not address secondary or macromolecule metabolism. When known, the genes encoding the subunits of an enzyme will be listed, as well as the map position of a gene on the *E. coli* chromosome. The DB will describe every chemical compound involved in each reaction, listing synonyms for the compound name, the molecular weight of the compound, and in many cases its chemical structure. We estimate that the DB will describe roughly 500 enzymes and reactions, and 1000 chemical compounds. The DB will also contain the location of restriction sites in the *E. coli* genome for 8 restriction enzymes, and the locations of Kohara miniset clones. Genomic data in EcoCyc will be obtained from K. Rudd of the NIH.

EcoCyc will have two classes of uses. First is as an electronic reference for biologists. Second is for complex computations. To support the use of EcoCyc as an reference, we will implement graphical interfaces to the EcoCyc data that allow biologists to view reaction pathways, information about individual reactions, enzymes, and compounds, and the genetic map. One computational use of the metabolism DB relates to the design of variants of biochemical pathways. The EcoCyc data will also be useful to scientists who perform quantitative and qualitative simulations of metabolic pathways.

VARIATION AMONG ENTERIC BACTERIA OF GENES OF THE RESTRICTION CLUSTER (ICR): MRR-HSDRMS-MCRBC.

J.E. Kelleher and E.A.Raleigh. New England Biolabs, 32 Tozer Rd., Beverly, MA 01915 USA

E. coli K-12 expresses four restriction functions that regulate entry of foreign DNA into the cell. Three of these, Mrr, *Eco*K and McrBC, are encoded by six genes--*mrr, hsdRMS, mcrBC*--in a tight cluster of 14 kb, which we call the Immigration Control Region (ICR). Previous evidence obtained by sequence analysis of *hsd* genes from different enteric bacteria (Murray laboratory, Edinburgh) had suggested that different parts of this locus might have evolved in different ways, in a mosaic structure with conserved and variable segments. Sequence analysis in our laboratory had also presented evidence suggesting that part of the region (*mcrBC*) might have been imported from an environment of low G+C DNA base composition. As part of our interest in mechanisms that limit genetic exchange, we have begun an analysis of variation at this locus in 19 species of Enterobacteriaceae, using Southern blotting. Preliminary findings confirm that the observed sequence variation at this locus cannot easily be explained on the basis of conventional evolutionary trees.

THE IDENTIFICATION OF STRONG GYRASE SITES USING F PLASMID PROTEIN CcdB

Katherine E. Kézdy*, Philippe Bernard, Martine Couturier, and N. Patrick Higgins. Department of Biochemistry, University of Alabama, Birmingham, AL 35294 and Laboratoire de Génétique, Département de Biologie Moléculaire, Université Libre de Bruxelles, rue des Chevaux, 67, B-1640 Rhode Saint Genèse, Belgium.

DNA gyrase is important in DNA replication, repair, and recombination, but important sites at which gyrase binds and acts have not been easy to study. An unusual DNA gyrase binding site in bacteriophage Mu is essential for efficient viral replication. This strong gyrase site(SGS) is located at the center of the Mu genome between the late genes G and I. The quinolone antibiotic Enoxacin traps gyrase in a cleavable complex at the Mu SGS both *in vivo* and *in vitro*. We recently found a cellular protein (CcdB) that traps gyrase in a cleavable complex at the SGS. CcdA and CcdB proteins are encoded by the F plasmid and these factors are important in F segregation. Cells that express CcdB but not CcdA protein become filamentous and produce aberrant nucleoids. This CcdB-mediated perturbation of DNA structure and formation of anucleate cells suggests that CcdB interferes with chromosome replication and/or segregation. Recent genetic studies and ability of CcdB to trap gyrase in a cleavable complex identifies gyrase as the target of CcdB protein killing in vivo. Using CcdB as a gyrase poison, a strong gyrase site has been identified in the plasmid pSC101. This highly specific gyrase inhibitor is a new tool to identify important gyrase sites in the *E. coli* chromosome.

DYNAMIC METHODS FOR FRAGMENT ASSEMBLY IN LARGE-SCALE GENOME SEQUENCING PROJECTS

A. Kryder*, G. Lewandowski, J. Meidanis, A. Rang, S. Wyman, D. Joseph. Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, 53706, USA.

Large-scale genome sequencing projects present a unique set of problems not found in smaller sequencing efforts. Key to the success of large sequencing efforts are the mathematical and computational tools for organizing and analyzing large quantities of genetic sequence data. We have developed a package designed specifically for handling fragment assembly problems that arise in large-scale sequencing projects. Our system is unique in that it maintains the layout information for the fragment assembly using a set of dynamic data structures and algorithms. This permits new data to be very quickly added and analysis to be carried out at any point in the assembly process.

HORIZONTAL TRANSFER AND THE EVOLUTION OF COBALAMIN SYNTHESIS AMONG ENTERIC BACTERIA

J.G. Lawrence* and J.R. Roth. Department of Biology, University of Utah. Salt Lake City, Utah 84112.

The vitamin B_{12} (cobalamin) synthetic genes (*cob*) clustered at 41' comprise nearly 1% of the *Salmonella typhimurium* chromosome. Although cobalamin is produced *de novo* by *Salmonella* under anaerobic conditions, it participates in only four, nonessential enzymatic reactions, including degradation of propanediol. Among enteric bacteria surveyed, a majority could synthesize cobalamin only if provided with cobinamide, the corrin precursor. Aside from *Salmonella*, only species of *Klebsiella* and *Citrobacter* could synthesize cobinamide and cobalamin *de novo*. However, homologues of the *Salmonella* cluster of synthetic genes were not present in any other enteric bacterium, including those with *de novo* synthetic capacity. Therefore, the cobalamin synthetic gene cluster has been introduced into the *Salmonella* chromosome from an exogenous source. The adjacent, cobalamin dependent, propanediol utilization operon (*pdu*) is coregulated with the *cob* operon and may also be of exogenous origin. In support, DNA sequence analysis reveals incongruities between the *pdu/cob* operons and typical *Salmonella* genes.

THE *fuc* AND *rha* LOCI OF *Escherichia coli* SPECIFY ANALOGOUS METABOLIC FUNCTIONS BUT THE ENCODED PRODUCTS SHOW NO APPARENT SEQUENCE HOMOLOGY

Z. Lu* and E. C. C. Lin. Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston MA 02115

The *fuc* and *rha* gene clusters, located at minutes 60.2 and 87.7 respectively on the chromosome, specify the inducible metabolism of L-fucose and L-rhamnose, which are structural analogs. Each of the clusters encodes a permease, an isomerase, a kinase, and an aldolase involved sequentially in the dissimilation of the respective pentose. The two metabolic pathways merge after the aldol cleavage step, forming common products. One product is further acted upon anaerobically by an oxidoreductase, encoded by *fucO*. Both sets of the structural genes are positively regulated. L-Rhamnose cross induces *fuc* expression, thereby allowing the recruitment of the oxidoreductase. Prevalent molecular evolutionary theory would predict that the *fuc* and *rha* genes evolved from common ancestors through duplications. Amino acid sequence analysis, however, provided no evident homology between these two sets of proteins.

AN IN VIVO SYSTEM FOR STUDYING Z-DNA FORMATION IN E. COLI CHROMOSOME

S. Lukomski,*and R. D. Wells. Institute of Biosciences and Technology, Texas A&M University, 2121 Holcombe Blvd., Houston, TX 77030

Left handed Z-DNA in (pur-pyr) sequences was detected by *in vivo* and *in situ* assays in recombinant plasmids. Herein, we describe a system which enables the extension of the structural studies into the bacterial chromosome.

A DNA methylase assay was developed based on the concept that appropriate sequences that adopt Z helices *in vivo* are not methylated in the left handed conformation but are methylated in the right handed B structure. A cloned MHaeIII (GGCC) was employed to study the methylation of chromosomal NotI (GCGGCCGC) sites. We showed that all naturally occurring NotI sites (20-23 per chromosome) were effectively methylated. For comparison, a NotI site flanked by (pur-pyr) stretches of different lengths was cloned within the Tn5 transposon and integrated into the *E. coli* chromosome with the suicide vector pRT733. Genomic DNAs from several Tn5 mutants were prepared in microbeads and the insertions were mapped in different locations on the chromosome. The introduced $(GC)_n$ tracts were stable and were inserted as single copies.

TRANSCRIPTION AS A DETERMINANT OF LOCAL DNA TOPOLOGY IN *E.COLI.*

A.S. Lynch* and J.C.Wang. Department of Biochemistry & Molecular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138.

In an effort to further elucidate physiological roles of topoisomerase I, and the mechanisms by which compensatory mutations facilitate growth in its absence, we have undertaken a series of studies to investigate cellular processes which may in vivo result in the formation of its preferred in vitro substrate, namely highly negatively supercoiled DNA. The expression of plasmid-borne genes encoding either integral cytoplasmic membrane proteins or exported proteins is observed to give rise to highly negatively supercoiled forms in cells lacking topoisomerase I. The results are consistent with predictions of the 'twin-domain' theory of transcriptional supercoiling, wherein effective cellular anchoring of RNA polymerase is mediated by co-transcriptional interactions of nascent polypeptides with factors involved in the integration into, or translocation across, the cytoplasmic membrane of the cell. Further studies suggest that the accumulation of high levels of negative supercoils into circular templates requires the presence of topological barriers which prevent the passive diffusion of supercoils and therein the twinning of supercoiled domains of opposite sign. We are currently characterizing sequences derived from the *E.coli* genome which appear to mediate either transcriptional anchoring events or the formation of barriers to supercoil diffusion.

POSTTRANSCRIPTIONAL REGULATION OF KATF SYNTHESIS AND ITS CLIENT GENE PRODUCTS

M.McCann, J.Kidwell, and A.Matin*. Department of Microbiology and Immunology, Stanford University, Stanford, CA 94305, USA

When *Escherichia coli* is starved for carbon, some 50 starvation proteins are induced, 32 of which require the putative σ factor KatF. The latter proteins confer generalized stress resistance on *E. coli*. Operon and protein reporter fusions were constructed to the *katF* gene containing different sizes of its coding region. Glucose starvation induced the operon fusions by about twofold, but the extent of induction of the protein fusions depended on the size of the coding region: that to base pair 786 exhibited eightfold induction, but to one with a shorter fragment only a twofold induction. The major promoter for *katF* transcription resembles the σ^F recognized consensus sequence. KatF mRNA secondary structure suggests posttranscriptional regulation by interaction between its translation initiation and 3' regions, although other mechanisms -- KatF autoregulation and stability -- also play a role. The induction of KatF-dependent genes *katE* and *pexB* was regulated at the transcriptional level and was coincident with KatF synthesis under glucose starvation. However, *pexB* regulation under nitrogen starvation was entirely at the posttranscriptional level. This multilevel regulation of KatF and its client genes probably facilitates the differential induction of KatF-controlled genes by different stresses.

HASHING AND BACK PREDICTION METHODS FOR COMPUTING SEQUENCE SIMILARITY FOR FRAGMENT ASSEMBLY

J. Meidanis, D. Joseph*. Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, 53706, USA.

In large-scale sequencing projects, the computational task of assembling sequenced fragments usually begins with a pairwise comparison phase whose goal is to detect overlapping fragments. We present a new overlap detection method and show that it is much faster than other methods, yet accurate enough to assemble fragments in one contig in most cases. The technique works best with genomes in which repeated elements are rare (e.g., *E. coli*).

AMBIENT-TEMPERATURE-STABLE REAGENTS FOR MANUAL AND AUTOMATED DNA SEQUENCING

Edmund Ogutu*, David Walker, Ponnusamy Ramanujam, Harry Osterman, Chris Lively, James Koelbl, James Jolly, Scott Duthie, Brent Burdick, Peter Bell. R&D Department, Pharmacia P-L Biochemicals, Inc., Milwaukee, WI 53202, USA

Enzymatic DNA sequence determination relies on labile DNA polymerases and nucleotides that require sub-ambient-temperature storage. In addition, dispensing enzymes and nucleotides from aqueous solution is time-consuming and may lead to pipetting errors. We have developed a novel method for the stabilization of reagents for enzymatic DNA sequencing which permits ambient-temperature storage and reduces the number of pipetting steps by 25%. Carbohydrate polymers were utilized to form glassy matrices of labeling mix (dNTPs and buffered DNA polymerase) and termination mixes (dNTP/ddNTP; G, A, T, C). These mixes were found to be stable for extended periods of time at room temperature. Both T7 DNA polymerase and the Klenow fragment of *E. coli* DNA Polymerase retained full functional activity after six months of storage at 22°C. FPLC® analyses of stabilized nucleotides showed no significant breakdown of dNTPs after several months of storage. We have utilized these ready-to-use reagents in sequencing both double and single-stranded DNA templates. Results of manual and Automated Laser Fluorescent (A.L.F.™) DNA sequencing show data that are equivalent to those obtained using conventional sequencing reagents.

LACK OF CONSERVATION OF SOME *TER* REGION SEQUENCES

I.R.Oliver and Millicent Masters. Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh, EH9 3JR, Scotland

The terminus region of *E. coli* although deficient in essential or even identified genes does contain identifiable coding units (Moir et al J. Bacteriol. 174:2102 (1992)). A possible explanation of these facts is that terminus DNA consists of sequences that vary considerably between strains and species which, otherwise, are closely related. To test this idea we compared, by hybridization, the conservation of DNA from the *trg* region with that of DNAs from conserved parts of the chromosome. Hybridization was to DNAs of several Enterobacterial species and to DNA from 10 selected strains of the ECOR collection. We found that terminus region DNA was indeed less conserved than control DNAs and contained sequences of several kb in length that were entirely missing both from several Enterbacterial species and also from some of the ECOR strains.

NUCLEOTIDE MISINCORPORATION MODEL FOR RT-REACTION

Dao Nguyen*, Department of Mathematics
Douglas Cork, Department of Biology
Illinois Institute of Technology, Chicago, IL 60616

Abstract: A Markov process is introduced to study the *in vitro* reverse transcriptase reaction caused by an RT enzyme. This model is based upon considerations of probabilities of base misincorporation and interactions of neighboring nucleotides in environments that are strongly governed by patterns of the template-primer structure.

ANALYSIS OF THE *E. COLI* GENOME: INSIGHTS INTO OTHER ORGANISMS.

Guy Plunkett III*, Donna L. Daniels, Valerie Burland, and Frederick R. Blattner.
Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA.

In the course of analyzing the sequence of *E. coli*, it is not surprising to find genes whose products are conserved over large evolutionary distances. It is these very sequence similarities that aid the identification of newly sequenced *E. coli* genes. However, this flow of information is shown to be a two-way process, with data from the *E. coli* genome project shedding light on the functions of uncharacterized sequences from other organisms. Previously sequenced (but unidentified) genes or gene fragments from other organisms can be identified by similarity to products of *E. coli* genes. New information regarding the functions of previously described peptides can also arise from such similarities. Specific examples of such "spinoffs" from the region currently being sequenced will be described.

IN VIVO AMPLIFICATION OF SPECIFIC SEQUENCES OF ESCHERICHIA COLI

G. Pósfai, M. Koob, N. Hasan and W. Szybalski
McArdle Laboratory, University of Wisconsin, Madison, WI 53706, USA

Genome sequencing usually requires cloning of random, large pieces of a chromosome in a heterologous host. In contrast, we intend to amplify ordered, large segments of the chromosome in the organism itself, eluding difficulties concerning the size and stability of the cloned segment and eliminating the need for sorting the cloned pieces. First, we need a set of strains marked with a single mobile genetic element, each at a different location. Next, a replication origin and a sequence which permits excision are inserted in the marked site by homologous recombination. Third, a pair of such strains is crossed to obtain a doubly marked strain. Finally, providing an inducing signal, the segment between the marked sites is excised and replicated in the cell.

The feasibility of such a scheme is shown first in E. coli, as a model system. A complete set of strains, each marked with a single Tn10 transposon at a different location is available (Singer et al., 1989, Microbiol. Reviews 53 1-24). We have constructed the plasmids which ensure efficient insertion of the yeast FRT sequences (Huang et al., 1982, J. Mol. Biol. 161 33-43) and a π replication origin (Stalker et al., 1990, Nucl. Acids Res. 19 443-448) into the E. coli chromosome at the Tn10 sites. Doubly marked strains are obtained by P1 transduction. Providing the Flp and π proteins in trans, the segment between the two marked sites will be excised and replicated. This scheme may provide a tool to study the effects of amplification or deletion of large segments of the E. coli chromosome as well. Currently we are testing the efficiency of the system.

MENAQUINONE (VITAMIN K_2) BIOSYNTHESIS: 2-KETOGLUTARATE DECARBOXYLASE AND SHCHC SYNTHASE ARE ENCODED BY THE SAME GENE IN ESCHERICHIA COLI

V. Sharma, C. Palaniappan, M. E. S. Hudspeth, and R. Meganathan*. Department of Biological Sciences, Northern Illinois University, DeKalb, IL, 60115

In menaquinone biosynthesis the formation of 2- succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylic acid (SHCHC) from isochorismate and 2-ketoglutarate requires both SHCHC synthase and 2-ketoglutarate decarboxylase (KDC). The DNA sequence of menD, part of the menDBCE cluster located at 48.5 min, has been shown to encode SHCHC synthase. KDC is a TPP-requiring enzyme and is distinct from the 2-KGDH complex. By genetic and enzymatic complementation we have identified a region of the chromosome which confers both enzymatic activities. DNA sequence analysis of this region identified a 0.48Kb ORF overlapping menD, but which lacks an RBS and is incapable of elevating KDC activity. However, a reexamination of the reported upstream menD sequence identified an additional nucleotide in the SHCHC synthase coding region. This corrected menD sequence incorporated the KDC ORF into an extended reading frame. Deletion analysis of the new reading frame confirmed that menD encodes a bifunctional enzyme.

AUTOMATED SOLID PHASE DNA SEQUENCING

Mathias Uhlen and Thomas Hultman, Department of Biochemistry and Biotechnology, Royal Institute of Technology, S-10044 Stockholm, Sweden.

A solid phase approach have been developed for automated sequencing of plasmid and genomic DNA obtained directly by the polymerase chain reaction. Using the biotin-streptavidin interaction, two different robotic work stations (Beckman and ABI) have been designed to perform the various steps in the protocol (magnetic separation, primer annealing, sequencing reactions etc). An integrated scheme for large scale sequencing has been developed involving the analysis on automated laser fluorescent electrophoresis units. The use of the system for sequencing parts of the *E. coli* genome will be discussed.

GLOBAL REGULATION OF GENE EXPRESSION IN *E. COLI*

S. E. Chuang*, D. L. Daniels, and F. R. Blattner. Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA

Global transcription responses of *E. coli* to various stimuli or genetic defects were studied by measuring mRNA levels in about 400 segments of the genome. This was done by hybridization to DNA dot blots made with overlapping lambda clones spanning the genome of *E. coli* K12. Conditions examined included IPTG induction, heat shock, osmotic shock, starvation for various nutrients, cells going into stationary phase, anaerobic growth in a tube, growth in the gnotobiotic mouse gut, and effects of pleiotropic mutatons *rpoH*, *himA*, *topA*, and *crp*. Most mapped genes known to be regulated by a particular situation were successfully detected. In addition, many chromosomal regions containing no previously-known regulated genes were discovered that responded to various stimuli. This new method for studying globally regulated genetic systems in *E. coli* combines detection, cloning and physical mapping of a battery of co-regulated genes in one step.

A GREEDY ALGORITHM FOR ALIGNMENT OF SIMILAR MULTIPLE SEQUENCES

Alexey S. Kondrashov. Laboratory of Genetics, University of Wisconsin, Madison, WI, 53706, USA

A new algorithm for alignment of many sequences is proposed. It works in steps which number cannot two times the length of the longest sequence. During the i-th step, (1) the i-th symbol is added to consensus and (2) a letter, corresponding to this i-th symbol, is found in each sequence. Problem (1) is solved using 'extrapolated consensus': decision on which symbol to add to consensus is based on 'how successfully' the remaining parts of sequences can be aligned. Problem (2) is solved by comparison of 'extrapolated consensus' with different regions of the compared sequences. The algorithm works properly when sequences are similar and are initially roughly prealigned and can be used for rapid and preciese refinement of alignments obtained by other methods.

END

DATE
FILMED
2/2/93