LBL--24830

DE88 007529

LBL-24830

Sequence Dependent Structure and Thermodynamics of DNA Oligonucleotides and Polynucleotides: UV Melting and NMR Studies

F.M. Aboul-ela
(Ph.D. Thesis)

Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

December 1987

Abstract

Thermodynamic parameters for double strand formation have been measured for the twenty-five DNA double helices made by mixing deoxyoligonucleotides of the sequence dCA_3XA_3G with the complement dCT_3YT_3G. Each of the bases A, C, G, T and I (I=hypoxanthine, designated as I by analogy with the hypoxanthine containing nucleotide, inosine) have been substituted at the positions labeled X and Y. The results are analyzed in terms of nearest neighbors. At room temperature the sequence $\binom{5'-A-A-3'}{3'-T-T-5'}$ is similar in stability to $\binom{-A-G-}{-T-G-}$, $\binom{-G-A-}{-G-T-}$, $\binom{-A-G-}{-T-C-}$, $\binom{-G-A-}{-C-T-}$; $\binom{-A-T-}{-T-A-}$ and $\binom{-T-A-}{-A-T-}$ are least stable. At higher temperatures the sequences containing a $G{\cdot}C$ base pair become more stable than those containing only $A{\cdot}T$. All molecules containing mismatches are destabilized with respect to those with only Watson-Crick pairing, but there is a wide range of destabilization. At room temperature the most stable mismatches are those containing guanine ($G{\cdot}T$, $G{\cdot}G$, $G{\cdot}A$); the least stable contain cytosine ($C{\cdot}A$, $C{\cdot}C$). At higher temperatures pyrimidine-pyrimidine mismatches become the least stable. $I{\cdot}C$ pairs were found to be less stable than $A{\cdot}T$ pairs in these duplexes. Large neighboring base effects upon stability were observed. For example, when (X, Y)=(I, A), the duplex is eightfold more stable than when (X, Y)=(A, I). Independent of sequence effects the order of stabilities is: $I{\cdot}C > I{\cdot}A > I{\cdot}T \approx I{\cdot}G$. All of these results are discussed within the context of models for sequnce dependent DNA secondary structure, replication fidelity and mechanisms of mismatch repair, and implications for probe design.

The cooperativity of the (righthanded helix)$B \rightleftharpoons$ (lefthanded helix) Z transition in $poly[d(^{5methyl}C-G)]$ has been investigated. A theoretical discussion is presented using an Ising model formalism similar to those used in the past to describe the helix to coil transition in proteins. Based on this formalism a statistical "order-order transition" model is used to simulate the distribution of B- and Z- form tracts at the midpoint of

the $B - Z$ equilibrium as a function of polymer length. Measured van't Hoff enthalpy values for the $B \rightleftharpoons Z$ transition in polymers of different lengths are compared with predictions based on the statistical model and a similar "order-disorder" model. The order-disorder model assumes that end residues favor the B-form over the Z-form. For both models the data are best fit for a cooperative unit of over 800 base pairs. Experimental observation of faster $B \rightarrow Z$ transition rates with increasing polymer lengths can be explained by a mechanism rate limited by nucleation within the polymer instead of at the ends, as predicted for an order-disorder system. A direct relationship between rates of the $B \rightarrow Z$ transition and the van't Hoff enthalpy values of the $B \rightarrow Z$ transition reflects a dependence of kinetics and cooperativity upon the energy of the nucleation event.

The duplex deoxyoligonucleotide *d(GGATGGGAG)·d(CTCCCATCC)* is a portion of the gene recognition sequence of the protein transcription factor IIIA (TFIIIA). The crystal structure of this oligonucleotide was shown to be A-form. The present study employs Nuclear Magnetic Resonance (NMR), optical, chemical and enzymatic techniques to investigate the solution structure of this DNA 9-mer. NMR COSY experiments indicate 16 of the 18 residues are predominantly south ($C_{2'}$-endo) sugar conformation. NMR NOESY indicates glycosidic angles in the range predicted for B-form DNA as opposed to A-form. Related DNA and RNA self-complementary 18-mer sequences, *d(GGATGGGAGCTCCCATCC)*, with U substituted for T in RNA, were studied by circular dichroism. Circular Dichroism (CD) spectra support B-form structures for the DNA 9-mer and the DNA 18-mer, and A-form for the RNA 18-mer. High trifluoroethanol concentrations induce a B to A-form transition in the DNA oligonucleotides. We find no evidence to support an A-form conformation for the TFIIIA recognition sequence *d(GGATGGGAG)·d(CTCCCATCC)* in solution.

In memory of my grandfather, Bernard Wygle and my uncle, Ahmed Aboul-Ela

Acknowledgements

Many individuals, too numerous to mention, have made generous and crucial contributions to this work. Unfortunately, space does not permit an exhaustive list of those who provided advice and support, so I must select a few whose contributions deserve special mention.

First on this list must be my advisor, Nacho Tinoco, who provided ideas, facilities and support. I am grateful for his efforts to find projects that even I would be interested in, his patience with numerous obnoxious remarks, and most of all for giving me the opportunity to visit Sicily.

The methodology of a work like this is heavily influenced by the intellectual environment in which it was carried out. Thus in this case the Tinoco lab has made a collective contribution through the ideas which have arisen from the scientific interaction and communication taking place between its members. Moreover, Barbara Dengler and David Koh passed on the fruits of the labors of others who worked in the lab before my time. Barbara provided instruction on experimental details of melting experiments and working with the Gilford. David synthesized DNA by hand and by machine and in some cases also performed purifications. In addition, Barbara performed the under-appreciated job of dealing with bureaucratic matters, while 'Yung Man' was a constant source of sometimes bizarre but often helpful advice on subjects ranging from Chinese restaurants to housing. Helen Lok is another important long-time member of the lab, a point which was demonstrated painfully during some of her vacations (except when her husband proved an able substitute).

In addition to Nacho - Terry Walker, Gabriele Varani and Frank Martin have contributed to the figures and text. Frank also synthesized most of the molecules studied in chapter II. Terry essentially conceived and performed the experiments discussed in the last part of chapter III and our subsequent discussions influenced the data analysis

and theoretical discussion in that chapter. Moreover, he was a frequent target of experimental questions. Terry and Gabriele were also collaborators during the experiments described in chapter IV. Gabriele's contribution has been especially important in the analysis and interpretation of NMR data and I wish him the best of luck in continuing this project. However, he needs to spend more time sleeping and less time in lab, lest he set a bad example for certain impressionable youngsters.

In addition to those mentioned above, several others have kindly read and edited portions of the text. Jackie Wyatt and Bill Thurmes were especially helpful in this regard. Bill also helped by directing the Tinoco lab welfare system, which has eliminated the once common site of scientists rummaging through the lab's trashcans searching for "spare data".

Discussions with Terry, Jody Puglisi and Chaejoon Cheng, among others, inspired much of the material in the appendices. Jody also helped make late night work sessions bearable by keeping me in constant contact with my mother.

My forerunner in the lab, Jeff Nelson, made rapid melt data acquisition and analysis possible by writing the first programs to automate the process in the lab. Phil Cruz wrote the programs for collecting melt data on the Apple. Phil also taught all of us a good deal about NMR and nucleic acid chemistry. Steve "hex dance" Wolk wrote and/or compiled many of the programs used for melt data analysis and provided helpful NMR advice, especially regarding "nuts and bolts" practical information.

The members of the NMR study group (Steve, Jody, Phil, Chuck Hardin, Pete Davis, Nacho and the members of the Wemmer group) gave me some motivation to learn basic NMR theory. Professor Wemmer and his group patiently answered questions and helped grapple with problems outside the study group as well. Chuck has been an excellent source of a biochemist's viewpoint on things. Pete rivals Jody as the nickname champion of the lab, but I wish he would get a hyphen in his name.

Numerous conversations with my officemates have kept up my morale over the years. Jackie, Gabriele and Jim Corbett have listened patiently and sympathetically to my whining and complaining. Jim is an excellent listener, while Jackie has mercifully acted as an apologist for the usually pathetic condition of my desk.

Brian Wimberly induced many episodes of laughter and silliness. I still have not heard anyone argue with my contention that he is one of the brighter students to join the lab within recent memory, and I wish him the best in his upcoming exam.

During my early years in the lab, the older grad students (Joe Kao, Emil Scoffone, Bruce Weir, Kathy Hall and David Keller) were always ready to drop anything to help me learn the ropes, while the postdocs (Bill Michols, Skip Shimer and Art Williams) took a 'postdoctorly' interest in the direction of my research.

Marcos Maestre, Milan Tomic, and the members of the Hearst group have been helpful. Bill Ross, Cynthia Phillips, Madhu Wahi, Edie Leonhardt and Sarah Schofield will be remembered.

The support of many non-scientist friends has been most important to me, including, among others, Atef, Shay, Ali, Mesut, Shawn, Harisankar, Rajiv and Mahmoud. Finally, I want to thank my parents for all of their patience, support and encouragement.

Table of contents

Chapter I

"This (physics) is a science that investigates bodies from the point of view of the motion and stationariness which attach to them. It studies the heavenly bodies and the elementary (substances), as well as the human beings, the animals, the plants, and the minerals created from them."

Ibn Khaldun "The Muqaddimah"

ca. 1477 (trans. F. Rosenthal)

A twentieth century "man in the street" would probably ask Ibn Khaldun and a biophysicist the same question, "what does physics have to do with biology?" The question is certainly relevant to a biophysicist, who might find it useful to phrase the question in the following way; "what can the approach(es) of physics and physical chemistry contribute to the study of and understanding of biology?" Physics attempts to describe the real world in terms of a set(s) of mathematical principles. This set of principles and the overall structure of the set is called a *formalism*, of which there are many in physics.

This thesis presents applications of one of the formalisms of physics (thermodynamics) along with several of its experimental techniques (Ultraviolet, Circular Dichroism, Nuclear Magnetic Resonance spectroscopies) to the study of a biologically significant molecule (deoxyribonucleic acid or DNA). The approach is to understand the structure, the dynamics, and factors that determine the structure of DNA. It is hoped that such knowledge will ultimately help biologists to better describe the biological functions of DNA. Of course, this should not imply that physical studies are worthwhile only with the assurance of a "biologically relevant" result, since the physical chemistry or practical applications of the system may be worthwhile in themselves. Yet for most biophysicists the real motive for their research (and their funding) is to learn something about biology. At the very least by defining what DNA is physically

capable of doing, certain biological models can be eliminated. In other words, the theoretical framework will not be applied to the biological problem directly. Rather it will be applied to physical problems whose solutions could provide biological insight.

The thermodynamic formalism and its application to problems of biological interest

The formalism used in most of the studies presented below is that of classical equilibrium thermodynamics (1). The formalism is built upon the set of four fundamental postulates known as the zeroth, first, second, and third laws of thermodynamics. The first three of these state; Equilibrium can be defined, or, in other words, temperature exists (zeroth). The energy of an isolated system is conserved (first). It is possible to define a quantity which is a function only of the macroscopic state of the system, and which reaches its maximum at the equilibrium state for an isolated system (second). A macroscopic state is one that is presented in the language of everyday experience as opposed to a microscopic description in the language of atoms and molecules. The quantity defined in the second law is commonly called the entropy. The rest of the formalism can be derived from these fundamental assumptions.

This formalism stands on its own for macroscopic systems at equilibrium, but it does not predict the microscopic behavior of atoms and molecules. However, Boltzmann proved in his famous H-theorem (3) that the maximum entropy postulate can be derived for a macroscopic system from a microscopic model. The conclusions of the H-theorem may be summarized as follows; Consider a system consisting of a very large number of identical particles which interact only through "elastic collisions". Elastic collisions conserve momentum (defined as mass × velocity=$M\vec{V}$) and kinetic energy (=$\frac{1}{2}MV^2$). It is also assumed that the probability of two particles with momenta i and j colliding to produce two particles with momenta k and l ($M\vec{V}_i + M\vec{V}_j = M\vec{V}_k + M\vec{V}_l$) is equal to the probability of the reverse process in which a pair of particles with momenta k and l collide to produce particles with

momenta i and j. Then consider a function, H, defined as $logW(N_1, N_2, N_3...)$. N_i is the number of particles in the i^{th} accessible momentum state for the individual particles. $W(N_1, N_2, N_3...)$= the total number of configurations of the system (or *microstates*) in which N_1 particles are in momentum state 1, N_2 are in state 2, and so forth. Then it can be shown that, starting with an arbitrary configuration of the system with a given distribution of particles in their individual states, the function H=$logW$ increases with time until reaching a maximum at equilibrium. At equilibrium $\frac{N_i}{N_j} = e^{-\frac{(E_i - E_j)}{RT}}$, where E_i is the kinetic energy of the i^{th} state, R is known as Boltzmann's constant, 1.98717 cal K^{-1} Mol^{-1}, and T is defined as the temperature.

Note that the H-theorem suggests a possible physical interpretation of the entropy. One may view the entropy as a consequence of the fact that in a macroscopic measurement one's knowledge of the microscopic configuration of the system is incomplete. In this case, since the particles making up the system are indistinguishable, one can measure only the *total* number of particles in each state without knowing which state a *specific* particle is in. The entropy of a macroscopic state is identified with $K \times H$, so that it is proportional to the number of microscopic configurations or microstates which, from the point of view of the measurement, all look like that particular macrostate. The entropy goes to a maximum for a large number of particles in equilibrium because different microstates, provided they satisfy the conservation of energy and momentum, are equally probable. This means that the macrostate that one detects is most likely to be the one which corresponds to the largest number of microstates.

Statistical Mechanics is the formalism which deals with this connection between microscopic or atomic theory and macroscopic thermodynamics (3). Though macroscopic thermodynamics needs no statistical mechanical justification, derivations such as Boltzmann's H-theorem, probably one of the most elegant proofs in the history

of science, provide valuable insight into the physics of macroscopic systems. For example, according to the H-theorem, there is always a certain probability that, even after having settled into equilibrium, a macroscopic system at a given moment will be found in a state which does not correspond to its equilibrium state. Such fluctuations, as they are called, become more important as the number of components becomes smaller. In the "thermodynamic limit", in which the number of components of the system approaches infinity, the probability of the existence of a nonequilibrium state is negligible and the equilibrium "maximum entropy" assumption can be said to be satisfied exactly.

All of the experiments presented in this thesis have been performed on solutions containing on the order of 10^{13} or more DNA and solvent molecules. Considering the system in the thermodynamic limit is then an excellent approximation, and the application of the equilibrium thermodynamic formalism is fully valid. One caution needs to be mentioned. The probes used to determine the state of the system are for the most part sensitive to the state of the DNA. The configuration of the solvent molecules has not been probed directly, however, any thermodynamic parameters measured describe the state of the whole system. Consequently, there is a "hidden" contribution to measured parameters due to the effect of solvent, which complicates any microscopic, physical interpretation of the data. The measured parameters are still valid, but for the complete system including solvent, not just the DNA.

Extrapolating the results of such experiments in order to draw conclusions about biological systems is more problematic, for at least four reasons. First, biological systems are extraordinarily complex. DNA is never found in a living cell merely floating in an aqueous solution, rather it is almost always complexed with a number of proteins, membranes, and other cell components. Second, biological systems are open systems, whereas thermodynamic laws such as the conservation of energy apply to

closed systems. The third problem is actually a consequence of the second. Biological systems do not have to exist in equilibrium, since they are open systems and can input energy from their environment in order to maintain a nonequilibrium steady state. Attempts to apply nonequilibrium formalisms (4, 5) to biological problems has thus far resulted in limited success with real systems (6). The fourth problem is that at the cellular level, where DNA carries out its biological functions, biological systems are heterogeneous. In other words, a cell does not contain a large number of identical DNA molecules. This means that fluctuations away from equilibrium and nonequilibrium structures are potentially of great importance.

Taking all of these factors into consideration, it is clear that describing biological processes using thermodynamics is a difficult task. Nonetheless, thermodynamic studies are bound to increase understanding of the relationship between microscopic structure and biological function. The thermodynamic formalism is particularly powerful simply because unlike all of the other major formalisms in physics (other than classical mechanics) it deals with easily measurable macroscopic properties. Thus it is a most convenient tool for indirect investigation of what are actually microscopic properties, especially when complemented by spectroscopic probes. Whatever microscopic or molecular property one may choose to study, it is usually macroscopic data which are accessible. For example, an experiment designed to probe structure often can only probe the average structure resulting from a thermodynamically controlled distribution of structures. Ultimately, the physical characterization of a biological macromolecule using thermodynamics and physical probes will delineate what biological processes require input of energy from outside sources and which can occur spontaneously. All of this describes only part of the potential importance of thermodynamics to the study of molecular biology.

DNA structure, DNA structural polymorphism and "biological relevance"

DNA stores the genetic information in the cell (7). The structure of DNA proposed by Watson and Crick is known as B-DNA. It consists of two strands wrapped around each other to form a right handed double helix. Each strand contains a sequence of individual units-each of these units contains one of the four standard bases, adenine (A), cytosine (C), guanine (G) or thymine (T)-and these individual residues are linked to each other through a sugar and phosphate backbone. The bases are complementary to each other, so that normally A pairs with T and G pairs with C. Consequently, each strand can act as a template for the replication of a daughter strand containing the same sequence as the complementary strand. In this way the genetic information, which is stored in the base sequence, can be passed on during cell division.

However, the B-DNA structure of Watson and Crick is not the only structure which DNA can adopt (8). The possible alternatives include structures which maintain the base complementarity of B-form DNA and others which do not. Among the alternative structures which maintain base complementarity is A-DNA. A-DNA is also a right handed double helix, however its structure is dramatically different from B-DNA in the following respects (among others); the conformation of the sugar residues is different, the tilt of the base pairs with respect to the helix axis is much more dramatic and of opposite sense in A-form as compared to B-form, the relative widths of the two helical grooves are different, and the winding angle between neighboring base pairs is smaller in A-form than in B-form. Very soon after Watson and Crick published their work X-ray crystallographers found that most DNA fibers when dehydrated formed A-DNA, while hydrated fibers were B-form. It was therefore assumed that in solution DNA would be B-form. In the nineteen seventies it was discovered that alternating $d(C-G)_n$ sequences could form a left handed double helix in crystals and in high salt solutions. The new structure was named Z-DNA, due to a zig-zag structure in the phosphate

backbone. DNA can also form A, B or Z helices in which some of the opposing bases are not complementary in the Watson-Crick sense. Such non complementary base oppositions are called mismatches. DNA can also exist as single strands. Even for mixtures of complementary strands. there are conditions (usually high temperature and low concentration) in which the single stranded state is thermodynamically favored. Several additional DNA structures have been identified. Even B-DNA has been found to exhibit surprising local variability in structure according to base sequence.

It is not known which conformation(s) DNA favors *in vivo*. Traditionally, the B-form was assumed dominant since it is favored by hydration and cells contain mostly water by volume. In fact, it is necessary for DNA to undergo transitions between structures in order to perform its biological functions. For example, the double helix must unwind to single strands for the genetic information in the DNA base sequence to be read (7). There are at least three factors present within the cell which may induce DNA to convert from canonical B-form to other structures. 1) DNA in the cell is complexed with proteins and highly compacted. These factors may remove the scaffolding of water molecules attached to the DNA surface which stabilize the B-form. 2) DNA in the cell is under topological stress. Most DNA extracted from cells is found to be underwound. The stress produced by this underwinding of DNA can be relieved by the conversion of a segment to the unwound single stranded form, the underwound A-form, or the oppositely wound Z-form. 3) There is a sequence dependence to DNA structure. After all, each of the four standard bases is a unique compound with unique properties. Hence it is not surprising that when different bases are stacked next to each other they favor different structures. *In vitro* physical studies have established the ability of all these factors to influence DNA structure.

The potential biological ramifications of DNA structural polymorphism in living cells are enormous. Though the "genetic code", that is, the code used by the cell's

machinery to read the genetic information from the DNA base sequence, is well understood, it is still not well understood how the expression of genes is controlled. The DNA base sequence in an individual's cells is identical throughout an organism, yet the expression of various genes and therefore the function of a cell in the eye, for example, is different from that of a cell in the liver. Moreover, the expression of genes in a single cell varies during different stages of development, and even in one celled prokaryotes rates of gene expression change in response to environmental conditions. Conformational flexibility of DNA provides several possible mechanisms for the regulation of gene expression. Sequence dependent conformational variability is a possible mechanism for recognition of consensus DNA sequences by proteins which are known to be involved in gene regulation. Moreover, a conformational transition in a DNA segment could act as a "conformational switch" turning genes off and on. Some more specific speculative models are mentioned in the introductions to chapters II-IV.

It is therefore of interest to study the fine structure of DNA; to understand how DNA structure depends on base sequence and to determine the characteristics of the transitions between the major families of DNA structures such as A, B, Z and single stranded DNA. It is also of interest to determine the effect of mismatched bases on DNA structure and thermodynamics, since a mismatch in a cell's DNA may lead to a mutation when the DNA is replicated (7).

The scope of this work

Several physical techniques are available to investigate DNA structure and thermodynamics. This work makes use of three types of spectroscopy; ultraviolet (UV), circular dichroism (CD), and proton nuclear magnetic resonance (^{1}H NMR).

Ultraviolet absorbance is a convenient technique to monitor conformational transitions in DNA as a function of temperature. A large decrease in absorbance of 260 nm

wavelength radiation, known as a hypochromic effect, is observed in DNA upon the annealing of the two single strands to form a double helix. A similar hypochromicity is observed at 295 nm during a Z-DNA to B-DNA transition.

Circular dichroism spectroscopy, which measures the difference between the absorbance of right and of left circularly polarized radiation (divided by the total absorbance) as a function of wavelength, is very sensitive to subtle features of the geometry of base stacking. It is an excellent "fingerprint" for A, B and Z forms of DNA.

Nuclear magnetic resonance detects the magnetization of spins in atomic nuclei induced in the presence of a magnetic field. NMR studies in this thesis involve detection of magnetization of proton spins (^{1}H NMR). Of particular interest are two processes which involve exchange of magnetization between protons; the nuclear Overhauser effect (NOE), which involves short range through space coupling between spins and reveals information about distances between nearby protons, and J-coupling, a through bond coupling which is sensitive to bond angles. NOEs and J-coupling provide the most sensitive probes available for details of molecular structure in solution.

Temperature dependent UV absorption is applied to the study of the double helix to single strand transition in chapter II. Thermodynamic parameters have been measured for the set of 25 short double helices of the base sequence $dCA_3XA_3G+dCT_3YT_3G$; X, Y=A, C, G, T and I, where I is the base nucleotide analog inosine. The results provide insight into the effect of mismatches on thermodynamic stability of double helical DNA, as well as the effect of base sequence on the thermal stability of Watson-Crick paired DNA. Measurements are presented in $1M$ NaCl solution and in $150mM$ KCl, $30mM$ $MgCl_2$ solution in order to explore the effect of ionic environment on relative stabilities. The experiments presented in chapter II have been published elesewhere (9, 10). Chapter III extends the approach of obtaining thermodynamic parameters from

UV temperature curves to the $B \rightleftharpoons Z$ transition in a polymer, $poly[d(^{5me}C - G)]$. A theoretical discussion of the cooperativity of the $B \rightleftharpoons Z$ transition is presented using a statistical mechanical formalism based on those established for the helix to coil transition in proteins. Experimental results obtained from UV studies of $poly[d(^{5me}C - G)]$ samples fractionated according to size are then evaluated within the context of the theoretical discussion. The experimental work presented in chapter III along with some discussion is also contained in (11). Chapter IV presents a ^1H NMR study of the solution structure of a DNA fragment $dGGATGGGAG + dCTCCCATCC$, which forms a portion of the binding site for a protein which participates in the developmental control of genes in a frog, *Xenopus Laevis*. Most of the material in Chapter IV has also been submitted for publication (12).

References

1. Callen, H. B. (1960) Thermodynamics, John Wiley and Sons, New York.

2. Reif, F. (1965) Fundamentals of Statistical and Thermal Physics, McGraw-Hill,

3. Landau, L. D. and Lifshitz, E. M. (1980) Statistical Physics, 3rd ed., part 1, Pergamon Press, New York.

4. Katchalsky, A. and Curran, P. F. (1965) Nonequilibrium Thermodynamics in Biophysics, Havard University Press, Cambridge.

5. Prigogine, I., (1968) Introduction to Irreversible Processes, 3rd ed., Interscience Publishers, New York.

6. Winfree, A. T. and Strogatz, S. H. (1984) Nature 311, 611 – 615.

7. Watson, J. D. (1987) Molecular Biology of the Gene, Addison-Wesley.

8. Saenger, W. (1983) Principles of Nucleic Acid Structure, Springer-Verlag, New York.

9. Aboul-ela, F., Koh, D., Martin, F. H. and Tinoco, I., Jr. (1985) Nuc. Acids Res. 13, 4811 – 4824.

10. Martin, F. H., Castro, M. M., Aboul-ela, F. and Tinoco, I., Jr. (1985) Nuc. Acids Res. 13, 8927 – 8938.

11. Walker, G. T. and Aboul-ela, F. (1987) submitted to J. Biomol. Struc. Dynamics.

12. Aboul-ela, F., Varani, G., Walker, G. T. and Tinoco, I., Jr. (1987) submitted to Nuc. Acids Res.

Chapter II

A. Introduction

The thermodynamic stability of mismatched bases affects the probability of incorporating the wrong base during replication, and of repairing the mistake during proofreading by the polymerase (1). Thus, thermodynamics is important in the study of mechanisms of mutations. Moreover, thermodynamic studies can help in understanding the sequence dependence and the polymorphism in secondary structure of DNA, which has been shown clearly by Dickerson (2). Distinct secondary structures may be involved in the recognition of base sequence by proteins which bind to DNA (see the introduction to chapter IV).

This chapter presents and discusses the measurement of thermodynamic parameters for the helix to coil transition of 25 double helices of the sequence $dCA_3XA_3G + dCT_3YT_3G$, where X, Y are all possible combinations of the four Watson Crick base pairs, A, C, G, T, and the base analog hypoxanthine (designated as I for the hypoxanthine containing nucleotide, inosine). The data are presented in four parts. First the data for the four Watson-Crick paired helices of the series is presented along with a discussion of the results within the context of the sequence dependent thermodynamics and structure of DNA, and particularly the properties of the $(dA_n)\cdot(dT_n)$ sequence. Then sets of parameters for the twelve mismatched molecules of the series which contain the standard bases is discussed along with the implications for the role of mistakes in DNA replication and proofreading in the process of mutation. Data for nine molecules which contain inosine are presented in the third section. Whereas measurements in the first three sections were all performed in buffer solutions containing $1M$ NaCl, the final section presents measurements for three of the Watson-Crick paired sequences from the above series in buffer containing $150mM$ KCl and $30mM$ $MgCl_2$, conditions more closely approximating "physiological" salt concentrations.

B. Watson-Crick paired molecules

Background

The binding of drugs and proteins to DNA can induce structural transitions in DNA, and these induced structural transitions may be involved in the regulation of gene expression. A well studied example is the interaction of RNA polymerase with DNA during the process of transcription (3). The binding site of *E. coli* RNA polymerase (within a region known as the promoter) contains two regions of semi conserved base sequence. However, a single sequence does not function as a recognition site for the protein and its co-factors throughout the entire genome. Rather the protein recognizes and initiates transcription at several sites with varying efficiencies (3). Since melting out of a portion of the DNA is believed to occur during the formation of an "open promoter" complex which precedes the initiation of transcription, it is likely that sequence dependent variability in the thermodynamic stability of recognition sites could affect their promoter strength (4). Moreover, as the polymerase proceeds with transcription, it apparently denatures the region of DNA in contact with it (5). The polymerase is also known to pause at certain sites (6), possibly influenced by the thermodynamic stability of the DNA at the pause site.

Evidence for a biological role for DNA structural flexibility comes from "action at a distance" (7) in control of gene expression, in other words, the regulation of activity (i.e. binding of a regulatory protein) at one site on a gene through the binding of a protein at a distant site. For example, the binding of CAP protein in the lac operon of *E. coli* promotes RNA synthesis at a site as much as 100 base pairs away. The mechanism(s) of action at a distance are not understood, however it could occur through a variety of mechanisms mediated by DNA conformational transitions under torsional stress (7) and/or by protein-protein or protein-DNA interactions mediated by DNA bending (8, 9). DNA bending has been observed in sequences similar to those

used in this study (10, see below).

For some years systematic studies of double strand formation have been done using oligonucleotides of specific base sequences, with the goal of being able to predict thermodynamic parameters for DNA and for RNA secondary structure (4, 11-18, see also references in 15). The usual method has been to measure melting temperatures (T_m) at several concentrations by monitoring absorbance as a function of temperature and to obtain thermodynamic values from a van't Hoff analysis. The oligonucleotide studies have led to values for free energies, enthalpies, and entropies for double strand formation in RNA and DNA based on nearest neighbor interactions.

Experimental procedures

Deoxyoligonucleotides were synthesized by the phosphoroamidite method (19). Purification was by RPC-5 chromatography after deblocking. Melting curves were obtained by a method similar to that described earlier (16). The buffer in all cases contained $1M$ NaCl, $0.1mM$ EDTA $10mM$ phosphate in H_2O at pH 7.

Thermodynamic values were obtained using the van't Hoff method (17). Absorbance (A) vs. temperature (T) curves for several concentrations of one duplex are shown in figure $II-1$. For a two state model and assuming equimolar concentrations of non-self complementary strands, the equilibrium constant can be written as

$$K = 2f/(1-f)^2C_t \qquad (II-1)$$

Where f is the fraction of strands in the double-stranded state and C_t is the total concentration of all single strands. At any temperature the single stranded fraction can be obtained from

$$1-f = [A(T) - A_d(T)]/[A_s(T) - A_d(T)]$$

where $A_s(T)$ and $A_d(T)$ are the absorbance of the single strand (upper baseline) and double strand (lower baseline) at temperature T. The method for obtaining baselines is explained below (see also appendix II). The equilibrium constant can also be written

as

$$K = exp(-\Delta G^0/RT) = exp(-\Delta H^0/RT + \Delta S^0/R) \qquad II-2$$

At the melting temperature, T_m, $f = 1/2$ and we can combine Eqs. (II-1) and (II-2) to write

$$Rln(C_t/4) = (\Delta H^0/T_m) - \Delta S^0$$

Thus if the difference in standard enthalpy, ΔH^0, and the difference in standard entropy, ΔS^0, between the double strand and the single strand are assumed to be independent of temperature, they can be obtained from a plot of lnC_t versus $1/T_m$. Such a plot is shown in Figure $II-2$.

The baselines were obtained from a linear least squares fit to ten points chosen near 0^0C for the lower baseline and near 65^0C for the upper baseline. The same upper baseline was used for the melting curves at different concentrations for the same molecule; this minimizes the effect of choice of baseline on the thermodynamic parameters. Because hypochromicity was found to increase slightly with concentration (apparently due to aggregation), the lower baseline intercept for each experiment was chosen based on the absorbance recorded at 0^0C. Data points from the melting curves at the lowest concentrations were used to obtain all upper baselines, and data from curves taken at the highest concentrations provided the slopes for lower baselines. For helices that melt at temperatures too low to provide data for lower baselines, calculations were done using an assumed flat lower baseline. For all helices, the standard free energy was calculated from the relation between ΔG^0 and T_m.

$$\Delta G^0(T_m) = RT_m ln(C_t/4)$$

At 25^0C ΔG^0 is obtained by extrapolation from the least squares fit for $ln(C_t/4)$ vs. $1/T_m$ to the concentration, C_t, for which the melting temperature is 25^0C.

Results and discussion

Measured thermodynamic parameters for double helix formation are given in Ta-

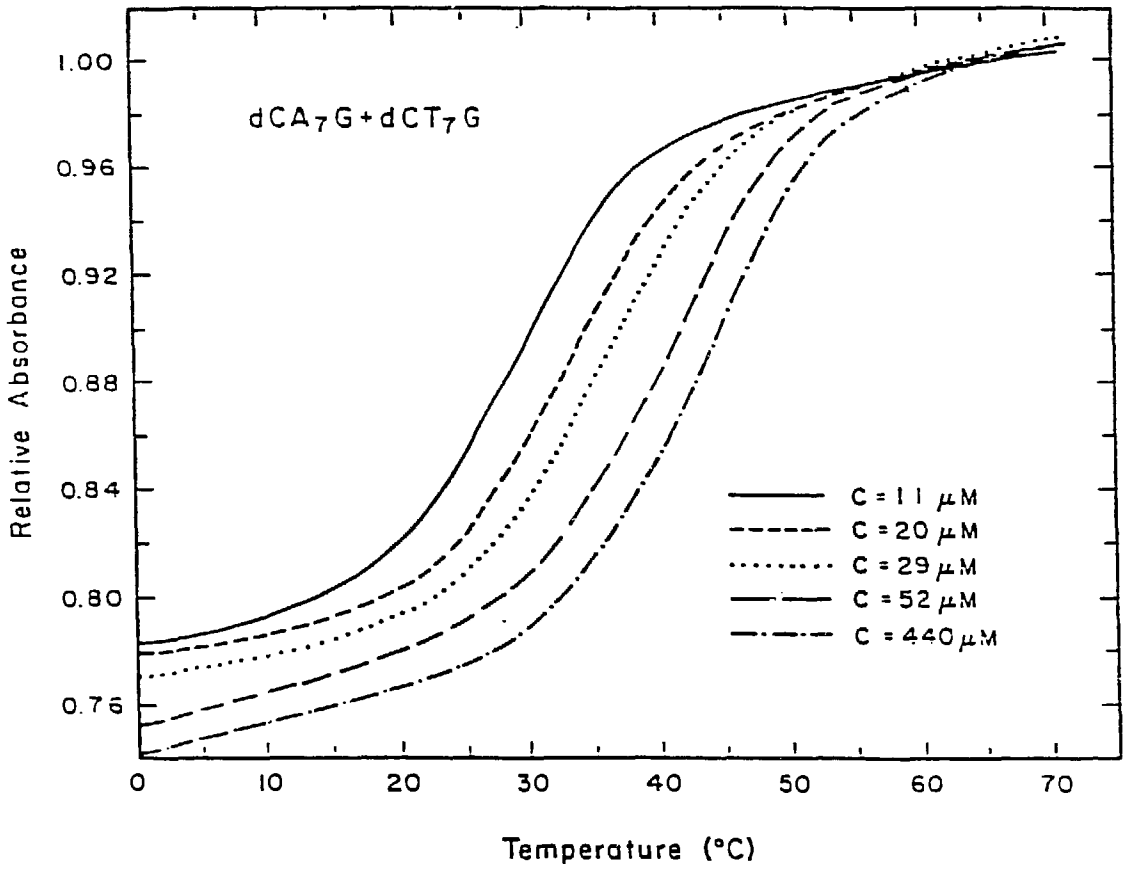
ble II-1 for all molecules of the sequence $dCA_3XA_3G{\cdot}dCT_3YT_3G$ where all possible combinations of A, C, G and T are substituted for X and Y. This set of sequences was chosen in order to favor all or none melting and to facilitate comparison with earlier work (16-18, 20). Also included is the helix with one less $A{\cdot}T$ base pair ($dCA_6G{\cdot}dCT_6G$) and two helices with an extra nucleotide on one strand ($dCA_3CA_3G{\cdot}dCT_6G$ and $dCA_6G{\cdot}dCT_3CT_3G$). The values of ΔH^0, ΔS^0, ΔG^0 (25^0C) are referred to standard condition (1 molar concentration of each single strand reacting to form 1 molar concentration of duplex) in 1 M NaCl, pH 7, 10 mM phosphate, 0.1 mM EDTA. The melting temperatures, T_m, are given in this buffer for a total single strand concentration of 400 μM. The duplexes are arranged in order of thermodynamic stability as measured by their free energy of formation from the single strands at 25^0C.

The four Watson-Crick paired duplexes are, as expected, considerably more stable than those containing a non-Watson-Crick base opposition. The duplex containing seven $A{\cdot}T$ pairs with the A's on the same strand is of comparable stability to the duplexes which replace an $A{\cdot}T$ pair with a $G{\cdot}C$ pair. The other $A{\cdot}T$ duplex, in which the series of A's is interrupted by a T residue is less stable by +1.0 to 1.6 kcal mol^{-1} for $\Delta G^0(25^0C)$; this corresponds to an order of magnitude decrease in the equilibrium constant for double strand formation. The order of duplex stability is $dCA_3CA_3G{\cdot}dCT_3GT_3G > dCA_7G{\cdot}dCT_7G \approx dCA_3GA_3G{\cdot}dCA_3CA_3G > dCA_3TA_3G{\cdot}dCT_3AT_3G$.

A nearest neighbor analysis of $dCA_3XA_3G + dCT_3YT_3G$

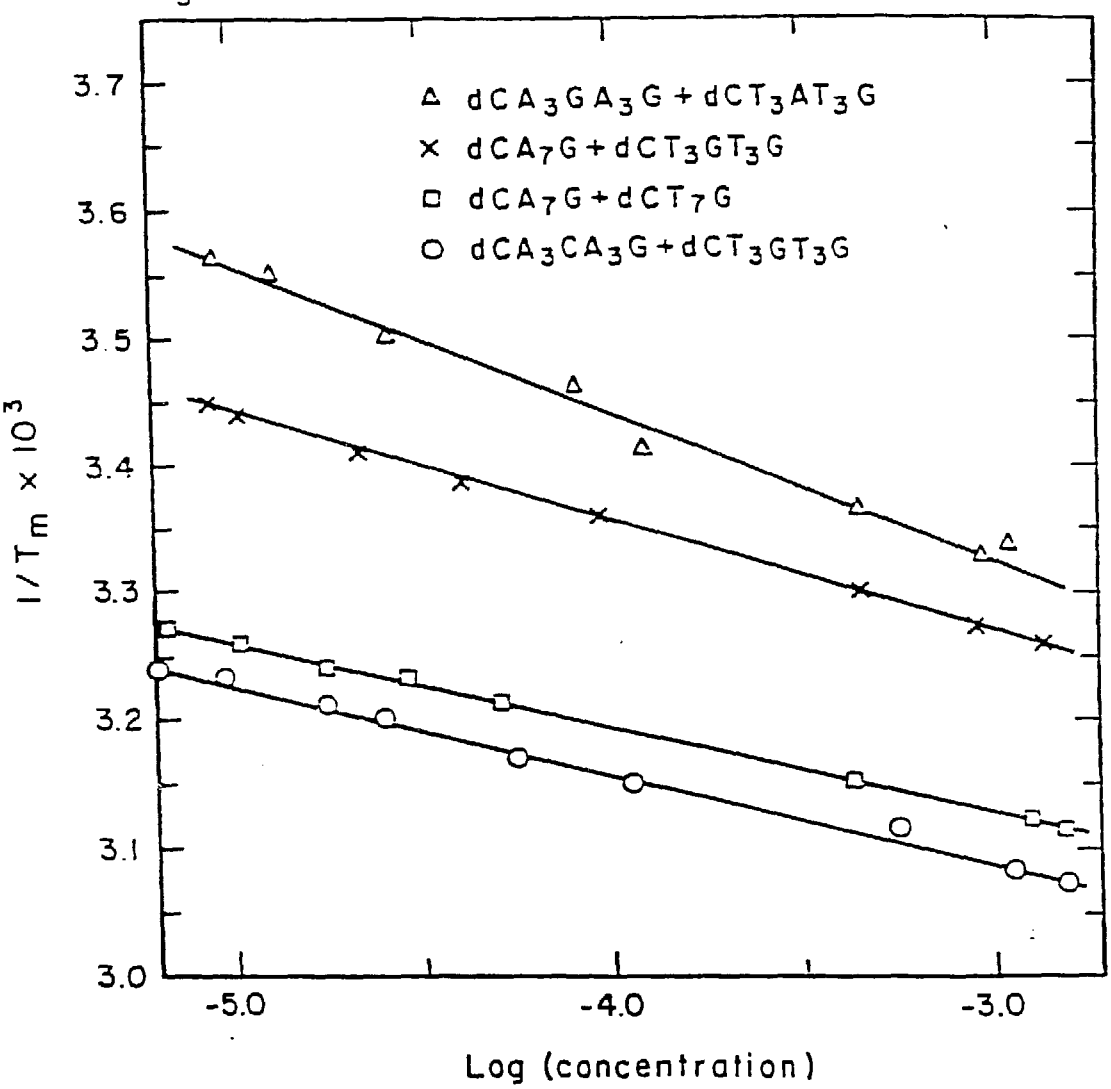
Estimation of helical stability of nucleic acids has often been based on analysis of oligonucleotide duplexes in terms of nearest neighbor contributions (12, 14). The key assumption is that thermodynamic properties of an oligonucleotide are the sum of the properties of neighboring base pairs taken two at a time. A nearest-neighbor

Figure II-1



Melting curves showing absorbance vs. temperature for five concentrations of $dCAG_7 \cdot dCT_7G$ in 1 M NaCl, pH 7, 10 mM phosphate, 0.1 mM EDTA. The total concentrations of single strands range from 11μM to 440 μM. All the curves are normalized to absorbance of 1 at 65°C.

Figure II-2



van't Hoff plots of $1/T_m$ *vs.* log C_T where T_m is the melting temperature and C_T is the total strand concentration. The lines shown are the best least squares fit to the data.

Table II-1 Van't Hoff Thermodynamic Values for Double Helix Formation of $dCA_3XA_3G + dCT_3YT_3G$ in 1 M NaCl, pH 7.

X·Y	$\Delta G°$(kcal mol^{-1})[a] 25°C	50°C	$\Delta H°$ (kcal mol^{-1})[b]	$\Delta S°$ (cal $deg^{-1} mol^{-1}$)[c]	T_m(°C) C_T=400μM[d]
C·G	-10.1	-5.5	-64.5	-183	48°
A·T	- 9.6	-4.7	-68.0	-196	45°
G·C	- 9.5	-5.0	-62.8	-179	50°
T·A	- 8.5	-4.3	-58.6	-168	41°
-·-[e]	- 7.8	-3.3	-59	-172	37°
T·G	-6.5	-2.4	-55.6	-165	31°
G·G	-6.3	-2.4	-53.5	-158	30°
G·A	-6.2	-2.3	-52.6	-156	30°
G·T	-5.8	-2.4	-46.7	-137	27°
A·G	-5.3	-2.4	-39.9	-116	24°
C·T	-5.3	-1.3	-53.2	-161	24°
T·C	-5.0	-1.2	-50.0	-151	22°
A·A	-5.0	-2.3	-36.9	-107	21°
T·T	(-5.0)[h]	-0.8	(-54.6)	(-167)	(22°
C·-[f]	-4.9	-0.9	-53.0	-161	22°
C·A	(-4.6)	-1.6	(-40.3)	(-120)	(19°
C·C	(-4.5)	-0.2	(-55.3)	(-171)	(20°
A·C	(-4.4)	-1.8	(-35.8)	(-106)	(19°
-·C[g]	(-4.2)	-0.8	(-45.0)	(-135)	(16°

[a]Estimated precision in $\Delta G°$ is ±0.1 kcal mol^{-1}

[b]Estimated precision in $\Delta H°$ is ±3 kcal mol^{-1}

[c]Estimated precision in $\Delta S°$ is ±9 cal deg^{-1} mol^{-1}

[d]Estimated precision in T_m is ±1°

[e]$dCA_6G·dCT_6G$. Data from reference 20

[f]$dCA_3CA_3G·dCT_6G$. Data from reference 20

[g]$dCA_6G·dCT_3CT_3G$.

[h]Data in parentheses are significantly less accurate. An estimated flat lower base line was used to obtain the (1-f) vs. T curve.

Table II-2 Nearest-neighbor Contributions to Double Strand Formation in 1M NaCl, pH 7.[a]

Nearest Neighbor	ΔG°,25°C (kcal mol^{-1})	ΔH° (kcal mol^{-1})	ΔS° (cal $deg^{-1} mol^{-1}$)
$\begin{matrix}-A-A-\\ -T-T-\end{matrix}$	−1.5±0.2	−10.2	−29
$\frac{1}{2}\left(\begin{matrix}-A-T-\\ -T-A-\end{matrix} + \begin{matrix}-T-A-\\ -A-T-\end{matrix}\right)$	−1.0±0.2	−5.6	−15
$\frac{1}{2}\left(\begin{matrix}-A-G-\\ -T-C-\end{matrix} + \begin{matrix}-G-A-\\ -C-T-\end{matrix}\right)$	−1.5±0.2	−7.6	−21
$\frac{1}{2}\left(\begin{matrix}-A-C-\\ -T-G-\end{matrix} + \begin{matrix}-C-A-\\ -G-T-\end{matrix}\right)$	−1.8±0.2	−8.5	−22

[a]The values given are for the reaction $\begin{matrix}-W_1 & W_2-\\ -C_1 & C_2\end{matrix} \rightleftharpoons \begin{matrix}W_1- & W_2-\\ -C_1- & C_2-\end{matrix}$

analysis of the oligonucleotide used in this study is

$$\binom{CA_3XA_3G}{GT_3YT_3C} = [\,\binom{C-A}{G-T} + 4(\,\binom{A-A}{T-T}\,) + \binom{A-G}{T-C}\,] + (\,\binom{A-X}{T-Y} + \binom{X-A}{Y-T}\,) \quad (II-3)$$

The terms in square brackets equal the nearest neighbor contribution of $CA_5G{\cdot}CT_5G$. Instead of using the experimental data for this duplex, we use a least squares fit to the experimental data (18, 19) for $CA_nG{\cdot}CT_nG$ (n=5, 6, 7). This gives a best nearest-neighbor equivalent for $CA_5G{\cdot}CT_5G$. By subtracting the thermodynamic parameters for this $CA_5G{\cdot}CT_5G$ equivalent from measured values in Table II-1, we obtain the nearest-neighbor thermodynamic contributions for the terms in parentheses. For example

$$\Delta G^0(\,\binom{A-X}{T-Y} + \binom{X-A}{Y-T}\,) = \Delta G^0(dCA_3XA_3G{\cdot}dCT_3YT_3G) - \Delta G^0(dCA_5G{\cdot}dCT_5G) \quad (II-4)$$

Comparing only these very similar duplexes we obtain data which do not require assumptions about free energies of initiation or about end effects.

Table II-2 gives $\Delta G^0(25^0C)$, ΔH^0 and ΔS^0 for the nearest-neighbor contribution to formation of Watson-Crick base pairs in a double strand. The $(\binom{A-A}{T-T})$ contribution to ΔG^0 at 25^0C is much greater than the average obtained from $\frac{1}{2}\,(\,\binom{A-T}{T-A} + \binom{T-A}{A-T}\,)$, and is more nearly equal to $\frac{1}{2}\,(\,\binom{A-C}{T-G} + \binom{C-A}{G-T}\,)$ and $\frac{1}{2}\,(\,\binom{A-G}{T-C} + \binom{G-A}{C-T}\,)$. Note that this means replacing an $A{\cdot}T$ base pair by a $G{\cdot}C$ base pair does not always increase the stability of the double helix in DNA. The ΔG^0 data in Table II-2 are consistent with those obtained from a much more extensive set of oligonucleotides measured by Markey and Breslaur (12). The ΔG^0 values are more directly related to the measurements ($\Delta G^0 = -RTlnK$) and are thus more accurate than values of ΔH^0 and ΔS^0, particularly for the less stable duplexes.

The anomolous properties of dA_ndT_n

Although ΔH^0 and ΔS^0 values are less reliable than ΔG^0 the magnitudes of the effects seen in Table II-1 and II-2 for Watson- Crick base pairs merit comment. The most negative enthalpy (favorable) and most negative entropy (unfavorable) occur for formation of $dCA_7G{\cdot}dCT_7G$. The helix with the center $A{\cdot}T$ reversed ($dCA_3TA_3G{\cdot}dCT_3AT_3G$),

has a much less favorable enthalpy and a more favorable entropy; the net effect is a less favorable free energy. Since the publication of this data (15), the anomolously low enthalpy of the $dA_n dT_n$ sequence has been confirmed (21). The special stability of the $\left(\begin{smallmatrix}-A-A-\\-T-T-\end{smallmatrix}\right)$ sequence may be linked to the unique properties of *polydA·polydT*. Experiments using the band shift method (22-24) to measure helical repeat lengths of DNA sequences in supercoiled plasmids have shown that the sequence $dA_n dT_n$ forms a helix with a pitch of $10(\pm)0.1$ base pairs, compared to a pitch of $10.6(\pm)0.1$ for all other sequences measured. Moreover, Crothers and coworkers (10, 25) have suggested that the sequences $dA_5 dT_5$ and $dA_6 dT_6$ in a restriction fragment have a non-standard secondary structure, causing a bend at the junction between the $dA_n dT_n$ sequences and B-DNA sequences. Other unusual properties found for *polydA·polydT* include cooperative binding of intercalating drugs (26, 27), the lack of capacity to form chromatin in the presence of nucleosomes (28), the lack of capacity to form A-DNA in low humidity fibers (29), and the capacity to induce specific antibody production (30). These properties may be related to the *polydA·polydT* structure observed in fibers (31, 32), which apparently has an exceptionally narrow minor groove. Dickerson (33) has postulated that the relatively narrow minor groove in B-DNA is related to the formation of an ordered water structure known as the "spine of hydration" which is expected to form most strongly in *polydA·polydT*. The very large entropy for the melting of $dCA_7G + dCT_7G$ is consistent with this speculation, since an ordered water structure would be associated with a low entropy. Moreover, the exceptionally low enthalpy observed for an uninterrupted run of A residues is consistent with the speculation of Breslauer et al (27) that the binding of netropsin and various intercalators induces an endothermic conformational change to 'standard' B-form in *polydA·polydT*. It is proposed (27) that this conformational transition requires an enthalpic input which offsets the favorable enthalpy of binding. Table II-1 also shows values for the standard free

Table II-3 Destabilization of Double Helices by Base-Base Mismatches or Wobble Base Pairs[a]

Mismatch/ Wobble	ΔG°, 25° (kcal mol^{-1})	ΔH° (kcal mol^{-1})	ΔS° (cal deg^{-1} mol^{-1})
-A-T-A- -T-G-T-	0	-8.1	-27
-A-G-A- -T-G-T-	0.2	-6.0	-20
-A-G-A- -T-A-T-	0.3	-5.1	-18
-A-G-A- -T-T-T-	0.7	0.8	1
-A-A-A- -T-G-T-	1.2	7.6	22
-A-C-A- -T-T-T-	1.2	-5.7	-23
-A-T-A- -T-C-T-	1.5	-2.5	-13
-A-A-A- -T-A-T-	1.5	10.6	31
-A-T-A- -T-T-T-	1.5	-7.1	-29
-A-C-A- -T-A-T-	1.9	7.2	19
-A-C-A- -T-C-T-	2.0	-7.8	-33
-A-A-A- -T-C-T-	2.1	11.7	32

[a]The values are obtained by subtracting nearest-neighbor contributions present in $dCA_5G \cdot dCT_5G$ from the data present in Table II-1. The values from a least squares fit to $dCA_5G \cdot dCT_5G$, $dCA_6G \cdot dCT_6G$, $dCA_7G \cdot dCT_7G$ are ΔG° = -6.5 kcal mol^{-1}, ΔH° = -47.5 kcal mol^{-1}, ΔS° = -138 cal deg^{-1} mol^{-1} for the nearest neighbor approximation to $dCA_5G \cdot CT_5G$.

energy at 50^0C using the equation

$$\Delta G^0(25^0C) - (T - 25^0C) \times \Delta S^0$$

Note that at higher temperatures the two $G{\cdot}C$ duplexes are more stable than both $A{\cdot}T$ duplexes. Again, this is consistent with the finding (34) that the anomolous bending of the $A_n^{\cdot}T_n$ sequence disappears at high temperature. If the conformation of $dA_n^{\cdot}dT_n$ changes as n increases beyond $n = 2$, the nearest neighbor parameters obtained here may be more relevant to $dA_n^{\cdot}dT_n (n \geq 5)$ than to sequences with only two or three neighboring $A{\cdot}T$ pairs.

C. Mismatched or wobble bases

Background; fidelity of replication

The base pair complementarity of DNA discovered by Watson and Crick insures the faithful replication and propagation of genetic information, however mispaired or noncomplementary bases are sometimes incorporated (1, 35). If allowed to propagate to the next generation a mismatched pair of bases incorporated during replication will lead to a mutation in the amino acid sequence of the protein product. Several enzymes are present in the cell which recognize and correct mistakes in replication, thus increasing the fidelity (36). Until recently nothing was known about the mechanisms for recognizing mistakes in replication, but it is now becoming possible to compare thermodynamic, structural, and genetic data (35). The relative frequency of mutations due to various mismatches can be affected by several factors, including the relative thermodynamic stability of the mismatches (which may or may not be similar in the cell and in aqueous solution), kinetic factors, and ability of repair enzymes to recognize specific mismatches. The relationship between thermodynamics and kinetics of mismatch formation is discussed in (35), while experimental data on kinetics are presented in (37). This section presents thermodynamic parameters for the twelve mismatched oligomers of the series $dCA_3XA_3G + dCT_3YT_3G$ in solution and discusses the

implications for understanding fidelity of replication and mechanisms of DNA repair. Since the publication of this data (see 15, which also includes earlier references), other thermodynamic (38, 39) and kinetic (37) studies have appeared which have reported results consistent with those presented here.

Methods and results

Melting experiments were performed and analyzed as described above. The mismatches and wobble base pairs are destabilizing to widely varying degrees relative to the perfect duplex (Table II-1). For example, the duplex with a $G{\cdot}T$ wobble has a T_m of 31^0C at $400\mu M$ strand concentration, whereas an $A{\cdot}C$ mismatch results in a T_m of 19^0C. The most stable non Watson-Crick base oppositions in the duplexes are $G{\cdot}T$, $G{\cdot}G$ and $G{\cdot}A$; the least stable are $T{\cdot}T$, $A{\cdot}A$, $A{\cdot}C$, $C{\cdot}C$ and $C{\cdot}T$. The thermodynamic contributions of all these base-base oppositions can be treated in a nearest-neighbor analysis as an internal loop of two bases regardless of whether or not the bases are hydrogen bonded. The results (given in Table II-3) indicate the wide range of stability of "an internal loop of two bases". The ΔG^0 values show that $G{\cdot}T$, $G{\cdot}G$ and $G{\cdot}A$ are the most stable; this is consistent with hydrogen bonding which has been reported in $G{\cdot}T$ and $G{\cdot}A$ (40-42) and suggest that $G{\cdot}G$ also forms hydrogen bonds. Sequence effects due to stacking are very important as shown by the slight destabilization caused by $\binom{-A-G-A-}{-T-A-T-}$, compared to the large stabilization caused by $\binom{-A-A-A-}{-T-G-T-}$ which posseses the same hydrogen bonding possibilities. The other base oppositions destabilize the helix with an unfavorable standard free energy at 25^0C relative to $CA_5G{\cdot}CT_5G$ varying from 1.0 kcal mol^{-1} to 1.9 kcal mol^{-1}. In general the most unfavorable base to have in a mismatch is C; G tends to be the least destabilizing. The order of stability is approximately $G{\cdot}T > G{\cdot}G > G{\cdot}A > C{\cdot}T > A{\cdot}A = T{\cdot}T > A{\cdot}C = C{\cdot}C$, but is dependent on the surrounding sequence. The order may be somewhat different in sequences which are not $dA_n{\cdot}dT_n$.

Comparison of relative stabilities of mismatches to relative rates of misincorporation and repair in vitro and in vivo

Several recent studies have made use of genetic assays for repair of site specific mutants to obtain qualitative data on relative repair efficiencies of mismatches *in vivo* (43-48) and *in vitro* (49-52). In vivo studies can examine relative mismatch repair efficiencies for post replicative proofreading systems (44-48). Typically a mutation is identified in a bacteriophage which produces a phenotypic marker, for example, the ability to lyse the host cell. A point mutation may produce a stop signal (i.e. a nonsense mutation) preventing synthesis of a protein that is critical to the lytic process (53) or affect binding strength of a protein "switch" whose binding induces lysis (43) by occuring within the recognition site. The mutated progeny are easily identified through their ability to lyse the bacterial host cell under conditions in which lysogeny (a state of what might be called peaceful coexistence between the phage and the host bacteria) ordinarily would occur or vice versa. DNA from mutant colonies is collected, denatured (47), (or the phage DNA is isolated during a single stranded stage in the phages' life cycle (43)) and each strand annealed to a wild type strand to form a mismatched duplex. The heteroduplex is then transfected into a host, and after several generations the progeny are assayed to determine the efficiency of repair (43).

By carrying out such experiments within repair deficient strains of *E. coli*, the relative mismatch repair efficiencies of several post replicative proofreading enzymes have been investigated (45-47). The products of genes *mut H*, *mut L*, and *mut S* are known to be involved in methylation directed mismatch repair in *E. coli* (54). Their repair efficiencies for mismatches do not correlate with thermodynamic stabilities. Radman and co-workers (47) report that *Mut L* repairs $G{\cdot}G, G{\cdot}T, A{\cdot}C, T{\cdot}T, A{\cdot}A$, and $G{\cdot}G$ with high efficiency, while $A{\cdot}G, C{\cdot}T$, and $C{\cdot}C$ were repaired with much lower efficiency. Moreover, some repair enzymes are probably specific for certain mismatches (46).

The repair efficiency is highly sequence and orientation dependent (45, 48). Similar findings have been reported for the *hex* repair system in *Streptococcus pneumoniae* (44).

Prokaryotic DNA polymerases also possess a $3' - 5'$ exonuclease proofreading activity (36). This activity in *T4* DNA polymerase has been tested with eight mismatches, including some with reversed orientation with respect to neighboring bases (51). The relative rates of efficiency of repair are $A{\cdot}A > A{\cdot}C > A{\cdot}G, G{\cdot}A, G{\cdot}G, A{\cdot}C, A{\cdot}C, T{\cdot}T >> T{\cdot}G$. Considering sequence effects these data are not inconsistent with a thermodynamically controlled mechanism, but the quantitative disagreement between relative repair efficiencies and thermodynamic stabilities is strong enough to merit skepticism. In *vitro* experiments can test for the relative rate of incorporation of mismatches by the polymerizing activity alone. This experiment has actually been performed using a eukaryotic polymerase (52), and the error rates ($G{\cdot}G > (G{\cdot}T + C{\cdot}T) > A{\cdot}G >> C{\cdot}A \geq G{\cdot}A \geq T{\cdot}T$) roughly correlate with thermodynamic stability. Similarly qualitative agreement is observed in *E. coli* replication fidelity (49, 50).

DNA ligase is involved in several stages of replication, especially on the lagging strand (36). Preliminary evidence suggests that restriction fragments with single mismatch containing cohesive ends are ligated with an efficiency roughly predicted by the thermodynamic stability of the mismatch (55) ignoring neighboring sequence effects.

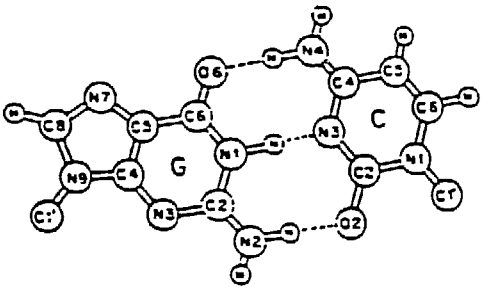
It is possible that the thermodynamic parameters reported here in aqueous solutions are very different from those present *in vivo* where DNA is highly compacted and complexed with proteins (35). However, the qualitative agreement between relative thermodynamic stabilities in aqueous solution and relative error rates in replication and ligation suggests that the hydrophobic environment of a protein may not dramatically affect the energetics of the DNA. With that qualification in mind, the overall picture that emerges is as follows. The processes of replication and repair of mismatches

are under the control of several enzymes, some perhaps not yet discovered. Some of the enzymes involved recognize mismatches merely by their thermal instability, their activity is thermodynamically controlled. Other enzymes (including all post replicative proofreading enzymes characterized so far) repair different mismatches with varying efficiencies and show strong sequence effects. It appears that some of these repair enzymes specialize in repairing specific mismatches or perhaps even mismatches at specific sites (45).

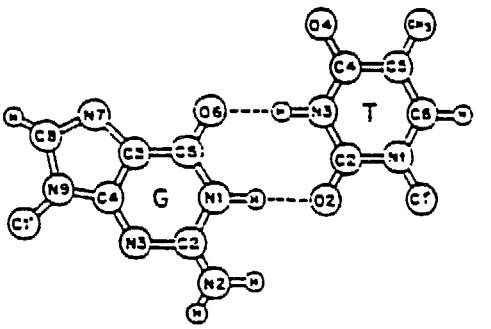
Possible structures of mismatches

The apparent ability of repair enzymes to recognize mismatches suggests that mismatches may have unusual structures which are recognized by the enzyme. Possible hydrogen bonding schemes for all base-base oppositions are shown in Figures *II* – 3 to *II* – 5. Watson-Crick base pairing is included to allow a qualitative comparison of the *C*1′ – *C*1′ distances and orientations of the different pairs. In recent years several studies have appeared proposing base pairing schemes for mismatches based on crystal structures (56-59) and NMR (40-42, 60-63). The most studied combination has been *G*·*A* (40, 42, 48, 56, 57). Two crystal structures have resulted in two different proposed base pairing schemes. While Priv'e et al (56) observed the *G*(*anti*)·*A*(*anti*) shown, Brown et al report (57) a *G*(*anti*)·*A*(*syn*) structure in a different sequence at somewhat lower resolution. The structure shown in figure *II* – 3 is also consistent with that proposed on the basis of NMR data (40, 42). However an NMR investigation by Fazakerly et al (48) found evidence for very different structures in two sequences. The neighboring sequence dependence observed for the structure of *G*·*A* mismatches is consistent with the orientation dependence observed in the thermodynamic measurements as well as the neighboring sequence dependence of repair efficiencies. The *G*·*T* structure shown has been well established by NMR (41, 63) and crystallography (58). The *G*(*anti*)·*G*(*syn*) base pair has not been seen to our knowledge, but the hydrogen bonding

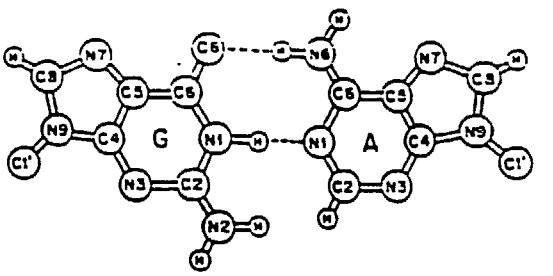
Figure II-3 Postulated base pairing schemes for base pairs involving guanine. The drawings are to scale; G·C is shown to indicate relative distances between C1' atoms and to give relative C1'–N bond orientations.



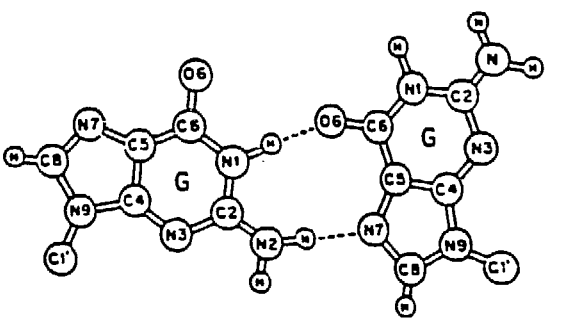
Watson Crick G (*anti*) · C (*anti*)



Wobble G (*anti*) · T (*anti*)



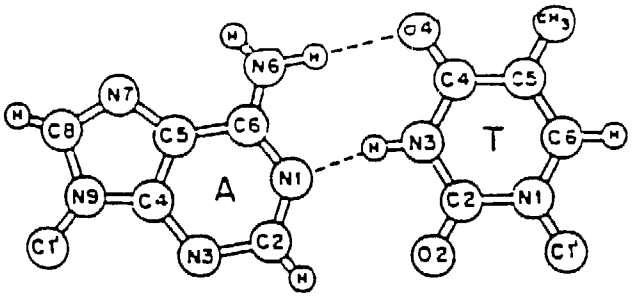
G (*anti*) · A (*anti*)



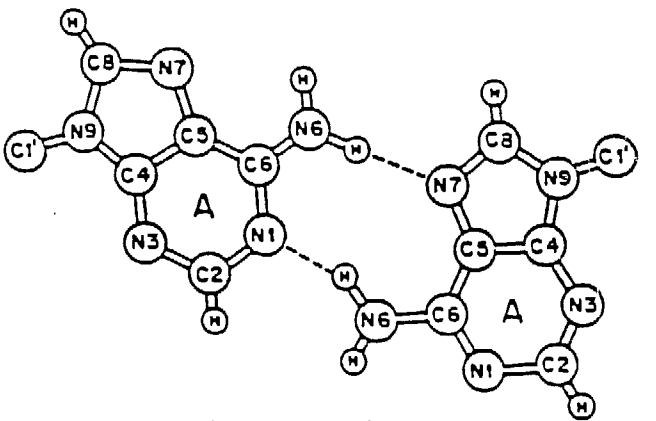
G (*anti*) · G (*syn*)

Figure II-4

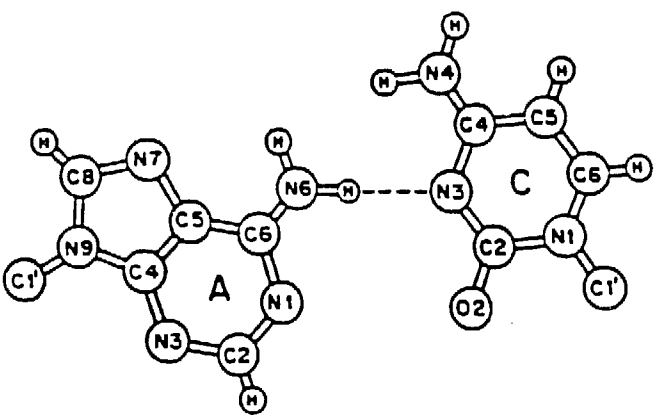
Postulated base pairing schemes for base pairs involving adenine.



Watson Crick A (*anti*) · T (*anti*)

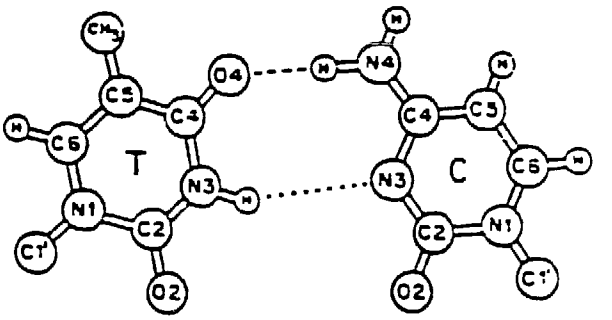


A (*anti*) · A (*anti*)

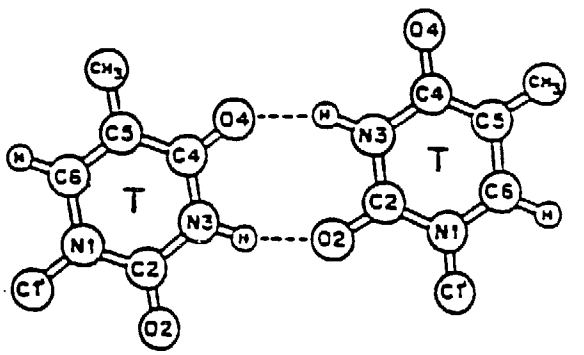


A (*anti*) · C (*anti*)

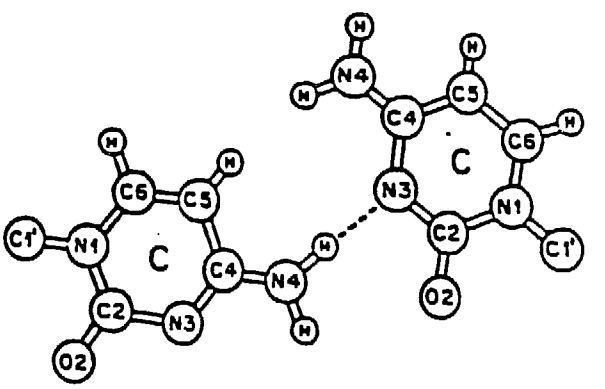
Figure II-5 Possible pyrimidine-pyrimidine base pairs. There is no experimental evidence for these pairs. The geometries are very different from Watson-Crick base pairs.



T(*anti*) · C(*anti*)



T(*anti*) · T(*anti*)



C(*anti*) · C(*anti*)

shown is postulated in a four-stranded *polyG* structure (64), in low pH polymer solutions (65), and evidence for some form of $G(anti)\cdot G(syn)$ is seen in telomeric DNA sequences (62). Another possibility is two hydrogen bonds of the type $N1-H\cdots O6$. In Fig. $II-4$ the $A\cdot C$ structure has been postulated by Patel et al. (60) and by Kollman (66), however a structure with two hydrogen bonds containing a protonated adenine has been observed by Hunter et al. (59) in a crystal. The $A(anti)\cdot A(anti)$ structure is speculation; another possibility is $A(anti)\cdot A(syn)$ with two $N6-H\cdots N1$ hydrogen bonds. However, Arnold et al (61) found that $A\cdot A$ does not form a stable H-bonded structure in dC_3AG_3. Figure $II-5$ contains the least stable base-base oppositions. Sugar-phosphate constraints may prevent any of these hydrogen bonds from forming, however, Kollman has postulated the $T\cdot C$ structure shown (66). Direct proof for any of these base pairs should come from NMR studies of the exchangeable protons to show hydrogen bonding, and NOE measurements to establish the *syn* or *anti* confromation.

The most plausible hydrogen bonding base pairs usually can be drawn for the base-base oppositions which are thermodynamically the most stable (ΔG^0 most negative). The data also show that the neighboring sequence is an important factor in stability. It is well known that base stacking is necessary for duplex stability. However, hydrogen bonding may be necessary to allow the bases close enough to each other inside the helix where they can stack.

It should also be realized that the free energies of different base pairs vary differently with temperature, so their relative stabilities at a typical hybridization temperature (see section D) may be quite different from those at 25^0C. The standard free energy at any temperature can be calculated from that at 25^0C using the standard entropy.

$$\Delta G^0(25^0C) - (T - 25^0C) \times \Delta S^0$$

In Table II-1 this equation has been used to provide values of $\Delta G^0(50^0C)$. Note that for the ΔG^0 values at the higher temperatures the sequence dependence of the various

base oppositions nearly vanish. The two $G{\cdot}C$ duplexes are more stable than the $A{\cdot}T$ duplexes. However, one should keep in mind that extrapolation to higher temperatures are least accurate for the least stable duplexes.

As our understanding of the conformations of the mismatches and of their effect on the thermodynamics improves, it may not be necessary to measure every possible nearest and next nearest neighbor sequence to predict the results. Evidence to test the hydrogen bonding schemes shown in Figs. $II-3$ to $II-6$ will be most helpful. It may be some time before quantitative comparison can be made with repair efficiencies, however qualitative comparisons already can identify enzymes with specialized activities.

D. Base Pairing Involving Deoxyinosine

Background; inosine and probe design

Hypoxanthine, the base found in the nucleosides inosine and deoxyinosine, behaves approximately as a guanine analog in nucleic acids. Inosine occurs naturally in the wobble position of the anticodon loop of some transfer RNA's, where it appears to pair with adenosine in addition to cytidine and uridine, the nucleosides which pair with guanosine in that position. Early studies on physical characterizations of inosine containing polynucleotides have been briefly reviewed (67).

Knowledge of the base-pairing energies of deoxyinosine with the four normal bases is of use in the design of oligonucleotide probes. If dI pairs with less specificity than the normal four bases, it could be placed at those positions in the probe where the base in the gene being sought is unknown. Such ambiguities arise when the genomic sequence is not known, but is being deduced from a known peptide sequence; the genomic sequence is ambiguous at positions where the genetic code is redundant. See (67, and references cited therein) for a discussion of various strategies for designing probes.

More information on stabilities of mismatches is needed to improve probe design. It would be most useful to find base analogs which would be less discriminatory in base pairing than the normal four bases, so unique sequence probes of greater and more predictable stability could be designed for gene isolation, or at least so that a smaller number of sequences in mixtures of probes would suffice. Inosine, because it seems to pair less strongly with C and more strongly with A than guanosine does, is a candidate for this purpose.

We have measured the stabilities of a set of deoxyoligonucleotide duplexes containing each of the four normal DNA bases paired with deoxyinosine. The contributions of matched and mismatched deoxyinosine base pairs to duplex stability have been calculated. Comparison to results obtained with similar duplexes containing only normal bases (described in sections A and B) allows evaluation of deoxyinosine and its possible utility in probes at positions of base ambiguity.

Methods and Results

Synthesis of deoxyinosine containing molecules is described in (67). Melting curves were obtained and analyzed as described above.

Thermodynamic parameters for the helix-coil transition of all nine deoxyinosine-containing duplexes were calculated from the absorbance curves as in sections B and C and are shown in Table II-4. In Table II-5 nearest neighbor contributions to double strand formation are listed for $dI{\cdot}dC$ pairs, whereas in Table II-6 nearest neighbor contributions have been calculated for the mismatched duplexes, treating the mismatched base pair as a two base internal loop, in accordance with the convention established in section B. In terms of nearest neighbor interactions, the duplexes may be considered as equivalent to $dCA_5G{\cdot}dCT_5G$ plus either two additional base pair stacking interactions, $\binom{AX}{TY}$ + $\binom{XA}{YT}$, or a two base internal loop, $\binom{A-X-A}{T-Y-T}$. We have chosen the former treatment for $dI{\cdot}dC$ pairs (matched base pairs) and have treated the other

dI oppositions as internal loops (mismatched bases), in order to maintain consistency with earlier treatments. The distinction between matched pairs and mismatched bases is merely formal and is summarized in sections B and C.

Table II-4 Van't Hoff Thermodynamic Values for Double Helix Formation of Deoxyinosine Containing $dCA_3XA_3G+dCT_3YT_3G$ in 1 M NaCl, pH7.

X·Y	ΔG°, 25° C (kcal mol^{-1})[a]	ΔH° (kcal mol^{-1})[b]	ΔS° (cal $deg^{-1}mol^{-1}$)[c]	T_m(°C)[d] C_T=400μM
I·C	-8.8	-66	-191	41°
C·I	-8.1	-58	-168	39°
-·-[e]	-7.8	-59	-172	37°
I·A	-7.5	-63	-186	35°
I·G	-6.3	-57	-168	30°
A·I	-6.3	-48	-141	30°
I·T	-5.9	-58	-176	27°
T·I	-5.8	-50	-147	27°
G·I	-5.7	-52	-154	26°
I·I	-5.7	-47	-140	27°

[a]Estimated precision in ΔG° is ± 0.1 kcal mol^{-1}

[b]Estimated precision in ΔH° is ± 0.3 kcal mol^{-1}

[c]Estimated precision in ΔS° is ± 9 cal deg^{-1} mol^{-1}

[d]Estimated precision in T_m is ± 1°

[e]dCA_6G ·dCT_6G. Data from Morden et al., (1983) Biochemistry 21, 428-436.

Table II-5 Neareast-neighbor Contributions of Deoxyinosine to Double Strand Formation in 1 M NaCl, pH 7.[a]

Nearest Neighbor	ΔG°, 25 °C (kcal mol^{-1})	ΔH° (kcal mol^{-1})	ΔS° (cal $deg^{-1} mol^{-1}$)
$\frac{1}{2}\left(\begin{smallmatrix}-A-I-\\-T-C-\end{smallmatrix} + \begin{smallmatrix}-I-A-\\-C-T-\end{smallmatrix}\right)$	-1.1 ± 0.2	-9.3	-27
$\frac{1}{2}\left(\begin{smallmatrix}-A-C-\\-T-I-\end{smallmatrix} + \begin{smallmatrix}-C-A-\\-I-T-\end{smallmatrix}\right)$	-0.8 ± 0.2	-5.3	-15

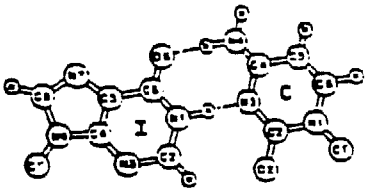
[a] The values given are for the reaction $\begin{smallmatrix}-W_1 & W_2-\\ -C_1 & C_2-\end{smallmatrix} \rightarrow -W_1-W_2- + -C_1-C_2-$

Table II-6 Destabilization of Double Helices by Deoxyinosine in Base-Base Mismatches or Wobble Base Pairs.[a]

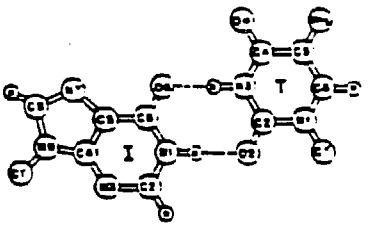
Mismatch/ Wobble	ΔG°, 25°C (kcal mol^{-1})	ΔH° (kcal mol^{-1})	ΔS° (cal deg^{-1} mol^{-1})
-A-I-A- / -T-A-T-	-1.0	-15.5	-48
-A-A-A- / -T-I-T-	+0.2	-0.5	-3
-A-I-A- / -T-G-T-	+0.2	-9.5	-30
-A-I-A- / -T-T-T-	+0.6	-10.5	-38
-A-T-A- / -T-I-T-	+0.7	-2.5	-9
-A-G-A- / -T-I-T-	+0.8	-4.5	-16
-A-I-A- / -T-I-T-	+0.8	0.5	-2

[a]The values are obtained by subtracting nearest-neighbor contributions present in $dCA_5G \cdot dCT_5G$ from the data given in Table II-4.

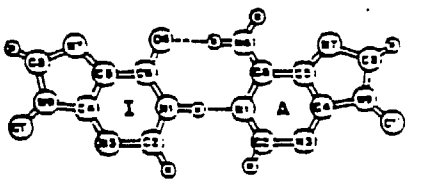
Figure II-6 Postulated base-pairing schemes for pairs containing deoxyinosine. The pairs I:C, I:T and I:A are exact analogs of G:C, G:T and G:A.



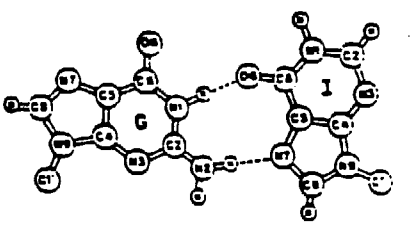
Watson-Crick I (*anti*) · C (*anti*)



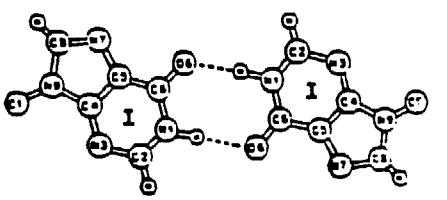
Wobble I (*anti*) · T (*anti*)



I (*anti*) · A (*anti*)



G (*anti*) · I (*syn*)



I (*anti*) · I (*anti*)

Discussion

The results reported in Table II-4 give information on the stability of base pairs containing deoxyinosine matched with each of the four normal bases, each in two orientations in the duplexes $dCA_3XA_3G{\cdot}dCT_3YT_3G$. The oligonucleotide duplex lacking the central $X{\cdot}Y$ pair, $dCA_6G{\cdot}dCT_6G$, is included for comparison; an "inert" $X{\cdot}Y$ base pair would give a duplex with the same stability as $dCA_6G{\cdot}dCT_6G$. As expected, $I{\cdot}C$ pairs are more stable than other I-containing pairs, but they contribute less stability than standard Watson-Crick pairs. The $I{\cdot}C$ pair is less stable in the duplex than $G{\cdot}C$ by an average of +1.4 kcal mol^{-1} in standard free energy at 25^0C and less stable than $A{\cdot}T$ by an average of +0.6 kcal mol^{-1} (see section B). Insertion of an $I{\cdot}A$ pair into the middle of a $dCA_6G{\cdot}dCT_6G$ duplex is only slightly destabilizing in one orientation, but the other mismatches are more strongly destabilizing. The decrease in stability from $dCA_6G{\cdot}dCT_6G$ ranges from +1.5 kcal mol^{-1} ($X{\cdot}Y = I{\cdot}G$, $A{\cdot}I$) to +2.1 kcal mol^{-1} ($X{\cdot}Y = G{\cdot}I$, $I{\cdot}I$) in standard free energy at 25^0C. Mismatches in the same sequence position not involving I are usually more destabilizing (section B); for example, a $C{\cdot}C$ or $A{\cdot}C$ mismatch reduces the stability by +3.3 kcal mol^{-1} in free energy at 25^0C. ΔS^0 values may be used to extrapolate the ΔG^0 values to higher temperature, as in section C. Most hybridization probe experiments are performed near 50^0C.

The relative stabilities of the various base oppositions depend on the number of hydrogen bonds that can be formed within the constraints of the glycosidic bonds, and on the stacking interactions with the neighboring bases. Possible hydrogen bonding schemes with inosine and the four standard bases are shown in Fig. $II-6$. We note that two hydrogen bonds per base opposition can be drawn for all pairs. This is in contrast to $A{\cdot}C$ and $C{\cdot}C$ for which at most one hydrogen bond is reasonable. The first three base pairs in Fig. $II-6$ are precise analogs of established G-containing pairs (Watson-Crick $G{\cdot}C$, wobble $G{\cdot}T$ and $G{\cdot}A$) (see section B). The $G{\cdot}I$ pair is drawn by

analogy with the *G·G* pairing proposed for a quadruple-stranded poly *rG* structure (see section B). The *I·I* pair is the most distorted from B-DNA geometry and is the least stable thermodynamically.

Thus far, to our knowledge, only the *I·A* base pair has been examined through NMR and crystallographic techniques. The structure shown in Fig. *II* – 6 has been observed in solution (68). In contrast, an *I·A* pairing involving a *syn A* has been reported in a crystal structure (69).

The implications of the results reported in this section for the design of oligonucleotide probes are addressed in (67) along with the sequence and orientation effects observed in Table II-4.

E. Effects of Physiological Salt Conditions on Thermal Stability

Background

It is well known that the concentration of ions in solution has a significant effect on the thermal stability of nucleic acid duplexes in solution (70). Most studies of thermodynamic stability of oligonucleotide duplexes in solution including those described above have been done in 1*M* NaCl (12-18, 67, 71). At that salt concentration, duplex formation is maximally stabilized with respect to hairpins or single strands, facilitating the measurement of parameters for duplex melting, but these parameters are of course valid only at 1*M* NaCl. Extrapolation to other NaCl concentrations is simple (70), but the effect of varying ions is not well characterized. It is therefore of interest to determine how the relative contributions to thermodynamic parameters from nearest neighbor pairs differs at 1*M* NaCl and under more "physiological" (72) salt conditions. Here we present thermodynamic measurements in "physiological" salt conditions.

Table II-7

Thermodynamic parameters for duplex formation for $dCA_3XA_3G+dCT_3YT_3G$ in buffer 1 containing standard salt conditions for the studies in sections A-C (1 M NaCl, 10 mM PO_4, 0.1 mM EDTA pH=7) and in buffer 2 containing more "physiological" salt conditions (150 mM KCl, 30 mM $MgCl_2$, 10 mM PO_4, 0.1 mM EDTA, pH=7)

X·Y	buffer 1	buffer 2	buffer 1	buffer 2	buffer 1	buffer 2
	ΔH^0	ΔH^0	ΔS^0	ΔS^0	ΔG^0	ΔG^0
C·G	-65	-62	-183	-178	-10.1	-8.6
A·T	-68	-69	-196	-205	-9.6	-8.2
T·A	-59	-61	-168	-181	-8.5	-7.2

Results and discussion

Thermodynamic parameters were measured as above for $dCA_3XA_3G + dCT_3YT_3G$ with the three Watson-Crick combinations of X and Y shown in Table II-7. The buffer contained $150mM$ KCl, $30mM$ $MgCl_2$, $0.1mM$ EDTA $10mM$ PO_4, pH= 7. In all three cases ΔG^0 at 25^0C is exactly 85% of that measured in $1M$ NaCl. Thus, barring an extraordinary coincidence, it is safe to conclude that relative free energies measured in $1M$ NaCl are applicable to physiological salt conditions for duplex DNA oligonucleotides. No clear trend is observed in ΔH^0 and ΔS^0 values. One caution must be added to the interpretation of this data; true "physiological conditions" imply more than a specific salt concentration, in fact the interaction of DNA with proteins within the cell may change relative thermodynamic stabilities dramatically. The significance of the results reported in Table IV-1 lies in the fact that under ordinary circumstances, relative free energies for duplex formation are not shifted by unusual *salt* conditions in the cell.

References

1. Lomant, A. J. and Fresco, J. R. (1975) in progress in Nucleic Acid Research and Molecular Biology, eds., Vol. 15, pp. 185 – 218, Academic Press, New York.

2. Dickerson, R. E. and Drew, H. R. (1981) J. Mol. Biol. 149, 761 – 786.

3. von Hippel, P. E., Bear, D. G., Morgan, W. D. and Mcswiggen, J. A. (1984) Ann. Rev. Biochem. 53, 389 – 446.

4. Gotoh, O. (1983) Adv. Biophys. 16, 1 – 52.

5. Gamper, H. B. and Hearst, J. E. (1982) Cell, 29, 81 – 90.

6. Morgan, W. D., Bear, D. G. and von Hippel, P. H. (1983) J. Biol. Chem., 258, 9565 – 9574.

7. Crothers, D. M. and Fried, M. (1983) Cold Spring Harbor Symp. Quant. Biol. 47, 263 – 269.

8. Borowiec, J. A., Zhang, L., Sasse-Dwight, S. and Gralla, J. D. (1987) J. Mol. Biol. 196, 101 – 111.

9. Ptashne, M. (1986) Nature 332, 697.

10. Koo, H-S., Wu, H-M. and Crothers, D. M. (1986) Nature 320, 501 – 506.

11. Vologodskii, A. V., Amirikyan, B. R., Lyubchenko, Y. L. and Frank - Kamenetskii (1984) J. Biomol. Struc. Dynamics 2, 131 – 148.

12. Breslauer, K. J., Frank, R., Blocker, H. and Marky, L. A. (1986) Proc. Natl. Acad. Sci., USA 83, 3746 – 3750.

13. Klump, H. (1985) Thermochimica Acta 85, 457 – 463.

14. Frier, S. M. et al. (1986) Proc. Natl. Acad. Sci., USA 83, 9373 – 9376.

15. Aboul-ela, F., Koh, D., Martin, F. H. and Tinoco, I., Jr. (1985) Nuc. Acids Res. 13, 4811 – 4824.

16. Borer, P. N., Dengler, B., Tinoco, I., Jr. and Uhlenbeck, O. C. (1974) J. Mol. Biol. 86, 843 – 853.

17. Martin, F. H., Uhlenbeck, O. C. and Doty, P. (1971) J. Mol. Biol. 57, 201 – 215.

18. Nelson, J. W., Martin, F. H. and Tinoco, I., Jr. (1981) Biopolymers 20, 2509 – 2531.

19. Beaucage, S. L. and Caruthers, M. H. (1981) Tetrahedron Lett. 22, 1859 – 1862.

20. Morden, K. M., Chu, Y. G., Martin, F. H. and Tinoco, I., Jr. (1983) Biochemistry 21, 428 – 436.

21. Wilson, W. D., Zuo, E. T., Jones, R. L., Zon, G. L. and Baumstark, B. R. (1987) Nuc. Acids Res. 15, 105 – 118.

22. Peck, L. J. and Wang, J. C. (1981) Nature 292, 375 – 377.

23. Rhodes, D. and Klug, A. (1981) Nature 292, 378 – 380.

24. Strauss, F., Gaillard, C. and Prunell, A. (1981) Eur. J. Biochem. 118, 215 – 222.

25. Wu, H-M. and Crothers, D. M. (1984) Nature 308, 509 – 513.

26. Wilson, W. D., Wang, Y-H., Krishnamoorthy, C. R. and Smith, J. C. (1985) Biochemistry 24, 3991 – 3999.

27. Breslauer, K. J. et al. (1987) Proc. Natl. Acad. Sci., USA (in press).

28. Prunell, A. (1982) EMBO 1, 173 – 179.

29. Pilet, J., Blichorski, J. and Brahms, J. (1975) Biochemistry 14, 1869 – 1876.

30. Diekmann, S. and Zarling, D. A. (1985) Nuc. Acids Res.

31. Alexeev, D. G., Lipanov, A. A. and Skuratovskii, I. Ya. (1987) Nature 325, 821 – 823.

32. Lipanov, A. A. and Chuprina, V. P. (1987) Nuc. Acids Res. 15, 5835 – 5844.

33. Dickerson, R. E., Drew, H. R., Conner, B. N., Wing, R. M., Fratini, A. V. and Kopka, M. L. (1982) Science 216, 475 – 485.

34. Diekmann, S. and Wang, J. C. (1985) J. Mol. Biol. 186, 1 – 11.

35. Tinoco, I., Jr., Wolk, S., Arnold, F. and Aboul-ela, F. (1987) in Structure and Dynamics Of Biopolymers, C. Nicolini, ed. (Plenum Press) 92 – 111.

36. Kornberg, A. (1980) DNA Replication, W. H. Freeman and Company, San Francisco.

37. Ikuta, S., Takagi, K., Wallace, R. B. and Itakura, K. (1987) Nuc. Acids Res. 15, 797 – 811.

38. Kawase, Y., Iwai, S., Inoue, H., Miura, K. and Ohtsuka, E. (1986) Nuc. Acids Res. 14, 7727 – 7736.

39. Gaffney, B. L. and Jones, R. A. (1987) Biophysical Journal 51, 497a.

40. Patel, D. J., Kozlowskii, S. A., Ikuta, S. and Itakura, K. (1984) Biochemistry 23, 3207 – 3217.

41. Hare, D., Shapiro, L. and Patel, D. J. (1986) Biochemistry 25, 7445 – 7456.

42. Kan, L., Chandresegaran, S., Pulford, S. M. and Miller, P. S. (1983) Proc. Natl. Acad. Sci., USA 80, 4263 – 4265.

43. Ferscht, A. R. and Knill-Jones, J. W. (1983) J. Mol. Biol. 165, 633 – 654.

44. Lacks, S. A., Dunn, J. J. and Greenberg, B. (1982) Cell 31, 327 – 336.

45. Schaaper, R. M. and Dunn, R. C. (1987) Proc. Natl. Acad. Sci., USA 84, 6220 – 6224.

46. Jones, M., Wagner, R. and Radman, M. (1987) J. Mol. Biol. 194, 155 – 159.

47. Dohet, C., Wagner, R. and Radman, M. (1985) Proc. Natl. Acad. Sci., USA 82, 503 – 505.

48. Fazakerly, G. V., Quignard, E., Guschlbauer, W., van der Marel, G. A., van Boom, J. H., Jones, M. and Radman, M. (1986) EMBO J. 5, 3697 – 3703.

49. Ferscht, A. R. (1979) Proc. Natl. Acad. Sci., USA 76, 4946 – 4950.

50. Ferscht, A. R. and Knill-Jones, J. W. (1981) Proc. Natl. Acad. Sci., USA 78, 4251 – 4255.

51. Sinha, N. K. (1987) Proc. Natl. Acad. Sci., USA 84, 915 – 919.

52. Rienitz, A., Grosse, F., Blocker, H., Frank, R. and Krauss, G. (1985) Nuc. Acids Res. 13, 5685 – 5695.

53. Skopek, T. R. and Hutchinson, F. (1982) J. Mol. Biol. 159, 19 – 33.

54. Lahue, R. S., Su, S-S. and Modrich, P. (1987) Proc. Natl. Acad. Sci., USA 84, 1482 – 1486.

55. Wiadekiewicz, R. and Ruiz-Carillo, A. (1987) Nuc. Acids Res. 15, 7831–7848.

56. Priv'e, G. G., Heinemann, U., Chandresegaran, S., Kan, L-S., Kopka, M. L. and Dickerson, R. E. (1987) Science 238, 498 – 504.

57. Brown, T., Hunter, W. N., Kneale, G. and Kennard, O. (1986) Proc. Natl. Acad. Sci., USA 83, 2402 – 2406.

58. Hunter, W. N., Kneale, G., Brown, T, Rabinovich, D. and Kennard, O. (1986) J. Mol. Biol. 190, 605 – 618.

59. Hunter, W. N., Brown, T. and Kennard, O. (1987) Nuc. Acids Res. 15,

6589 – 6606.

60. Patel, D. J., Kozlowski, S. A., Ikuta, S., and Itakura, K. (1984) Biochemistry 23, 3218 – 3226.

61. Arnold, F. H., Wolk, S., Cruz, P. and Tinoco, I., Jr. (1987) Biochemistry 26, 4068 – 4075.

62. Henderson, E., Hardin, C. C., Wolk, S. K., Tinoco, I., Jr. and Blackburn, E. H. (1987) Cell (in press).

63. Salisbury, S. A. and Anand, N. N. (1985) J. Chem. Soc. Chem. Commun. 10, 985 – 986.

64. Saisekharan, V., Zimmerman, S. and Davis, D. R. (1975) J. Mol. Biol. 92, 171 – 179.

65. Antao, V. P., Gray, C. W., Gray, D. M. and Ratliff, R. C. (1986) Nuc. Acids Res. 14, 10091 – 10112.

66. Keepers, J. W., Schmidt, P., James, T. L. and Kollman, P. A. (1984) Biopolymers 23, 2901 – 2929.

67. Martin, F. H., Castro, M. M., Aboul-ela, F. and Tinoco, I., Jr. (1985) Nuc. Acids Res. 13, 8927 – 8938.

68. Uesugi, S., Oda, Y., Ikehara, M., Kawase, Y. and Ohtsuka, E. (1987) J. Biol. Chem. 262, 6965 – 6968.

69. Corfield, P. W. R., Hunter, W. N., Brown, T., Robinson, P. and Kennard, O. (1987) 15, 7935 – 7949.

70. Record, M. T., Anderson, C. F., and Lohman, T. M. (1978) Quarterly Rev. Biophys. 11, 103 – 178.

71. Nelson, J. (1982) Ph. D. thesis, University of California, Berkeley.

72. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1983) Molecular Biology of the Cell, Garland Publishing, Inc., New York, P. 286.

Chapter III

A. Background

Poly[$d(C-G)$] was first found to undergo a conformational transition in high NaCl concentration solutions in 1972 (1). With the crystallization of $d(C-G)_3$ (2) as Z-DNA in 1979 it became known that alternating (CG) DNA sequences can form the left handed structure. Since then many reports have appeared dealing with the physical characterization of Z-DNA (reviewed in 3, 4) and with possible biological roles (4-7) for Z-DNA. Most attention has been focused on defining the structure of Z-DNA (2-4), establishing the conditions for its formation in various systems (3, 4, 8-11) and measuring thermodynamic and kinetic parameters for the transition (3, 8, 10). The underlying mechanism(s) (i.e. the nature of the interface, kinetic intermediates, the role of substituents and the cooperativity) are not well understood.

The studies reported in this chapter were initiated with the goal of investigating a thermodynamic property, the cooperativity, for the $B \rightleftharpoons Z$ transition in *poly*[$d(^{5me}C-G)$]. The cooperativity is a measure of the degree to which a given step in the transition increases the rate or the probability of further steps in the transition. It can be characterized by a parameter, N_0, the cooperative unit, which is defined as the average length of a B or Z tract in an infinite length polymer under conditions in which the polymer is in a "fifty-fifty" B, Z equilibrium. The cooperative unit is one of the parameters involved in the analysis of experiments in which intercalating drugs are used to induce a transition between B and Z forms (12, 13). The cooperativity of the $B \rightleftharpoons Z$ transition may also play a role in the process of genetic recombination. In the early stages of recombination, it is believed that strands from two double helical DNAs must join to form a Holliday junction (14), in which the topological linking number (the number of times two strands intersect each other in their two dimensional projection) of the two strands must be zero. Such a situation may lead to stretches of thousands of base pairs

in which half must be left handed and half must be right handed. If the left handed tracts are Z-form (the only known form of left handed DNA), then their average length will be a function of the cooperative length. The cooperativity also reveals information about the junction. If the cooperative length is found to change with temperature, that would be an indication that the unfavorable free energy of forming a junction has an enthalpic component, possibly due to the unstacking of bases. It is not known whether or not the formation of a junction requires or favors a local opening of the double helix (15).

The cooperativity for the $B \rightleftharpoons Z$ transition has been investigated for several $B-Z$ systems using calorimetric methods (16, 17) and salt titration curves (3, 18, 19). This chapter presents theoretical predictions for the length dependence of $B \rightleftharpoons Z$ transition curves for various values of N_0, the cooperative unit, and compares these predictions with data for $poly[d(^{5me}C-G)]$ in 0.35 mM $MgCl_2$, 50 mM NaCl, 5 mM tris pH 8. A value of 1200 ± 400 base pairs is estimated for the cooperative unit, and a model is presented to explain a previously unexpected correlation between measured van't Hoff enthalpies and kinetic lifetimes.

B. Theory

The experimentally accessible parameter here is the van't Hoff enthalpy obtained from the slope of lnK_{eq} versus $1/T$ as in chapter II. In order to derive predictions for the behavior of the van't Hoff enthalpy as a function of chain length some one-dimensional Ising models will be considered. The following discussion is derived from (20) and (21), the most general treatment of the problem is (22).

Ising models and the matrix method

The first Ising model was formulated in an attempt to develop the simplest conceivable model for ferromagnetic phase transitions (23, 24). A one-dimensional system was assumed with only nearest neighbor interactions. This system cannot undergo

a phase transition, and for some time it was believed that this was true for all one-dimensional systems (24). However, it has since been shown that a phase transition does occur in a infinite one-dimensional Ising system with a long range potential (25, 26). Poland and Scheraga (26) point out that though this system is physically unrealisitic, strictly speaking the same is true of the liquid-solid and liquid-gas phase transitions observed in three dimensions. In both cases the same requirement exists for the discontinuities in thermodynamic functions (and their derivatives) which define a phase transition; the system must be treated in the thermodynamic limit, i.e. it must be considered as an ensemble of an infinite number of components. In other words, neither system undergoes a true phase transition, even though an ice-water system, which contains on the order 10^{23} components, comes much closer than a helix-coil or $B \rightleftharpoons Z$ transition in a biopolymer. Therefore an Ising model approach will be capable of predicting the phase transition-like behavior observed in biopolymers.

In order to model $poly[d(^{5me}C - G)]$ as an Ising system, the polymer is represented as consisting of individual units (i.e. base pairs) each of which can exist in either the B or Z configuration. An equilibrium constant, S, is defined for the conversion of a single base pair at a boundary from B form to Z form.

$$...BBBZZ... \rightleftharpoons ...BBZZZ...$$

A cooperativity parameter, σ, is defined such that σS is the equilibrium constant for the nucleation reaction

$$...BBBBB... \rightleftharpoons ...BBZBB...$$

then σ is the equilibrium constant for the reverse nucleation reaction

$$...ZZZZZ... \rightleftharpoons ...ZZBZZ...$$

σ describes the extent of cooperativity of the system-it is related to the unfavorable energy for forming a boundary between B and Z forms. It can be shown that σ is related to the cooperative unit (20)

$$N_0 = 1 + \sigma^{-\frac{1}{2}}$$

All of the relevant thermodynamic parameters can be derived from the partition function

$$Q = \sum_i e^{\frac{-E_i}{RT}}$$

where the summation is over all accessible states of the system, and E_i is the energy of the i^{th} state. The power of using the partition function resides in the fact that any average quantity, $\overline{A}$ can be written as

$$\overline{A} = \frac{\sum_i A_i e^{\frac{-E_i}{RT}}}{\sum_i e^{\frac{-E_i}{RT}}} = \frac{\sum_i A_i e^{\frac{-E_i}{RT}}}{Q}$$

since the probability of the i^{th} state is proportional to $e^{\frac{-E_i}{RT}}$. Sometimes the term in the numerator can also be expressed as a function of Q (see below).

Now consider the partition function for a $B \rightleftharpoons Z$ transition in a system with, say, 2 residues, so that the system has $2^2 = 4$ possible states as illustrated in Figure $III-1$ (in general the number of states with N residues is 2^N, though some of these states will be identical due to symmetry)

$$Q(2) = S^2 + 2\sigma^{\frac{1}{2}}S + 1$$

The system with three residues has eight possible states

$$\begin{aligned} Q(3) &= S^3 + 2\sigma^{\frac{1}{2}}S^2 + \sigma S + \sigma S^2 + 2\sigma^{\frac{1}{2}}S + 1 \\ &= (1, S)\begin{pmatrix} 1 & \sigma^{\frac{1}{2}}S \\ \sigma^{\frac{1}{2}} & S \end{pmatrix}\begin{pmatrix} 1+\sigma^{\frac{1}{2}}S \\ \sigma^{\frac{1}{2}}+S \end{pmatrix} \\ &= (1, S)M^2\begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{aligned}$$

in general

$$Q(N) = (1, S)M^{N-1}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

This manner of generating Q is known as the matrix method. The physical meaning of M

$$M = \begin{array}{c} \\ B \\ Z \end{array}\begin{array}{c} \begin{array}{cc} B & Z \end{array} \\ \begin{pmatrix} 1 & \sigma^{\frac{1}{2}}S \\ \sigma^{\frac{1}{2}} & S \end{pmatrix} \end{array}$$

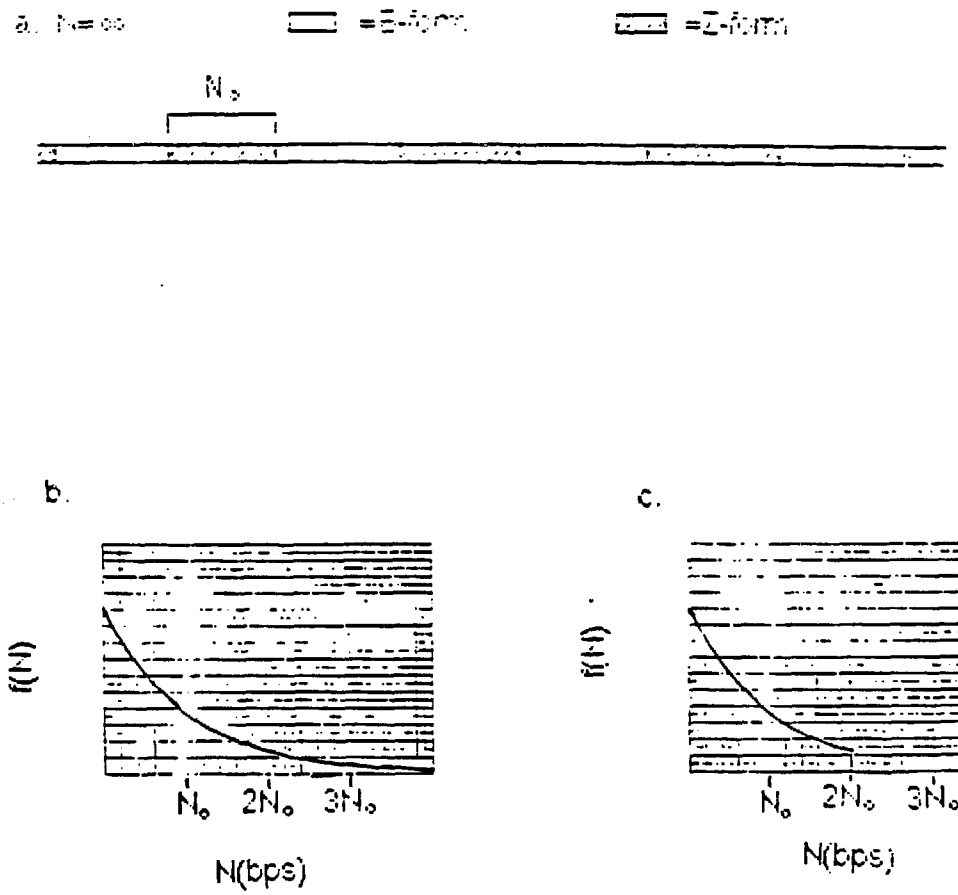
Figure III-1

Denumeration of states and their statistical weights ( ) for a *B-Z* system with (a) 2 and (b) 3 residues.

(a)	*BB*	*ZB*	*BZ*	*ZZ*
	(1)	$(\sigma^{1/2}S)$	$(\sigma^{1/2}S)$	(S^2)

(b)	*BBB*	*BBZ*	*ZBB*	*ZZZ*
	(1)	$(\sigma^{1/2}S)$	$(\sigma^{1/2}S)$	S^3
	BZB	*ZBZ*	*ZZB*	*BZZ*
	(σS)	(σS^2)	$(\sigma^{1/2}S^2)$	$(\sigma^{1/2}S^2)$

Figure III-2. Illustration of the distinction between the 'actual' cooperative unit, N_o, and the 'effective' cooperative unit, $\overline{N}$. (a) N_0 is the average length of a B or Z tract for an infinite length polymer under fifty-fifty equilibrium conditions. (b) distribution function for the fraction of B or Z tracts of length N for an infinite length polymer, $f(N)=\sigma^{1/2}/(1+\sigma^{1/2})^N$ (26). The mean is at N_o (c) distribution function for a polymer of length $2N_o$. Since the upper part of the distribution ($>2N_o$) is cutoff, the mean average length of a B or Z tract, $\overline{N}$, is less than the cooperative unit, N_o. For an order -disorder transition, the distribution function below N also changes (25).



is as follows: the n^{th} multiplication by the matrix M represents the addition of the n^{th} residue to the chain. If the $(n-1)^{th}$ residue was Z-form, a statistical weight of s is generated if the n^{th} residue is Z form or a statistical weight of $\sigma^{\frac{1}{2}}$ if the n^{th} residue is B. If the $(n-1)^{th}$ residue is B, a statistical weight of 1 is added for a B form n^{th} residue, otherwise the statistical weight is $\sigma^{\frac{1}{2}}s$. For a given N, the partition function may be calculated numerically by computer or the matrix M may be diagonalized by a similarity transformation, T

$$\Lambda = T^{-1}MT = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

λ_1 and λ_2 are the eigenvalues of M; the matrix T is constructed from the eigenvectors of M by standard methods (27). The authors of (27) derive

$$Q(N) = (1, \sigma s)(T\Lambda^{N-1}T^{-1}) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

based on the Zimm-Bragg model (21) for the helix to coil transition in proteins, where it is assumed that end residues favor the coil form. The equivalent expression for the case in which end residues do not favor either the B or Z forms is

$$Q(N) = (1, s)(T\Lambda^{N-1}T^{-1}) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The matrix M used here differs slightly from the one used by Zimm-Bragg

$$M(Zimm - Bragg) = \begin{pmatrix} 1 & \sigma s \\ 1 & s \end{pmatrix}$$

However, the eigenvalues of the two matrices are identical, the difference between the two expressions for Q lies in the pre-multiplication by a row vector. In the Zimm-Bragg case a factor of σ is introduced if the end residue is helical, reflecting the fact that the polymer ends tend to be coiled or disordered. The class of transitions which obey this assumption is called the class of "order-disorder" transitions.

Limiting behavior and the van't Hoff enthalpy

This section discusses the behavior of the van't Hoff enthalpy, $\Delta H_{vh} \propto T^2 \frac{d\theta}{dT}$ as a function of the number of residues. The fraction of Z form residues, θ, in a polymer of length N can be written as

$$\theta = \frac{\sum_{i=1}^{N} iP(i)}{N\sum_{i=1}^{N} P(i)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} i\Omega_{i,j}p(i,j)}{N\sum_{i=1}^{N}\sum_{j=1}^{N}\Omega_{i,j}p(i,j)}$$

where the summation, i, is over all possible numbers of Z residues within a chain, P(i) is the total probability of i residues within a chain being Z form, p(i, j) is the statistical weight of any particular state in which i residues are Z form and j residues are boundaries and $\Omega_{i,j}$ is the number of possible configurations in which i residues are Z and j are boundaries. Substituting

$$p(i,j) = \sigma^{\frac{j}{2}} S^{i}$$

yields

$$\theta = \frac{\sum_{i,j}^{N} i\Omega_{i,j}\sigma^{\frac{j}{2}} S^{i}}{N\sum_{i,j}^{N}\Omega_{i,j}\sigma^{\frac{j}{2}} S^{i}} = \frac{S}{N}\frac{\partial \ln Q}{\partial S} \qquad III-1$$

For short chains ($N \ll N_0$), when the transition is expected to be all or none, only those states which contain no junctions are expected to contribute to Q

$$Q = S^{N} + 1 \qquad III-2$$

Where differences in cooperativity for B and Z forms at the helix ends have been ignored as before and in contrast to the Zimm-Bragg case.

As a consequence of this assumption, the midpoint of the transition is predicted to occur when S=1 (in the Zimm-Bragg case the ends favor the coil form, thus the *overall* transition will reach the midpoint only when helix states are more stable than coil states for *internal* residues, so that S >1). Then substituting $III-2$ into $III-1$ and taking the derivative with respect to T yields

$$\frac{d\theta}{dT} = \frac{N}{4}\frac{dS}{dT}\Big|_{s=1} = \frac{N\Delta H^{0}_{bp}}{4T^{2}_{m}}$$

at the midpoint, where ΔH^{0}_{bp} is the enthalpy difference between B and Z forms for a single residue and T_m is the temperature at the transition midpoint. Therefore, the van't Hoff enthalpy is equal to the actual enthalpy $N\Delta H^{0}_{bp}$ for the transition for a chain of length N ($N \ll N_0$).

When N becomes much larger than N_0, end effects can be ignored and the system

becomes identical to the Zimm-Bragg case. The van't Hoff enthalpy is obtained from an "apparent equilibrium constant"

$$K = \frac{\theta}{1-\theta}$$

$$\frac{d\ln K}{dT} = \frac{1}{\theta(1-\theta)} \frac{d\theta}{d(\ln S)} \frac{d(\ln S)}{dT}$$

Substituting the Zimm-Bragg expression for θ (20)

$$\theta = \frac{1}{2}\left(1 + \frac{S-1}{((1-s)^2+4\sigma S)}\right)$$

$$\left(\frac{d\ln K}{dT}\right)_{\theta=\frac{1}{2}} = N_0 \frac{\Delta H_{bp}}{RT^2}$$

Recall $N_0 = \frac{1}{\sigma^{\frac{1}{2}}}$ =the cooperative unit, so that the van't Hoff enthalpy is

$$\Delta H_{vh} = N_0 \Delta H_{bp}$$

This means that in a very long polymer, the van't Hoff enthalpy corresponds to the enthalpy change that occurs when a cooperative unit of residues converts from Z form to B form.

Intermediate length polymers

In polymers for which N is of the same order of magnitude as N_0, no simple expression exists for the van't Hoff enthalpy. In this section two combinatorial models are described for predicting the behavior of $B \rightleftharpoons Z$ transition curves as a function of N and the cooperative unit N_0. The first model assumes that the van't Hoff enthalpy can be written as

$$\Delta H_{vh} = \overline{N}(N) \Delta H_{bp} \qquad III-3$$

where $\overline{N}(N)$ is the average length at the transition midpoint of a B or Z tract in a polymer with N residues. The difference between N_0 and $\overline{N}$ is illustrated in Figure $III-2$. It is also assumed that ends do not favor B or Z forms so that S=1 at the transition midpoint. $\overline{N}$ is then simulated directly based on statistical factors without use of a partition function. The alternative approach uses the Zimm-Bragg assumption that end residues favor one form over the other, and calculates expressions for Q and θ at S=1 (as a function of N) based on combinatorial considerations without the use

of matrices. The first approach will be referred to as an order-order model, the second will be called an order-disorder model.

The order-order case

The following conditions are assumed: (1) the source of cooperativity is the unfavorability of $B-Z$ junctions (2) the $B \rightleftharpoons Z$ transition is two state and intramolecular (3) the sample is monodisperse with respect to polymer length (4) the $B-Z$ equilibrium at the midpoint of the transition is represented by an intrinsic equilibrium constant, S, equal to unity for individual base pairs. This final assumption is not made in models for order-disorder transitions.

The average length of B or Z form tracts ($\overline{N}$) may be calculated for a polymer of length N from:

$$\overline{N}(N) = \sum_{j=0}^{N} \overline{N}(N,j) \cdot P(j) \qquad III-4$$

where j is the number of boundaries or junctions, P(j) is the probability of a polymer molecule having j junctions, and $\overline{N}(N,j)$ is the average length of B- or Z-form tracts in a polymer containing N residues j of which are junctions. Ignoring end effects

$$\overline{N}(N,j) = \frac{N}{(j+1)} \qquad III-5$$

The variable P(j) is calculated from

$$P(j) = p(j) \cdot \left[\frac{N!}{(N-j)!j!}\right] \qquad III-6$$

where p(j) is the normalized probability of any given configuration in which the polymer contains j junctions and the term

$$\frac{N!}{(N-j)!j!}$$

is the number of configurations in which a polymer molecule with N residues may contain j junctions. Adjustment of the above term to accomodate a constraint of 1 to 10 base pairs between two $B-Z$ junctions has a negligable effect for cooperative units greater than 500 base pairs. The term p(j) is the product of the probability that j specific base pairs are junctions ($\sigma^{\frac{1}{2}}$) and $N-j$ specific base pairs are not $(1-\sigma^{\frac{1}{2}})^{N-j}$

$$p(j) = \sigma^{\frac{1}{2}}(1-\sigma^{\frac{1}{2}})^{N-j} \qquad III-7$$

where $\sigma^{\frac{1}{2}}$ is the probability or equilibrium constant for a $B-Z$ junction when S=1. Finally, the cooperative unit corresponds to

$$N_0 = \sigma^{\frac{-1}{2}}$$

Thus, for a given cooperative unit (N_0) the average length of B- or Z-form tracts ($\overline{N}$) at the midpoint of $B-Z$ equilibrium can be calculated for a polymer of length N by substituting eqs $III-5,6,7$ into $III-4$. The program *bzcoop* (see appendix III) carries out this computation.

Order-disorder transitions

The combinatorial method has been applied to the order-disorder case in (26). With minimal approximations (e.g. requiring a monodisperse sample and large enough N to replace sums in Q by integrals) expressions are derived for Q and θ for the case S=1. If their expressions are adapted to the $B \rightleftharpoons Z$ transition with the B-form assumed to be the disordered state

$$\theta_{s=1} = 1 - \frac{\frac{\sigma N^2}{4}+1}{\frac{\sigma N^2}{2}+1} \qquad \text{for } [\theta]_{s=1} < 0.25 \qquad III-6$$

$$\theta_{s=1} = \tfrac{1}{2}(1 - \frac{1}{N\sigma^{\frac{1}{2}}}) \qquad \text{for } [\theta]_{s=1} > 0.25 \qquad III-7$$

The temperature $T_{s=1}$ at which S=1 can be obtained from the transition midpoint for a sample containing polymers much larger than N_0. Then $\theta(N)$ can be obtained from the fraction Z form measured at $T_{s=1}$ for a sample containing polymers of length N. The cooperativity, σ, is computed by inverting eqs. $(III-8,9)$ for measurements of $\theta_{s=1}$ at several values of N, and averaging the results.

C. Comparison with experiment and estimate of N_0

In this section experimental data for the $B \rightleftharpoons Z$ transition in *poly*[$d(^{5me}C-G)$] is evaluated in according to the above theoretical considerations. The data discussed here are also presented and discussed elsewhere (28). Figure $III-3$ shows a series of fraction Z-form versus temperature curves for *poly*[$d(^{5me}C-G)$] in 0.35 mM $MgCl_2$,

50 mM NaCl, 5 mM tris, pH 8 for samples of different lengths. $Poly[d(^{5me}C - G)]$ was fractionated according to size and purified as described (28). This procedure produced samples with different size ranges, however, samples were not monodisperse as assumed in section B. The curves shown in figure $III-3$ were produced by first monitoring the absorbance at 295 nm for the fractionated samples in the above buffer and normalizing A_{295} to 1 at 55 ^0C. The absorbance data were then converted to fraction Z form versus temperature using

$$f = [A(T) - A_b(T)]/[A_z(T) - A_b(T)]$$

where $A(T)$ is the observed absorbance at temperature T, and $A_z(T)$ and $A_b(T)$ are the Z-form (upper baseline) and B-form (lower baseline) at the temperature T. The Z-form baseline was assumed flat ($A_{295(normalized)} = 1$) while the slope for the B-form baseline was obtained by averaging the slopes of low temperature linear regions from all the curves. The intercept for the B-form baseline was chosen individually for each curve. Each curve was then smoothed and differentiated using the program *derivative* (appendices $II-III$). The van't Hoff enthalpy was calculated either using $\Delta H_{vh} = -4RT^2\frac{d\theta}{dT}$ (derivative method) or by inputting the smoothed data into the programs *freeuni* and *lstft* (appendices II, III). The second method converts θ versus T to $RTlnK$ versus T, then calculates a slope and intercept to obtain ΔS_{vh} and ΔH_{vh} respectively. ΔS_{vh} can be obtained from the derivative method using the relation

$$\Delta S_{vh} = \frac{\Delta H_{vh}}{T_m}$$

for a unimolecular transition where T_m is the midpoint temperature. The two calculated values for ΔS_{vh} and for ΔH_{vh} differed by less than 10% in every case. Presented values are the average of the two calculated values.

Values of ΔH_{vh} are plotted in figure $III-4$ along with a theoretical curve of $(\frac{\overline{N}}{N_0})$ for a range of values of N_0 based on the order-order model presented above. Eq. $(III-3)$ predicts that $\frac{\Delta H_{vh}}{\Delta H_{N\to\infty}} = \frac{\overline{N}}{N_0}$, ($\Delta H_{N\to\infty} = \Delta H_{vh}$ for an infinite length polymer) since

ΔH_{bp} is assumed independent of polymer length. A comparison of experimental data with the order-disorder model is presented in Table III-1. Values of θ and calculated values of σ are presented for each curve assuming three values of T_t, the temperature at which S=1. The values of T_t were chosen based on the transition midpoint of the longest polymer sample, in this case 39.4 ^0C. Since the longest polymer length sample was not of infinite length, it can be presumed that it undergoes its transition midpoint at a temperature slightly lower than that at which S=1.

The data in Table III-1 indicate that N_0 is best fit to a value between 1000 and 1500 base pairs for an order-disorder transition, while figure $III-4$ suggests a best fit of 800 – 1200 for an order-order transition.

The nature of the transition

From comparison of the fit of experiment to theory shown in figure $III-3$ to that found in Table III-1 it appears that the order-order model fits the data as well if not better than the order-disorder model. However, there are at least three reasons to believe that the system is actually behaving as an order-disorder system, with the B-form corresponding to the disordered state (e.g. B-form is favored for end residues). First, the increase in the transition midpoint temperature for decreasing polymer length clearly apparent in figure $III-3$ is predicted for an order-disorder transition, since B form ends will tend to induce internal residues to adopt the B state, and this effect is strongest for the shortest polymers. The order-order model presented above would predict an intersection of all curves at $\theta = \frac{1}{2}$. Second, DNase I digestion experiments were performed under conditions (3 mM $MgCl_2$) in which the polymer was predominantly Z-form as revealed by UV absorbance and circular dichroism (CD). When electrophoresed on an agarose gel and ethidium stained following digestion, the Z-form polymers were apparently intact, while similar experiments on B-DNA resulted in virtually complete cleavage. When the same experiment was performed on 5'^{32}P

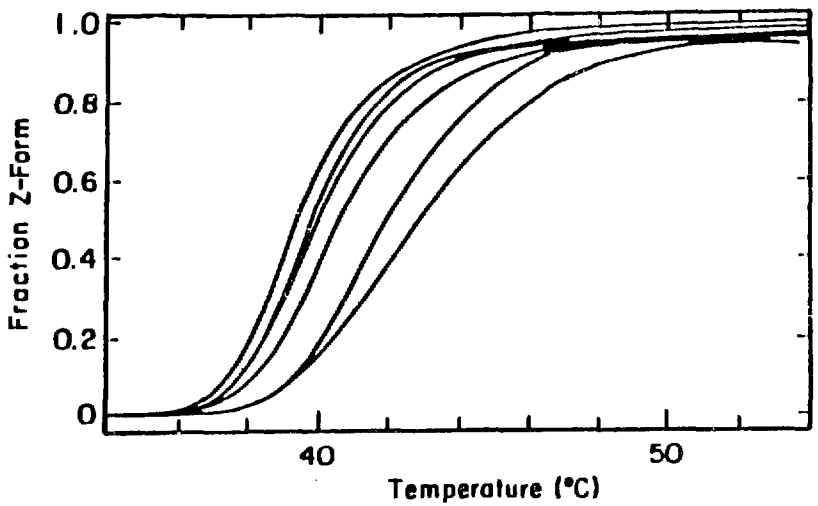
labeled samples and products were electrophoresed and autoradiographed, no label was visible. This result can be explained if the end residues were B-form, and therefore susceptible to digestion even under conditions in which the overall $B-Z$ equilibrium heavily favored Z-form. Third, experiments on a set of $d(^{5me}C-G)_n$ oligomers (n=3, 5, 9) showed that these oligomers require much more stringent conditons ($[MgCl]_2 > 100$ mM, 50 mM NaCl) in order to produce a significant contribution of the Z-form to the equilibrium. Thus, it seems that under the conditions of these experiments, end residues of $poly[d(^{5me}C-G)]$ have a strong preference for B-form as compared to internal residues, as assumed in the order-disorder model.

Kinetic measurements

Rate constants for the $B \rightleftharpoons Z$ transition were measured for the same set of sized polymer samples as above by jumping the $MgCl_2$ concentration from 0 to $3mM$ $MgCl_2$ and measuring the absorbance at 295 nm as a function of time. The A_{295} versus time curve was fit to a single exponential. The transition rate was found to increase as a function of polymer length (fig $III-5$, also 28) contradicting theoretical predictions (22). Repetition of the experiment at 35^0C, 40^0C, and 45^0C (28) led to the conclusion that the activation enthalpy for the transition is constant with respect to polymer length. Therefore, the observed effect of polymer length on the transition rate must result mainly from a pre-exponential (or "entropic") factor K_0 in the rate constant $K_r = K_0 e^{-\frac{H^*}{RT}}$, where H^* is the activation enthalpy and K_r is the rate constant for conversion of B-form to Z-form.

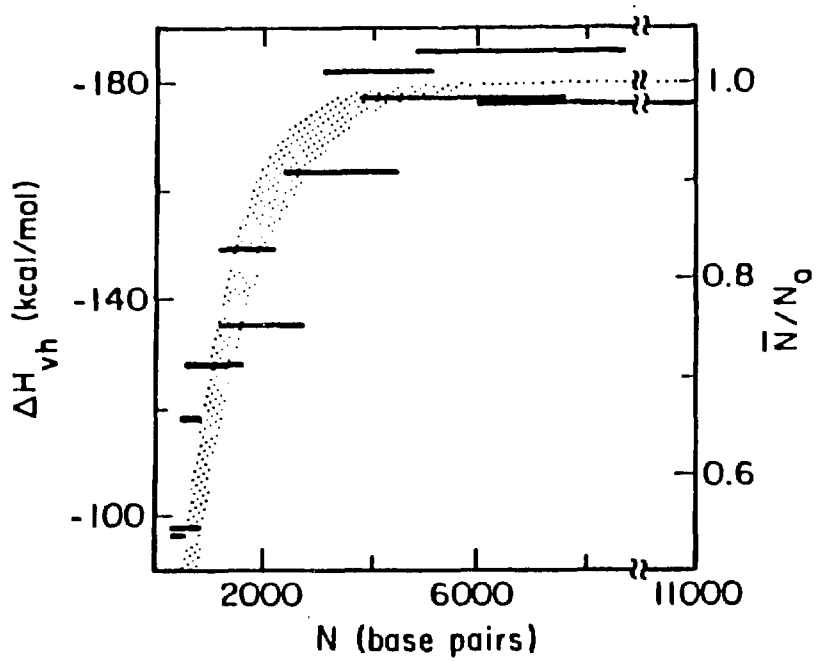
A remarkable correlation is observed between the kinetic and thermodynamic data (Table III-2). Empirically, the lifetime τ for the $B \rightleftharpoons Z$ transition in $poly[d(^{5me}C-G)]$ is found to be inversely proportional to the square of the van't Hoff enthalpy. For this reason, theoretical values of $\overline{N^{-2}}$, which is expected to predict ΔH_{vh}^{-2}, are plotted in figure $III-5$ alongside τ. The reasonable fit of both

Figure III-3



B-Z transition curves of poly(dG-m^5dC) in 0.35 mM $MgCl_2$, 50 mM NaCl, 5 mM TRIS, pH 8 as a function of polymer length. Fraction Z-form versus temperature for poly(dG-m^5dC) samples of varying length. For the curves from left to right, the respective polymer lengths ranges are 11000-6000, 7600-3800, 4500-2400, 2700-1200, 1600-600 and 800-300 base pairs.

Figure III-4



Van't Hoff enthalpy values (ΔH_{vh}) for the B-Z transition of poly(dG-m^{5}dC) in 0.35 mM $MgCl_2$, 50 mM NaCl, 5 mM TRIS, pH 8 as a function of polymer length. The horizontal bars represent experimental ΔH_{vh} values for poly(dG-m^{5}dC) samples of the indicated polymer length range. Calculated curves of ΔH_{vh} versus N and $\bar{N}/N_o$ versus N for N_o=800-1200 base pairs (see text) are represented by the shaded area.

Kinetic data for the B to Z transition of poly(dG-m^5dC) at the indicated temperatures (Celcius) in 3 mM $MgCl_2$, 50 mM NaCl, 5 mM TRIS, pH 8 as a function of polymer length. The horizontal bars represent experimental first order relaxation times (τ) for poly(dG-m^5dC) samples of the indicated polymer length range. Calculated curves of $\bar{N}^{-1}$ versus N for N_0=800-1200 (see text) are represented by shaded areas for each temperature.

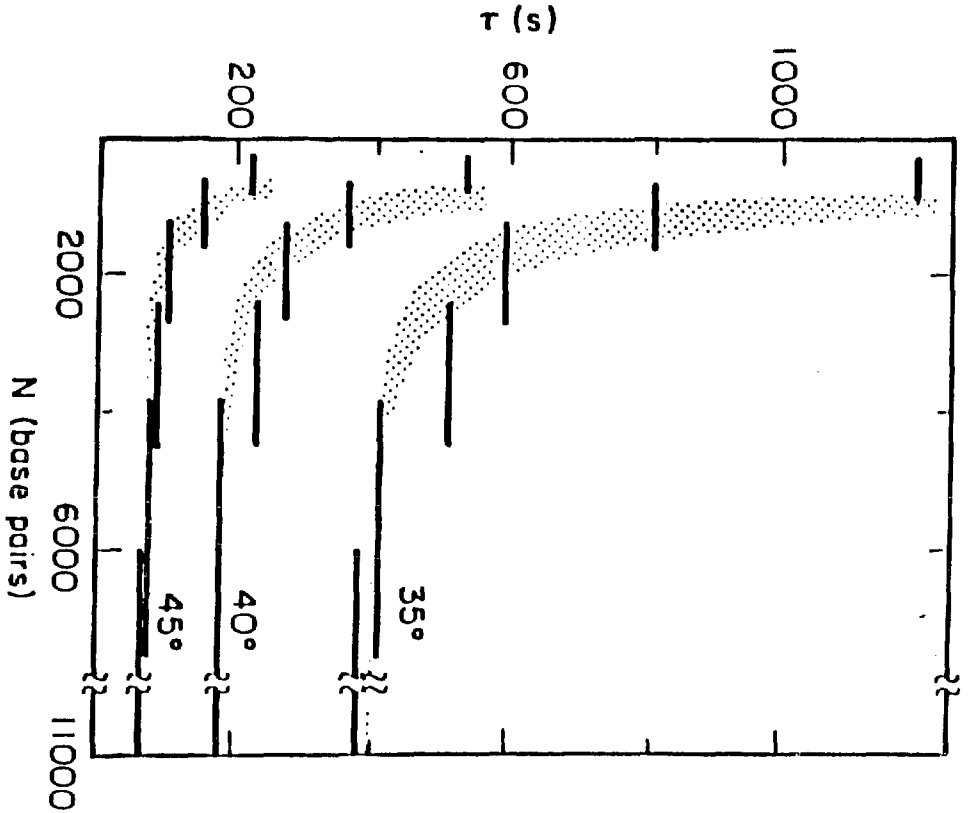


Figure III-5

Table III-1

Experimental values of σ and N_0 calculated according to the combinatorial method using an "order-disorder" model (26, pages 98-105). The formulas used (see text) are valid when the equilibrium constant for an individual residue, S=1. The temperature at which S=1 is determined from the transition midpoint temperature of a sample of infinite length polymers. The longest polymer length sample available had a transition midpoint at 39.4 °C. Since the polymer length of this sample was not infinite, $T_{S=1}$ is estimated to be slightly lower than 39.4 °C. Shown are estimates of σ for several estimated values of $T_{S=1}$. Values of σ have been multiplied by 10^7, values of N_0 by 10^{-2}.

	$T_{S=1}$=38.8 °C		$T_{S=1}$=39.0 °C		$T_{S=1}$=39.2 °C	
Ave. pol. length (bps)	σ	N_0	σ	N_0	σ	N_0
580	13	8.8	15	8.2	18	7.4
1050	3.6	17	4.6	15	5.7	13
1900	5.8	13	7.9	11	12	9.0
3450	4.0	16	3.8	16	4.5	15
5700	1.8	24	1.5	26	2.0	23
8500	7.0	12	1.4	27	2.9	18
	$\sigma_{ave}=5.8X10^{-7}$		$\sigma_{ave}=5.7X10^{-7}$		$\sigma_{ave}=7.5X10^{-7}$	
	(N_0=1314)		(N_0=1324)		(N_0=1153)	

Table III-2

Correlation of τ, the lifetime for the B to Z transition in 3 mM $MgCl_2$, with the square of the van't Hoff enthalpy, ΔH_{vh}, measured in 0.35 mM $MgCl_2$ for samples of different polymer lengths.

N_{ave}(bps)	τ(40 °C, sec^{-1}) ($X10^{-2}$)	ΔH_{vh}(Kcal Mol^{-1}) ($X10^{-2}$)	$\tau\Delta H_{vh}^2$ ($X10^{-6}$)
580	5.3(±5%)	0.98(±5%)	5.1
1050	3.6	1.28	5.9
1900	2.7	1.35	4.9
3450	2.3	1.64	6.2
5700	1.8	1.77	5.6
8500	1.8	1.76	5.6

thermodynamic and kinetic data to the same theoretical calculations observed in figures $III-4$ and $III-5$ confirms this correlation.

Kinetic and thermodynamic results can both be explained by a statistical model for nucleation

Though a survey of theoretical papers (21, 22, 27) found no predictions of the observed effect of polymer length on kinetics, the increase of transition rate with increasing polymer length can be explained by a nucleation limited mechanism. If the transition consists of two steps, nucleation and propagation, with the latter far faster than the former, the transition rate will be determined by the rate of nucleation and by the average number of base pairs which are converted during each nucleation event ($\overline{N}_k$). $\overline{N}_k$ will determine the total number of nucleations required. While the nucleation rate should be relatively independent of polymer length for long polymers, $\overline{N}_k$ is a function of the cooperativity and the length. Since each nucleation event requires the formation of two boundaries the transition rate will be proportional to the square of the number of base pairs flipped per boundary formed ($\overline{N}_k^2$). Note that $\overline{N}_k$ is an increasing function of polymer length for polymers of a length of the order of the cooperative unit.

Now, $\overline{N}_k$ is determined by the same factors as $\overline{N}$. It is therefore reasonable as a first approximation to identify $\overline{N}_k$, the average number of base pairs flipping from B to Z under non-equilibrium conditions heavily favoring Z-form with $\overline{N}$, the average number of base pairs in a B or Z tract under "fifty-fifty" equilibrium conditions. This leads to the prediction of a connection between τ, the lifetime for the transition and ΔH_{vh}. Specifically,

$$\tau \propto \frac{1}{\overline{N}^2}$$

$$\Delta H_{vh} \propto \overline{N}$$

$$\Delta H^2_{vh} T = constant$$

as observed in Table III-2.

D. Summary and conclusions

The studies presented in this chapter confirm that the $B \rightleftharpoons Z$ transition in DNA polymers can be analyzed according to the same type of Ising model formalisms which have been used in the study of proteins (26). The range measured for the cooperative unit for the transition in $poly[d(^{5me}C - G)]$ is somewhat larger than that measured in other $B-Z$ systems (3, 9, 18, 19) from salt titration curves and considerably larger than that measured calorimetrically for the same polymer in $1mM$ $MgCl_2$ with phosphate buffer (16). The cooperative unit may vary with sequence and buffer conditions. For example, any factor which stabilizes or destabilizes the $B - Z$ junction is expected to influence the cooperative unit. The size, structure, and thermodynamic properties of the $B - Z$ junction are not thoroughly characterized (10, 15), it is unclear what factors affect its stability. Calorimetric and optical measurements will also be affected by the sizes of the polymers in the sample-in particular, lengths should be greater than about three times the cooperative unit (Figure $III - 4$). Heating rates will also affect ΔH_{vh} obtained from optical and calorimetric experiments (Appendix I), but not the ΔH_{bp} measured calorimetrically.

In light of these experiments it is clear that in $d(^{5me}C-G)_n$ sequences, end residues have a much stronger propensity to form the B structure than internal residues. In general, the relative stabilities of the B- and Z-forms for end residues can be obtained from the values of θ at which a series of curves for different polymer lengths (such as those shown in figure $III - 3$) intersect (22).

$$\theta(intersection) = \frac{\sigma_z'}{\sigma_b'}$$

where σ_z' and σ_b' are the nucleation parameters for end residues for the Z and B forms respectively. Figure $III - 3$ demonstrates that for $poly[d(^{5me}C - G)]$ in 0.35 mM

$MgCl_2$ $\frac{\sigma_+'}{\sigma_b} \approx 0$, i.e. nucleation of Z-form is virtually impossible at polymer ends. For *poly*[*d*(*C*·*G*)] in NaCl solutions, $\frac{\sigma_+'}{\sigma_b} = 0.35$ has been reported (18).

The correlation observed between ΔH^2_{vh} and $\frac{1}{r}$ observed in table III-2 was reproducible and independent of temperature. Therefore it seems to indicate a common factor, perhaps the formation of a $B-Z$ junction, determining both the cooperativity and kinetics of the system. More precise investigations of this relationship and of the value of the cooperative unit would require more monodisperse samples. Also of interest would be the determination of the cooperative unit as a function of temperature over a broad temperature range. In this chapter the cooperative unit has been assumed independent of temperature, which is consistent with the linear Arrhenius plots (28) and van't Hoff plots observed for the temperature range $35^0C - 45^0C$. As mentioned in section IIIA, a temperature dependence of the cooperative unit would be indicative of an enthalpic component to the destabilization energy of the junction, which would be expected if the junction involves unpaired bases.

References

1. Pohl, F. M. and Jovin, T. M. (1972) J. Mol. Biol. 67, 375 – 396.

2. Wang, A. H. J., Quigley, G. J., Kolpak, F. J., Crawford, J. C., van Boom, J. H., van der Marel, G. and Rich, A. (1979) Nature 282, 680 – 686.

3. Jovin, T. M., Soumpasis, D. M. and McIntosh, L. P. (1987) Annual Rev. Phys. Chem. 38, 86-111.

4. Rich, A., Nordheim, A. and Wang, A. H. J. (1984) Annual Rev. Biochem. 53, 791 – 846.

5. Barton, J. K., and Raphael, A. L. (1985) Proc. Natl. Acad. Sci., USA 82, 6460 – 6464.

6. Kmiec, E. B. and Holloman, W. K. (1986) Cell 44, 545 – 554.

7. Jaworski, A., Hsieh, W.-T., Blaho, J. A., Larson, J. E. and Wells, R. D. (1987)

Science 238, 773 – 778.

8. Roy, K. B. and Miles, H. T. (1983) Biochem. Biophys. Res. Commun. 115, 100 – 105.

9. Hall, K. B. (1984) Ph. D. thesis, University of California, Berkeley.

10. Peck, L. J. and Wang, J. C. (1983) Proc. Natl. Acad. Sci., USA 80, 6206–6210.

11. Behe, M. and Felsenfeld, G. (1981) Proc. Natl. Acad. Sci., USA 78, 1619 – 1623.

12. Walker, G. T., Stone, M. P. and Krugh, T. R. (1985) Biochemistry 24, 8436 – 8439.

13. Chaires, J. B. (1985) Biochemistry 24, 7479 – 7486.

14. Stahl, F. W. (1979) Genetic Recombination, W. H. Freeman, San Francisco.

15. Kang, D. S. and Wells, R. D. (1985) J. Biol. Chem. 260, 7783 – 7790.

16. Chaires, J. B. and Sturtevant, J. M. (1986) Proc. Natl. Acad. Sci., USA 83, 5479 – 5483.

17. Klump, H. H. and Jovin, T. M. (1987) Biochemistry 26, 5186 – 5190.

18. Pohl, F. (1983) Cold Spring Harbor Symp. Quant. Biol. 47, vol. 1, 113 – 118.

19. Ivanov, V. I. and Minyat, E. E. (1981) Nucl. Acids Res. 9, 4783 – 4798.

20. Engel, J. and Schwarz, G. (1970) Angewandte Chemie 9, 389 – 400.

21. Zimm, B. H. and Bragg, J. K. (1959) J. Chem. Phys. 31, 526 – 535.

22. Schwarz, G. (1968) Biopolymers 6, 873 – 897.

23. Pathria, R. K. (1978) Statistical Mechanics, third edition, Pergamon Press, New York.

24. Landau, L. D. and Lifshitz, E. (1980) Statistical Physics, part 1, Pergamon Press, New York.

25. Dyson, F. (1971) Commun. Math. Phys. 21, 269 – 283.

26. Poland, D. and Scheraga, H. A. (1970) Theory of Helix-Coil Transitions in

Biopolymers, Academic Press, New York.

27. Cantor, C. R. and Schimmel, P. R. (1980) Biophysical Chemistry, W. H. Freeman, San Francisco.

28. Walker, G. T. and Aboul-ela, F. (1987) submitted to J. Biomol. Struc. Dynamics.

Chapter IV

A. Introduction

The most studied system of developmental regulation of gene expression in eukaryotes is probably the control of 5S RNA synthesis in the frog, *Xenopus Laevis* (1). There are two distinct types of 5S RNA genes in *Xenopus*, oocyte 5S genes and somatic 5S genes. Transcription of oocyte 5S RNA genes occurs at very different rates during different stages of development. Though oocyte 5S genes are present at approximately fifty times the abundance of somatic 5S genes, the production of oocyte 5S RNA surpasses that of somatic 5S RNA only during the early stages of embryogenesis (1).

Among the set of proteins which are known to participate in the regulation of this process are three "transcription factors", designated as transcription factors IIIA, B and C. Together, and perhaps in conjunction with other unknown components, these three factors interact with 5S DNA to form what is known as a "transcription complex" (2). The formation of this complex facilitates transcription by RNA polymerase. Moreover, the complex is stable to multiple passages by the enzyme (3).

The study of deletion mutants together with the sequencing of the TFIIIA gene and mapping of its DNA binding site has led to an interesting model for the structure of the protein and its interaction with the gene (4-6). The TFIIIA amino acid sequence contains nine repetitions of a 30 amino acid motif. These 30 amino acid units are thought to form nine zinc binding "fingers" which bind to a 54 base pair site in what is known as the internal control region of the gene. The repeating structure of the protein is paralleled by a repeating motif in the base sequence and in the DNase I digestion pattern of the DNA binding site (6, 7). Nuclease digestion and methylation protection experiments suggested that the gene binding site of TFIIIA exhibited A-form conformational features (6, 7). These enzymatic mapping experiments, in conjunction

with the ability of the protein to bind both the control region of the gene and the 5S RNA (8) led Klug and coworkers to propose that TFIIIA recognizes an A-form geometry (6).

Further support for an A-form TFIIIA binding site was provided by an A form crystal structure (9) for the deoxyoligonucleotide *dGGATGGGAG·dCTCCCATCC* which represents the strongest binding site (5, 6) of the gene (base pairs 81 to 89 of the gene). Similar findings of A-form crystal structure have been reported for other G.C rich DNA sequences (10, 11). However, nuclear magnetic resonance (NMR) and optical spectroscopy techniques (circular dichroism, Raman) have demonstrated that the conformation of a specific deoxyoligonucleotide can significantly differ in solution and crystal (12, 13). Indeed, recent circular dichroism (CD) studies have contradicted the proposal of A-form geometry for the 54 base pair 5S RNA gene sequence recognized by TFIIIA (14). Therefore, solution studies of the same deoxyoligonucleotide crystallized by Kennard and coworkers appear to be necessary to fully characterize the structure of this portion of TFIIIA recognition sequence.

NMR has been used succesfully to discriminate between different nucleic acid conformations in solution (13, 16). In particular, two-dimensional correlated spectroscopy (COSY) experiments are sensitive to the geometry of the sugar residue, while 2D-nuclear Overhauser enhancement (NOESY) experiments are sensitive to interproton distances. Distances derived from NOESY data can be compared with distances from X-ray crystallography for different DNA conformations. We have investigated the solution structure of the DNA oligonucleotide *dGGATGGGAG·dCTCCCATCC* using these NMR techniques and circular dichroism (CD), which is sensitive to the stacking of the bases (17).

In order to explore the basis of TFIIIA binding to both DNA and RNA, and specifically the possibility that TFIIIA is recognizing a common structure in both

DNA and RNA, comparison of the structural properties of the DNA 9-mer with an RNA oligomer of identical sequence was deemed advantageous. However, since it is easier to synthesize longer RNA oligomers using T7 RNA polymerase (18), we chose to synthesize a self-complementary 18-mer consisting of the two above strands in succession. DNA and RNA 18-mers *GGATGGGAGCTCCCATCC* (with U substituted for T in RNA) were synthesized and compared using circular dichroism, and chemical and enzymatic digestions. The structure of this TFIIIA recognition fragment has also been tested by examining the CD of the deoxyoligonucleotide at high trifluoroethanol (TFE) concentrations, since TFE is known to induce a B to A transition in DNA.

B. Materials and Methods

Oligonucleotide synthesis and purification

DNA oligonucleotides were synthesized on an Applied Biosystems 381A instrument. The RNA oligonucleotide was synthesized using T7 RNA polymerase and an oligomeric DNA template (18). Purification of the two DNA 9-mer strands was by reverse phase HPLC and desalting on a G10 column (Pharmacia) followed by extensive dialysis. The DNA and RNA 18-mers were purified by preparative 20% acrylamide gel electrophoresis under denaturing conditions (7*M* urea). The purity of the samples was checked by 20% acrylamide gel electrophoresis. Terminal 5' triphosphates, which are a product of synthesis by T7 RNA polymerase, were removed with calf intestinal phosphatase (Boehringer-Mannheim) followed by purification on Sep-pak cartridges (Millipore).

Circular Dichroism (CD)

CD spectra were recorded on a Jasco J500C spectropolarimeter at 25°*C* using 1 cm pathlength cuvettes. Nucleotide concentrations were 50 μM. Values of $\Delta\epsilon$ are expressed in terms of base pairs.

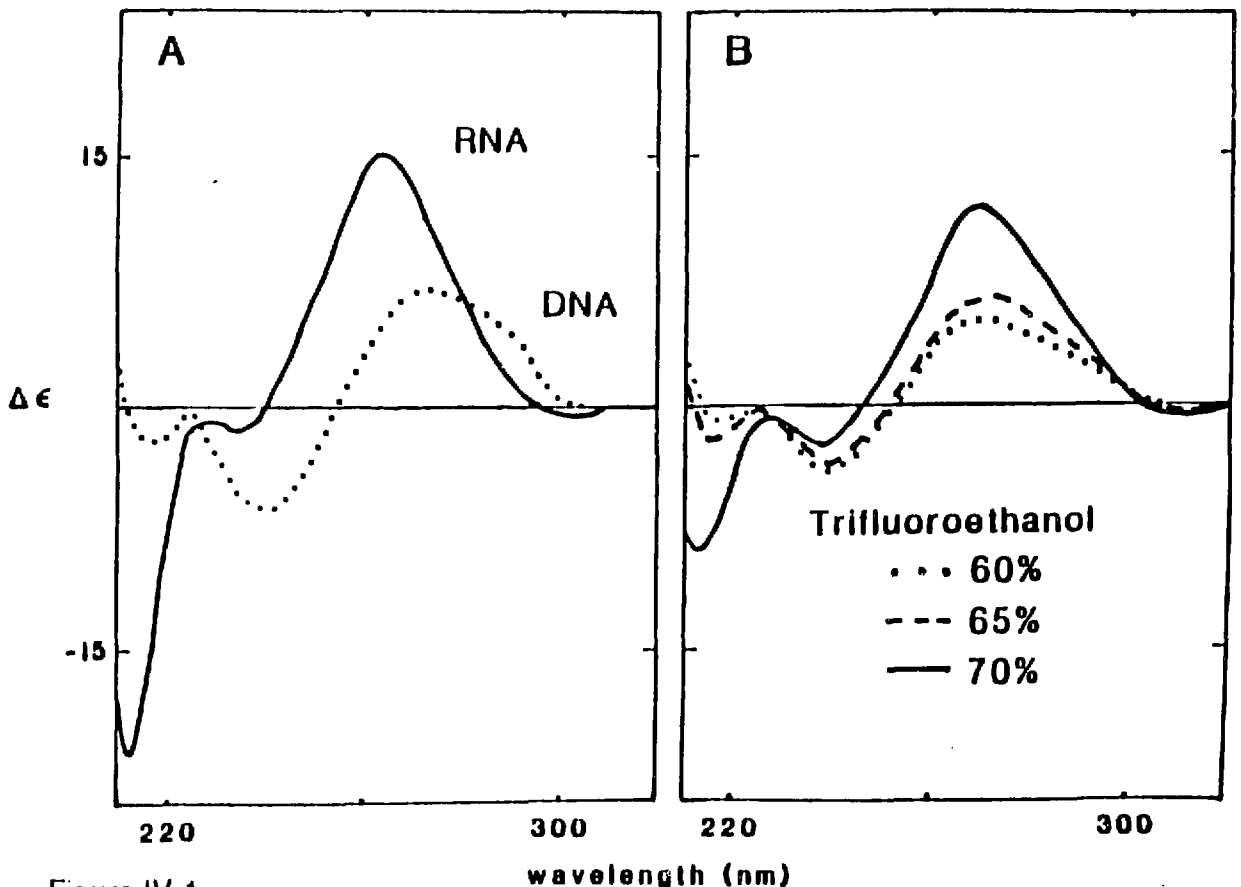


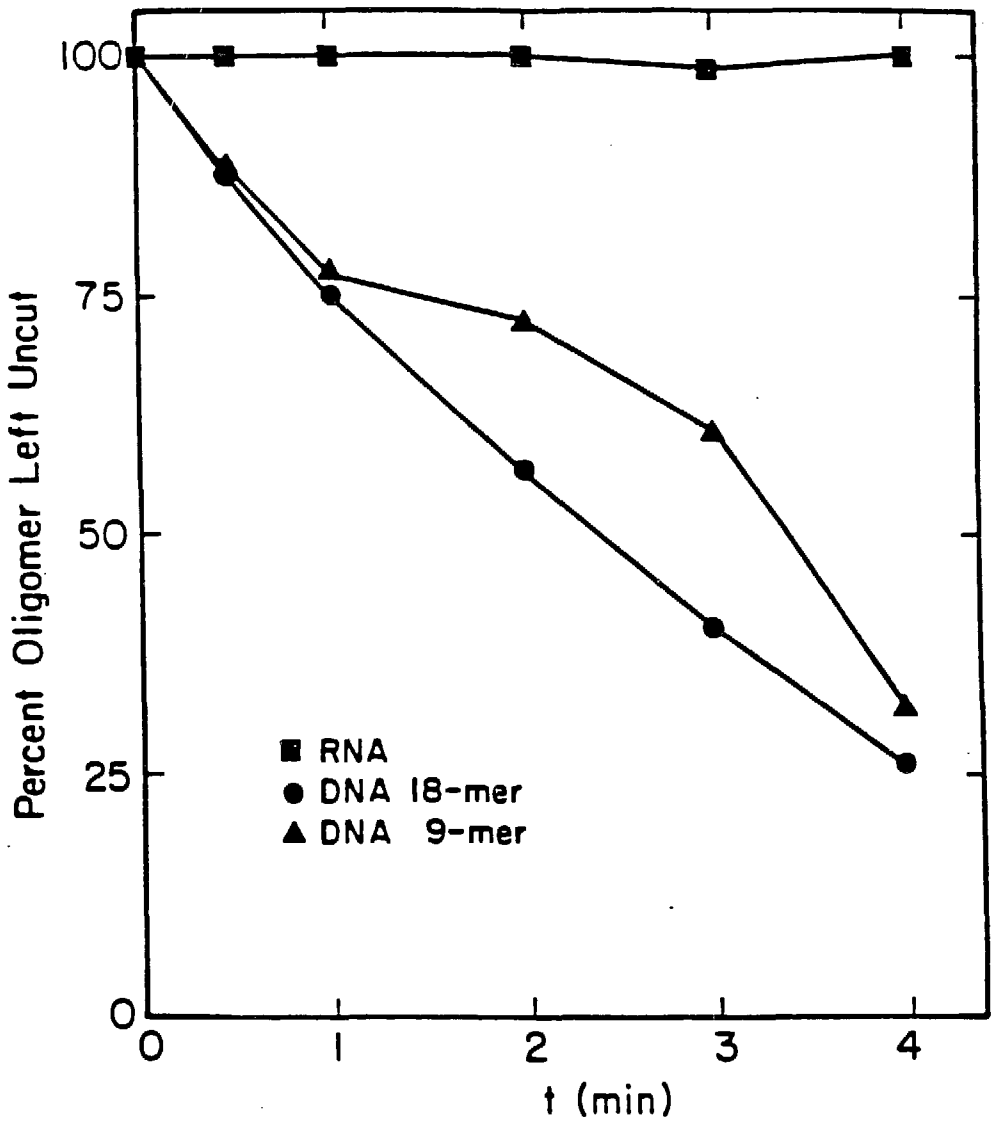
Figure IV-1

(a) CD spectra for the DNA 18-mer (·····) d(GGATGGGAGCTCCCATCC) and the RNA 18-mer (——) r(GGAUGGGAGCUCCCAUCC) in 50 mM NaCl, 8mM Na_2HPO_4, 0.1 mM Na_2EDTA, pH 7, 25 °C.

(b) B to A transition in the DNA 18-mer induced by trifluoroethanol (TFE). CD spectra for the DNA 18-mer in 2mM NaCl, 0.1 mM Na_2EDTA, 25 °C at the indicated TFE concentrations (v/v).

Figure IV-2

Cu-phenanthroline time course digestion of the RNA 18-mer r(GGAUGGGAGCUCCCAUCC) (■), DNA 18-mer d(GGATGGGAGCTCCCATCC) (●), and the DNA 9-mer d(GGATGGGAG)·d(CTCCCATCC) (▲). The percent of uncut oligonucleotide is reported vs. the reaction time. Cu-phenanthroline is specific for B-form helices (see 25).



Chemical digestion

Copper phenanthroline digestion studies with ^{32}P labeled oligonucleotides were at room temperature following the procedure of (19). Digestion mixtures were quenched as a function of time with 7 mM EDTA, 4 M urea (final concentrations). Each aliquot was run on a 20% acrylamide, 7 M urea denaturing gel. Full length oligonucleotide bands were cut and the amount of radioactivity counted on a liquid scintillation counter.

NMR

NMR samples were lyophilized several times with 99.8% D_2O and then diluted to approximately 2 mM in strands (or about 3.5 mM for the 250 ms D_2O NOESY and the H_2O NOESY) with 0.4 ml of 10 mM sodium phosphate, 50 mM NaCl, 0.1 mM Na_2EDTA (pH 7) in 99.96% D_2O (Aldrich). 2D-NMR spectra were recorded at 500 MHz on a General Electric GN-500 spectrometer at 30 ^oC, well below the melting temperature of the double strand (T_m=50 ^oC under the present conditions). Linewidths were somewhat broadened at lower temperatures, whereas signs of premelting were apparent in 1D-NMR spectra above 35 ^oC. Phase sensitive NOESY spectra at different mixing times were recorded using the TPPI method (20); the mixing times were 60, 120, 200 and 250 ms; the sweep width 4032 Hz. 450 FID's were collected and 2k complex data points recorded for every FID. Data were zero filled to 1k real points in t_1, and apodized prior to Fourier transformation using a skewed sine bell (phase shift 60^0, skewness 0.7) in both dimensions. Deviations from linearity in the cross peak intensities as a function of mixing time were observed in buildup rates for NOEs, but the qualitative trends described here were evident at all mixing times. In the following we shall therefore refer only to the data pertaining to the longest mixing time spectrum (250 ms). The phase sensitive COSY spectrum was recorded using the TPPI technique (20) with presaturation of the HDO peak. 750 FID's, each of 8K complex data points were collected; the sweep width was 4000 Hz. High digital resolution (1 Hz) was

desirable though not necessary to measure the coupling parameters. Data were zero filled to $4K$ real points in t_1 and a 30^0 phase shifted skewed sine-bell (skewness 0.7) was used for apodization in t_1 and t_2. The phase sensitive NOESY spectrum in H_2O was recorded using the TPPI technique with all three 90^0 pulses replaced by the $(90) - \tau - (-90)$ "jump and return" (21) pulse sequence. Attempts to replace only the final pulse of the sequence with the jump and return pulse resulted in the presence of two diagonals in the final 2-D spectrum. The presence of the second diagonal may have been due to the effect of the jump and return sequence on the relative phasing of the final pulse of the TPPI sequence, which presumably would modify the phase cycling. 500 FID's were collected and $4k$ complex data points recorded for each FID. The carrier frequency was set on the H_2O resonance. τ was 60 μsec. Data were zero filled to $2K$ real points in t_1 and apodized using a 60^0 phase shifted skewed sine-bell (skewness 0.7).

C. Results

CD

Circular dichroism is sensitive to stacking interactions and large differences exist between the CD spectra of A- and B-form nucleic acids (17, 22-24). Hallmarks of A-form RNA and A-form DNA are CD spectra with large positive magnitude at 270 nm, slight negative magnitude at 240 nm and large negative magnitude at 210 nm. In contrast, B-DNA of identical sequence generally exhibits a conservative CD spectrum with less positive magnitude around 270 nm, greater negative magnitude at 240 nm and near zero magnitude at 260 and 210 nm (23, 24). CD spectra for the DNA 18-mer $d(GGATGGGAGCTCCCATCC)_2$ and the RNA 18-mer $r(GGAUGGGAGCUCCCAUCC)_2$ are shown in Fig. *IV* – 1a. Comparison of the RNA and DNA spectra clearly support a DNA conformation distinct from the RNA conformation. The CD spectrum of the DNA 9-mer is also consistent with a B-form

conformation (data not shown). The DNA spectra are qualitatively similar to that reported for a 54 base pair fragment corresponding to the full TFIIIA binding site (14), suggesting that the B-form structure observed here is not an artifact due to end effects. CD spectra were calculated for these oligonucleotides according to the method of Gray and coworkers (23, 24) which uses an empirical basis set of DNA and RNA polynucleotide spectra . Observed and calculated spectra qualitatively agree, and support respective B- and A-form conformations for the DNA and RNA oligonucleotides.

CD studies have shown that trifluoroethanol (TFE) induces a B to A transition in DNA (17, 22). CD spectra were recorded for the DNA 18-mer in 60, 65 and 70% TFE (v/v), conditions under which the oligonucleotide remains double stranded (Fig. *IV* – 1*b*). The DNA 18-mer spectra exhibit increasing positive and negative ellipticity at 270 and 210 nm, respectively, with increasing TFE concentration. Comparison to the RNA 18-mer spectrum (Fig. *IV* – 1*a*) reveals an apparent B to A transition. The DNA 9-mer exhibits a similar B to A transition between 60 – 70% TFE (data not shown). This range is comparable to TFE concentrations required to induce the transition in other DNA sequences (22). Other conditions were tested for the capacity to induce a B to A transition in the 9-mer. The B-form appearance of the CD spectrum is conserved in the presence of high salt concentrations (up to 2*M* NaCl) and in the buffer used for crystallization by Kennard and coworkers (12 *mM* Na cacodylate, 12 *mM* Na acetate, pH 6.5 (9)). Spermine was also added to match more closely the conditions used in X-ray studies. No effect on the CD spectrum was observed for spermine concentrations up to 1.6 *mM*.

Chemical and enzymatic probes

The structures of the DNA and RNA 18-mers were probed with a variety of enzymes and the cleavage reagent 1,10-phenanthroline copper ion, which is reported to be specific for B-form helices (25). The DNA 9-mer and 18-mer were readily

cleaved by 1,10-phenanthroline copper, whereas no activity was observed with the RNA oligomer (Fig. *IV* – 2). As expected, the DNA was cleaved by DNase I and the restriction endonuclease Alu I, while double strand specific ribonuclease V1 efficiently cleaved the RNA (data not shown).

NMR

Assignments of the non-exchangeable aromatic proton resonances, as well as those of the H1', H2', H2" and methyl protons, were obtained from the NOESY spectrum, and cross-checked in the COSY spectrum, using standard sequential methods (26). The region of the NOESY spectrum corresponding to aromatic to H1' cross peaks is shown in Fig. *IV* – 3, together with the connectivity pathway used for assignments for one of the strands. The H2' resonances were distinguished from the H2" resonances on the basis of the different shapes of their cross-peaks to H1' in the phase- sensitive COSY (see Fig. *IV* – 4*a*). This distinction was confirmed by the relative intensity of the H1' to H2" vs. H1' to H2' cross peaks, the former always being stronger than the latter regardless of oligonucleotide conformation (27). For most residues, H2' was upfield from H2"; however, the chemical shifts of the H2' and H2" resonances were the same for the C18 residue, whereas H2" was downfield from H2' for the G9 residue. In addition, assignments for the H3' and H4', and a few of the H5' and H5" resonances were made using cross peaks to aromatic and sugar H1', H2' and H2" protons. Assignments are summarized in Table IV-1.

One of the major differences between A- and B-form DNA is the conformation of the sugar, which can be conveniently described by means of the pseudorotation phase angle (28). In A-form DNA, the sugar pucker is 3'-endo, corresponding to a pseudorotation angle of 18°. In B-form DNA the sugar pucker is usually found in the south family of conformers, frequently close to canonical 2'-endo (phase angle 162°). Magnitudes of the scalar couplings between nucleic acid sugar protons are very sen-

sitive to the sugar conformation. Two-dimensional correlated spectroscopy (COSY) is a suitable technique for evaluating coupling constants, but direct measurements are frequently impossible because of peak overlap and limited digital resolution. However, measurements of individual coupling constants is not necessary (29). Required information can be extracted from the knowledge of multiplet widths and splitting patterns in COSY spectra. The percent of time the individual sugar moieties are found in one of two major conformers while undergoing rapid conformational equilibrium can also be derived. Some cross peaks from the region of the COSY spectrum corresponding to the H1'/H2' and H1'/H2" sugar protons are shown in Fig. *IV – 4a*. The multiplet widths $\Sigma_{1'}$, $\Sigma_{2'}$, and $\Sigma_{2''}$ (defined in Fig. *IV – 4a*) along with the coupling constants $J_{1'2'}$ and $J_{1'2''}$ (measured as shown in Fig. *IV – 4b*) are reported in Table IV-2 for all but residue C18. It was not possible to measure the coupling constants for the C18 residue because H1'/H2' and H1'/H2" cross peaks are superimposed. The value of $\Sigma_{1'}$ is an excellent marker for the relative population of the south and north conformers. Large values of $\Sigma_{1'}$ (> 14.5 *Hz*, see Table IV-2) are a conclusive indication that the fraction of S-type conformer is greater than 80% for most residues (26).

Once the major conformer is determined, the different shapes of the COSY cross peaks and the multiplet widths distinguish unambiguously the H2' from the H2" resonances (Fig. *IV – 4a*, see also 29, 30). The splitting patterns of the H1', H2', and H2" are also qualitatively consistent with a larger population of the south conformer. Approximate percent south conformer and pseudorotation phase angles (29) have been determined for each residue (Table IV-1) based on measurements of coupling constants using the method of (29). Overall, with the exceptions of C17 and C10, multiplet widths and coupling constants are only consistent with a contribution less than 20% from the north (C3'-endo) conformer. For most residues, multiplet widths and coupling constants are most consistent with an average pseudorotation phase angle between 140°

and 180^0 and a relatively large amplitude of pucker (40^0). Qualitative agreement is also observed for most H1', H2' and H1', H2" cross peaks with simulated cross peaks (30) assuming a predominance of South conformer. High conformational flexibility at the 3' end of pyrimidine rich sequences, as seen here at C17, has been previously reported (31). Generally larger values of $\Sigma_{1'}$ and $J_{1'2'}$ indicate that the purine rich strand has less conformational flexibility than the pyrimidine rich strand (Table IV-2). This finding agrees with previous work on sequences containing homopurine-homopyrimidine tracts (31), but the conformational purity observed for our molecule is not seen in $d(C_3G_3)_2$, which also contains a tract of three consecutive guanines (Wolk et al., 1987 personal communication).

Figure IV-3

Sequential assignment of the 500 Mhz proton NMR spectrum using the aromatic to H1' connectivities observed in the NOESY spectrum at 250 ms mixing time. The aromatic to H1' connectivity pathway is shown for strand d(GGATGGGAG). Intranucleotide aromatic to H1' cross peaks are labeled. Assignments were confirmed using the connectivity path through the aromatic to H2', H2" region and the cross peaks between H1' and H2', H2" protons.

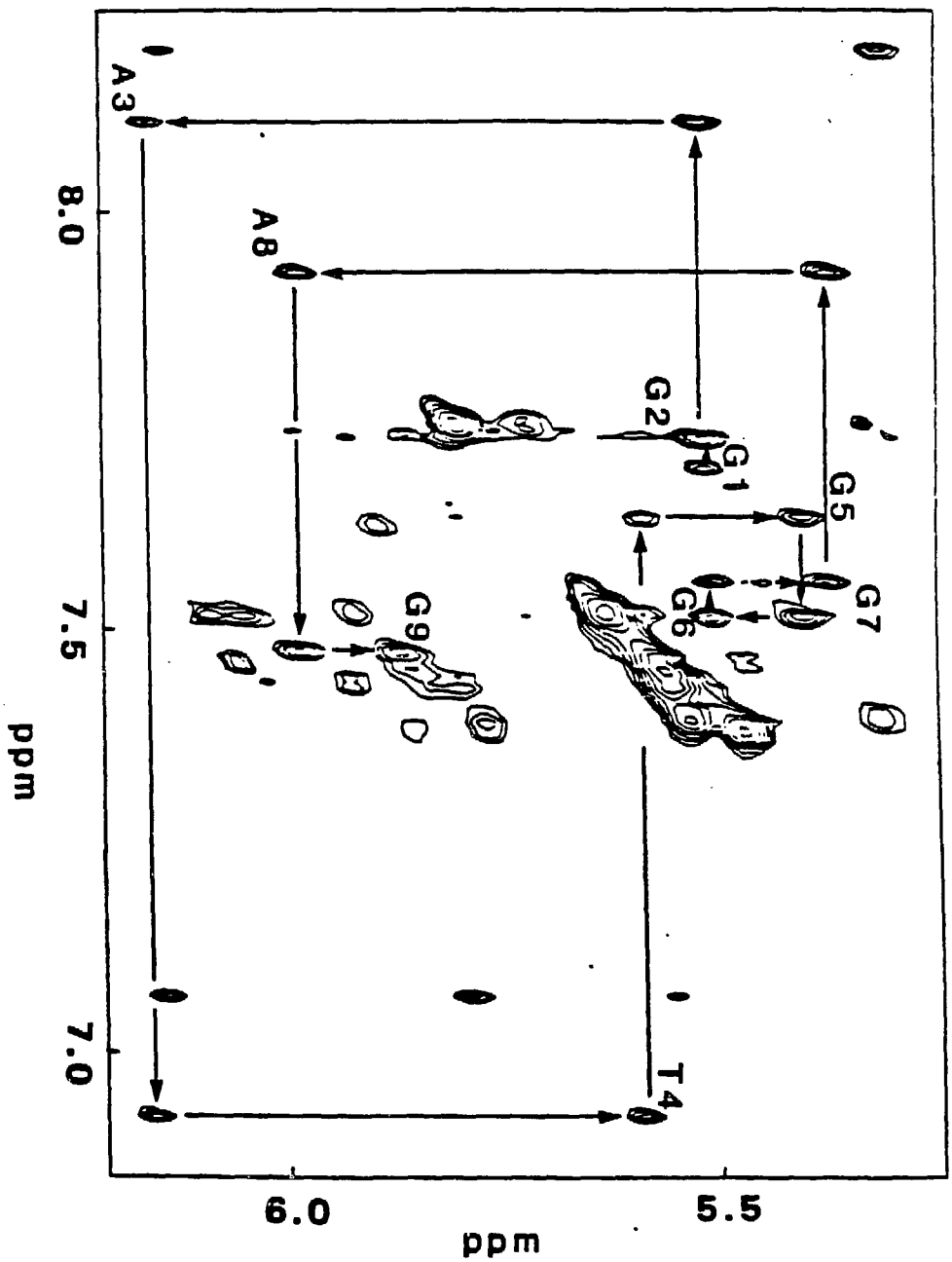
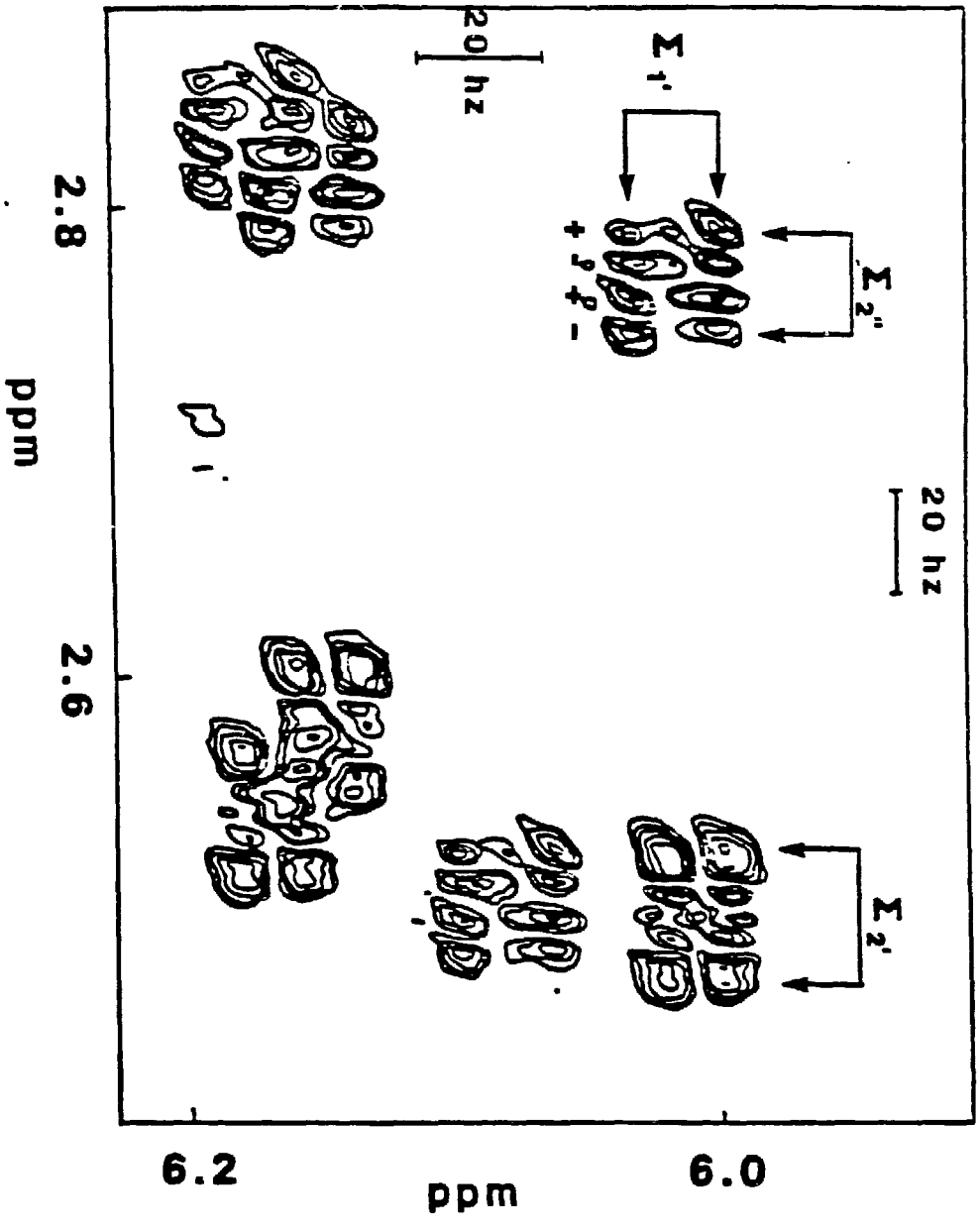
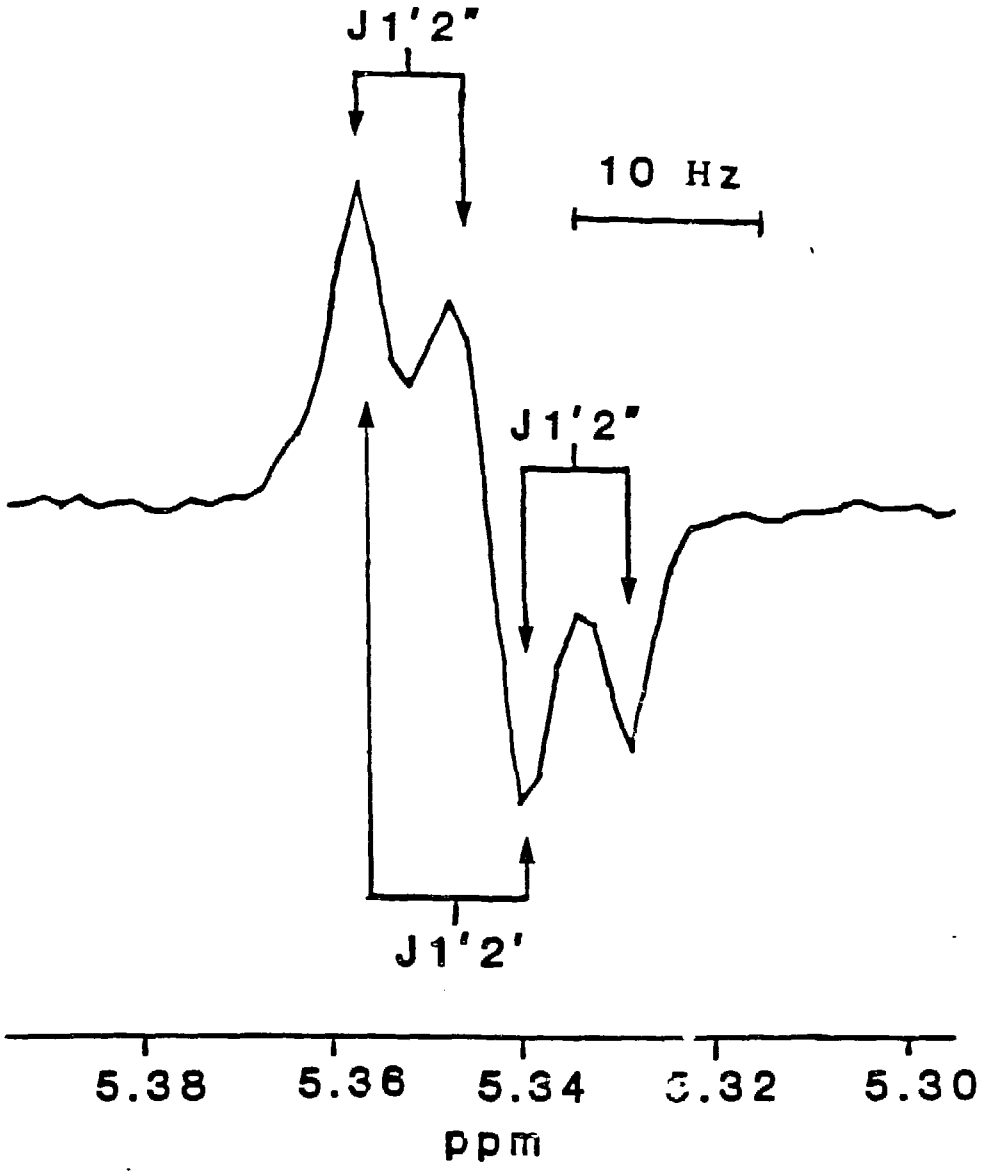


Figure IV-4



(a) Expanded view of part of the H1' to H2', H2" region of the COSY spectrum of d(GGATGGGAG)·d(CTCCCATCC). Starting from the upper left and moving clockwise the cross peaks correspond to A8 (H1' to H2"), A8 (H1' to H2'), T11 (H1' to H2"), A3, A15 (H1' to H2') (overlapped), and A3, A15 (H1' to H2") (overlapped). Note the difference in shape between cross peaks corresponding to H2' protons and those corresponding to H2" protons. Also indicated are the multiplet widths, $\Sigma_{1'}$, $\Sigma_{2'}$, and $\Sigma_{2''}$.

Figure IV-4



(b) Measurement of coupling constants $J_{1'2'}$ and $J_{1'2''}$. The 1D slice is through the H1', H2' COSY cross peak corresponding to residue C14.

Table IV-1

Chemical shifts of non-exchangeable protons in d(GGATGGGAG)·

d(CTCCCATCC) relative to TSP.

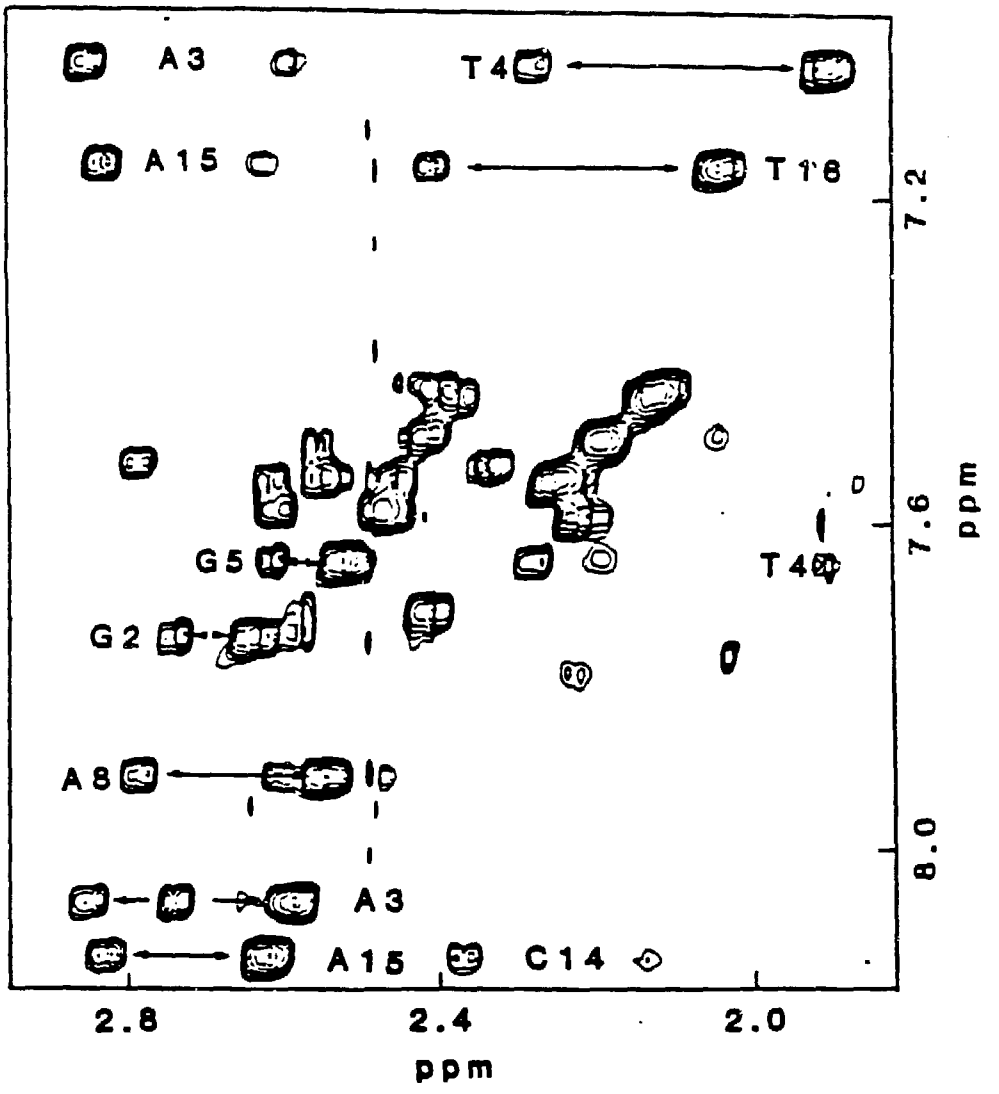
$^{5'}$ G_1 G_2 A_3 T_4 G_5 G_6 G_7 A_8 G_9 $^{3'}$
$^{3'}$ C_{18} $C_{17}T_{16}$ $A_{15}C_{14}$ $C_{13}C_{12}$ $T_{11}C_{10}$ $^{5'}$

Proton Residue	1'	2'	2"	3'	4'	6/8 (aromatic)	5/methyl
G1	5.54	2.35	2.53	4.70	4.13	7.71	
G2	5.54	2.62	2.72	4.91	4.27	7.75	
A3	6.17	2.55	2.84	4.93	4.37	8.13	
T4	5.62	1.79	2.20	4.73	4.03	6.94	1.27
G5	5.43	2.47	2.58	4.85	4.20	7.64	
G6	5.53	2.42	2.57	4.86	4.23	7.53	
G7	5.40	2.42	2.57	4.86	4.21	7.57	
A8	6.01	2.50	2.77	4.90	4.31	7.95	
G9	5.90	2.26	2.15	4.50	4.50	7.50	
C10	5.78	2.16	2.48	4.55	4.00	7.79	5.87
T11	6.08	2.19	2.51	4.80	4.16	7.53	1.59
C12	5.89	2.11	2.39	4.74	4.10	7.48	5.59
C13	5.79	2.04	2.34	4.71	4.07	7.39	5.49
C14	5.34	2.05	2.31	4.71	4.01	7.41	5.55
A15	6.16	2.59	2.82	4.89	4.30	8.20	
T16	5.82	1.95	2.35	4.73	4.06	7.09	1.34
C17	5.96	2.11	2.36	4.70	4.04	7.46	5.59
C18	6.13	2.15	2.15	4.44	4.17	7.58	5.71

Table IV-2 Scalar coupling constants (J) and multiplet widths (Σ_1) for the sugar protons of d(GGATGGGAG)·d(CTCCCATCC), together with the evaluated structural parameters (percent south conformer and approximate pseudorotation phase angle). From left, entries to the column represent H1'-H2' and H1'-H2" coupling constant, H1', H2' and H2" multiplet width, percent south conformer and pseudorotation phase angle. The amplitude of pucker is ≥ 40° for every residue. Uncertainty of the measured coupling constants and multiplet widths is ± 0.5 Hz. Uncertainty in the pseudorotation phase angle is ±25°.

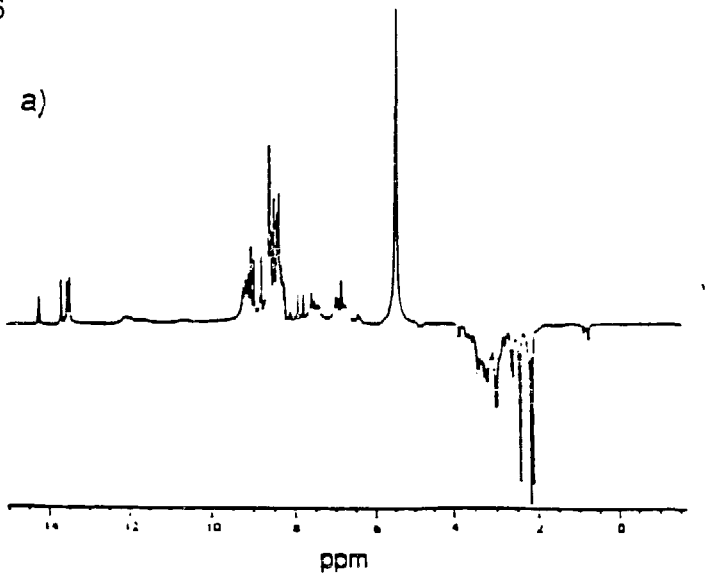
		$J_{1'2'}$	$J_{1'2"}$	$\Sigma_{1'}$	$\Sigma_{2'}$	$\Sigma_{2"}$	%S	Phase angle (°)
			Hz (±0.5)					
1	5'G	9.7	5.5	15.0	28.5	18.5	95±5	180
	G	10.6	4.9	15.4	/	19.8	100	160
	A	9.9	5.2	15.1	29.9	20.9	95±5	170
	T	9.6	5.1	14.7	30.5	21.5	90±10	150
5	G	10.9	4.7	15.3	30.0	18.8	100	/
	G	10.3	5.3	16.5	32.1	19.1	100	/
	G	10.7	4.9	15.6	28.3	20.8	100	/
	A	9.6	5.5	14.6	28.7	20.5	90±10	170
9	3'G	8.8	5.9	15.0	29.3	23.6	80±10	140
10	5'C	4.9	6.9	11.7	25.4	26.4	40±10	/
	T	9.8	4.9	15.3	30.0	20.7	95±5	170
	C	8.9	5.2	13.8	29.9	20.6	85±15	140
	C	9.0	5.6	14.6	29.4	20.9	90±10	170
	C	9.1	4.8	14.7	28.7	21.8	90±10	160
15	A	9.5	5.1	14.7	28.1	21.0	90±10	170
	T	8.8	5.8	14.7	30.3	20.9	80±10	120
	C	/	6.5	13.7	28.4	22.9	65±10	/
18	3'C	/	/	/	/	/	/	/

Figure IV-5



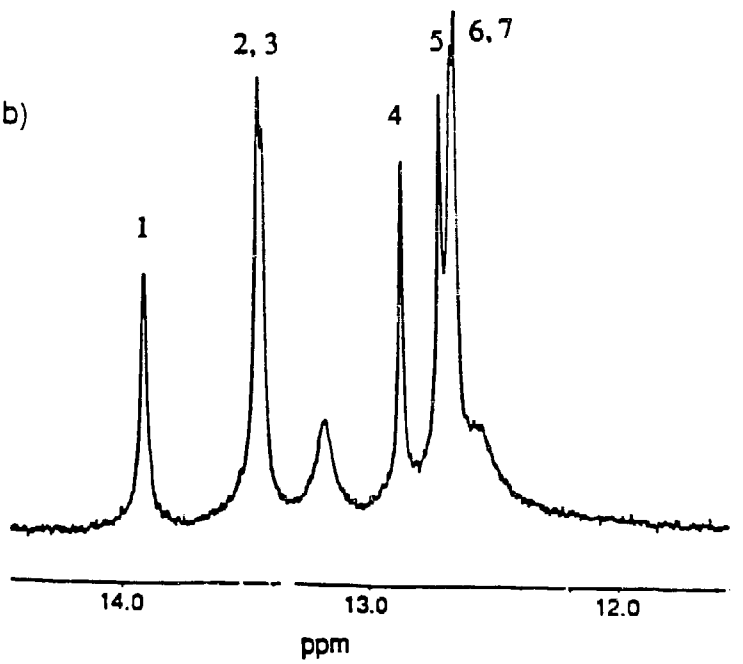
Region of the NOESY spectrum at 250 ms mixing time corresponding to aromatic to H2', H2" cross peaks. Four cross peaks are expected in this region for both A- and B-form DNA: The aromatic to its own H2', H2" protons and to the H2', H2" protons of its 5' neighbor. Cross peaks connected by arrows in the figure correspond to intranucleotide NOEs for the specified residues. For all labeled cross peaks the 2' cross peak is downfield (lower ppm) from the 2" cross peak. The fact that intranucleotide aromatic to 2" cross peaks are of comparable if not greater intensity than corresponding aromatic to H1' cross peaks (see Fig. IV-3) fixes the glycosidic angle to within the range expected for B-DNA.

Figure IV-6



a). 1-D spectrum of TFIIIA recognition fragment obtained by fourier transforming the first FID acquired during NOESY acquistion in H_2O. The "jump and return" pulse sequence was used to suppress the H_2O resonance.

b). Imino proton region of a 1-D spectrum of the same sample shown in (a). (obtained on a Bruker AM-500, 16 k complex data points, zero-filled to 64 k, 1 Hz line broadening.) Based on 1-D NOEs the assignments are 1) T11(H3), 2 & 3) T4 & T16 (H3) (overlapped), 4) G6(H1), 5) G5(H1), 6&7) G2 & G7(H1) (overlapped)



Nuclear Overhauser enhancement spectroscopy (NOESY) yields detailed information on the local structure of nucleic acid fragments since the distance between protons can be estimated from the magnitudes of the cross peaks between resonances. Distances between aromatic and sugar H2' and H2" protons are very different in A- and B-form nucleic acids (15). Intensities of intranucleotide cross peaks between resonances of these protons are very sensitive to the glycosidic angle (27). The region of the NOESY spectrum corresponding to aromatic to H2' and H2" cross peaks is shown in Fig. *IV* – 5. In this region of the spectrum, intranucleotide NOE's are generally stronger than internucleotide NOE's as expected for B-form, as opposed to A-form DNA. Intranucleotide cross peaks between aromatic and H2' and H2" are as strong or stronger than corresponding aromatic to H1' cross peaks, whereas aromatic to H2" cross peaks are more intense than those to H2' for internucleotide cross peaks. This pattern is typical of B-DNA (27); in particular, the relative magnitude of intranucleotide cross peaks indicates the glycosidic angle is near the value expected for B-DNA. Internucleotide H1' to H2' NOE's were not observed, and very weak NOE's were observed between aromatic and sugar H3', H4' protons (data not shown). The absence or weakness of intranucleotide aromatic to H3' cross peaks is another indication of B-form geometry, since these protons should be much closer in A-DNA (3 – 3.2Å) than in B-DNA (4.5 – 5Å) (27). In general, the pattern of internucleotide cross peaks is also consistent with B-form rather than A-form geometry. However, the internucleotide cross peaks show some evidence of variability in the structure as a function of sequence. This variability is suggested from the relative intensities of the NOEs between aromatic and neighboring sugar (H1', H2', and H2") protons (see Figs. *IV* – 3 and *IV* – 5).

The 1-D spectrum obtained from the first FID of the H_2O NOESY is shown in Figure *IV* – 6. Due to the jump and return sequence (21) peaks which resonate at

frequencies less than the carrier frequency appear negative, while those which resonate at frequencies above the carrier appear positive. Seven imino peaks are clearly visible, indicating that two base pairs are significantly destabilized. Only two internucleotide cross peaks are observed in the 2-D spectrum, A3 (H2) to G2 (H1) and A15(H2) to G5 (H1).

D. Discussion

Physical techniques and chemical and enzymatic probes were used as complementary tools for defining some features of the solution structure of the TFIIIA recognition fragment *dGGATGGGAG·dCTCCCATCC*. This DNA duplex of 9 base pairs and an 18 base pair DNA duplex *dGGATGGGAGCTCCCATCC* containing a palindromic repeat of the 9-mer sequence have CD spectra characteristic of B-form structure. The spectra are very different from the A-form CD spectrum of an 18-mer RNA of identical sequence, suggesting B-form base stacking in the DNA. An apparent B to A conversion of the DNA oligonucleotides over 60 – 70% TFE provides further evidence that in aqueous solution the fragment is in a conformation globally distinct from A-form. Likewise, enzymatic and chemical probe experiments confirm that this DNA sequence has properties distinct from those of A-RNA.

NMR data confirm that at the individual nucleotide level the DNA structure is B-form. The large values of the H1' multiplet widths ($\Sigma_{1'} > 14.5 Hz$) conclusively show that the sugar pucker is predominantly south for all but 2 residues. The values of coupling constants $J_{1'2'}$ and $J_{1'2''}$, together with the H2' and H2" multiplet widths, show that the sugar pucker is near 2'-endo as opposed to 3'-endo as expected for A-DNA. In general, the pattern of NOE cross peaks are consistent with a B-form overall conformation, using interproton distances for standard A- and B-DNA derived from X-ray data as references (27). Intranucleotide cross peaks between aromatic and sugar protons prove that the glycosidic angle is characteristic of B-form. Internucleotide

aromatic to sugar NOE intensities are also more consistent with B-form than with A-form geometry. This is most apparent from the relative intensities of the cross peaks between aromatic protons and neighboring H2', H2" protons. Superimposed on this general B-form pattern there are apparent local variations yet to be analyzed quantitatively. In summary, we find no evidence to support an A-form conformation for the DNA 9-mer in solution.

Since McCall et al. (9) determined that the same oligomer has an A-form structure in the crystal, several possibilities emerge. Crystal packing forces might induce a $B \rightarrow A$ transition for this TFIIIA fragment, as already suggested for other GC rich sequences (12, 13). Similarly, TFIIIA binding may induce a $B \rightarrow A$ transition in the gene, as proposed on the basis of unwinding data (32). However, addition of TFIIIA protein to a solution containing the full 54 base pair binding site induced no dramatic change in the high wavelength (270 nm) CD band (11). Furthermore, we have found that the TFE concentrations required to convert the 9-mer and 18-mer to A-form DNA are similar to those necessary for other DNA sequences. Therefore, this fragment of the 5S RNA gene does not appear significantly more labile toward A-form than random sequence DNA. On the other hand, a partial denaturation of the DNA upon TFIIIA binding would also explain the plasmid result (32), while producing little or no change in CD, since the CD spectrum of the single stranded 9-mer is qualitatively very similar to that of the 9-mer duplex (data not shown).

Local variations in the conformation of *dGGATGGGAG·dCTCCCATCC* from typical B-form may explain the enzymatic digestion data of Klug and coworkers and, perhaps, the structural features observed in the crystal. Recent combined NMR and molecular mechanics studies indicate that, within an overall B structure, values of base pair roll and slide may be more similar to A-form DNA between certain residues (16). Local A-form structural features may influence enzymatic digestion experiments (6,

7) and play a role in the specific recognition of the gene by TFIIIA. Variations of the structure within a general B-DNA geometry are suggested by variations in the relative intensities of internucleotide cross peaks (Fig. *IV* – 3). We are exploring the possible existence of local structural variations by obtaining a high-resolution solution structure for the TFIIIA fragment using distance geometry and other NMR-based methods (33). We are encouraged by the conformational purity revealed by the data in Table IV-1 since conformational equilibria tend to hinder methods for obtaining structures based on NMR.

References

1. Brown, D. D. (1980) The Harvey Lectures 76, 27 – 44.

2. Segall, J., Matsui, T. and Roeder, R. G. (1980) J. Biol. Chem. 255, 11986–11991.

3. Andrews, M. T. and Brown, D. (1987) Cell 51, 445 – 453.

4. Tso, J. Y, Van Den Berg, D. J. and Korn, L. J. (1986) 14, 2187 – 2200.

5. Sakonju, S. and Brown, D. D. (1982) Cell 46, 123 – 132.

6. Fairall, L., Rhodes, D. and Klug, A. (1986) J. Mol Biol. 192, 577 – 591.

7. Rhodes, D. and Klug, A. (1986) Cell 46, 123 – 132.

8. Hanas, J. S., Bogenhagen, D. F. and Wu, C.-W. (1984) Nucleic Acids Res. 12, 2745 – 2758.

9. McCall, M., Brown, T., Hunter, W. N. and Kennard, O. (1986) Nature 322, 661 – 664.

10. McCall, M., Brown, T. and Kennard, O. (1985) J. Mol. Biol. 183, 385 – 396.

11. Wang, A. H. J., Fujii, S., van Boom, J. H. and Rich, A. (1982) Nature 299, 601 – 604.

12. Benevides, J. M., Wang, A. H. J., Rich, A., Kyogoku, Y., van der Marel, G. A., van Boom, J., H. and Thomas, G. J., Jr. (1986) Biochemistry 25, 41 – 50.

13. Rinkel, L. J., Sanderson, M. R., van der Marel, G. A., van Boom, J. H. and

Altona, C. (1986) Eur. J. Biochem. 159, 85 – 93.

14. Gottesfeld, J. M., Blanco, J. and Tennant, L. L. (1987) Nature 329, 460 – 462.

15. Haasnoot, C. A. G., Westerink, G. A., van der Marel, G. A. and van Boom, J. H. (1983) J. Biomol. Struc. Dynamics 2, 345 – 360.

16. Nilges, M., Clore, G. M., Gronenborn, A. M., Brunger, A. T., Karplus, M. and Nilsson, L. (1987) Biochem. 26, 3718 – 3733.

17. Riazance, J. H., Baase, W. A., Johnson, W. C., Jr. , Hall, K., Cruz, P. and Tinoco, I., Jr. (1985) Nucleic Acids Res. 13, 4983 – 4989.

18. Milligan, J. F., Groebe, D. R., Witherell, G. W. and Uhlenbeck, O. C. Nucleic Acids Res., in press.

19. Kuwabara, M., Yoon, C., Goyne, T., Thederahn, T. and Sigman, D. S., (1986) Biochemistry 25, 2401 – 2408.

20. Ernst, R. R., Bodenhausen, G. and Wokaun, A., (1987) Principles of Nuclear Magnetic Resonance in One and Two Dimensions (Clarendon Press-Oxford).

21. Otting, G., Grutter, R., Leupin, W., Minganti, C., Ganesh, K. N., Sproat, B. S., Gait, M. J. and Wuthrich, K. (1987) Eur. J. Biochem. 166, 215 – 220.

22. Minchenkova, L. E., Scholkina, A. K., Chernov, B. K. and Ivanov, V. I. (1986) J. Biomol. Struc. Dynamics 4, 463 – 476.

23. Gray, D. M., Liu, J-J., Ratliff, R. L. and Allen, F. S. (1981) Biopolymers 20, 1337 – 1382.

24. Allen, F. S., Gray, D. M. and Ratliff, R. L. (1984) Biopolymers 23, 2639 – 2659.

25. Marshall, L. E., Graham, D. R., Reich, K. A. and Sigman, D. S. (1981) Biochem. 20, 244 – 250.

26. Hare, D. R., Wemmer, D. E., Chou, S. H., Drobny, G. and Reid, B. R. (1983) J. Mol. Biol. 171, 319 – 336.

27. Wuthrich, K. (1986) NMR of Proteins and Nucleic Acids (John Wiley and

Sons, Inc. New York).

28. Altona, C. and Sundaralingam, M. (1973) J. Am. Chem. Soc. 95, 2333 – 2344.

29. Rinkel, L. J. and Altona, C. (1987) J. Biomol. Struc. Dynamics 4, 621 – 649.

30. Widmer, H. and Wuthrich, K. (1987) J. Mag. Res. 74, 316 – 336.

31. Rinkel, L. J., van der Marel, G. A., van Boom, J. H. and Altona, C. (1987) Eur. J. Biochem. 166, 87 – 101.

32. Reynolds, W. F. and Gottesfeld, J. M. (1983) Biochem. 80, 1862 – 1866.

33. Patel, D. J., Shapiro, L. and Hare, D. (1987) Ann. Rev. Biophys. Biophys. Chem. 16, 423 – 453.

Appendix I: Approximations in Analyzing Melting Curves

A. Introduction

At least three major assumptions have been used in analyzing the melting curves presented in the text; (1) the transitions have been assumed to be two state (2) ΔH^0 and ΔS^0 have been assumed to be constant with temperature (3) the system has been assumed to be at equilibrium throughout the experiment.

B. The two state assumption and the constancy of ΔH^0 and ΔS^0

Calculation of thermodynamic parameters above has been carried out asssuming a so called two state equilibrium. For a bimolecular reaction, this means that the equilibrium can be described by

$$A + B \rightleftharpoons AB \qquad AI - 1$$

so that the equilibrium constant then can be written as in the derivation of the van't Hoff equation (see chapter II-B). However, if eq. $(AI - 1)$ is invalid, then the van't Hoff enthalpy does not correspond to the total enthalpy for the transition. This will be the case if there is any significant equilibrium population of intermediates. More complex models have been worked out to account for end fraying (see the introduction to Jeff Nelson's thesis). For very long polymers, the van't Hoff enthalpy corresponds to $N_0 \Delta H_{bp}$, where N_0 is the cooperative unit, and ΔH_{bp} is the enthalpy per base pair (see chapter III). Recently Werntges et al (1) found that helix-coil transitions in mismatch containing 18-mer deoxyoligonucleotides are not adequately described by a two state model. They developed a model which allows for loop formation initiated at the mismatch site. The purpose of this section is to describe methods of testing for simple two state behavior before resorting to more complex models. These tests are then used to show the validity of the all or none approximation (i. e. the two state model) for the studies presented in chapters $II - III$.

Consider again the van't Hoff equation

$$lnK_{eq} = -\Delta H^0/RT + \Delta S^0/R$$

Operationally, the fundamental assumption here is that one can define a single ΔH^0 and a single ΔS^0 which are constant over the course of the transition. Thus, if a van't Hoff plot of lnK_eq versus $1/T$ is found to be linear, then the model presented in eqs. $(II-1)$ and $(II-2)$ accurately predicts the observed behavior of the system, and the transition apparently involves only two *distinguishable* states. Other information must be used to establish what those two states or sets of states are, but the absorbance data clearly does not distinguish intermediates *in this case*, if they exist.

The validity of the all or none approximation for $dCA_3XA_3G + dCT_3YT_3G$

The first two methods for calculating van't Hoff thermodynamic parameters described in appendix II immediately suggest two tests for the linearity of the van't Hoff equation, and therefore of the validity of the two state model. One is the degree of linearity of the van't Hoff plot obtained from the T_m measured at various concentrations, the other is the degree of linearity observed in the van't Hoff plot constructed from a single transition curve. The latter can provide a direct measure of the constancy of ΔH^0 and ΔS^0 over the temperature range of the transition. A further check is the consistency of the thermodynamic parameters calculated from transition curves at various concentrations.

Fig. $II-2$ demonstrates the linear behavior of the van't Hoff plots obtained from the concentration dependence of the T_m for several deoxyoligonucleotides from the set $dCA_3XA_3G + dCT_3YT_3G$. The program Freeenergy (see appendix II) takes as input a set of fraction versus temperature data, and uses the equilibrium equation for a bimolecular reaction

$$K_{eq} = 2f/(1-f)^2C_t \qquad (AII-1)$$

$$\Delta G^0 = RTlnK_{eq} \qquad (AII-2)$$

to convert the data to the form $\Delta G^0 = RTlnK_{eq}$ versus T. Since $\Delta G^0 = \Delta H^0 - T\Delta S^0$,

Figure AI-1

Some examples of ΔG^0 versus T plots for $dCA_3XA_3G + dCT_3YT_3G$ calculated from the fraction versus temperature curves. $X\,Y$ are (a) $(G\,C)$, (b) $(I\,C)$, (c) $(T\,C)$, (d) $(G\,T)$. Only very small deviations from linearity are observed, suggesting that ΔH^0 and ΔS^0 are nearly constant over the temperature range of the transition.

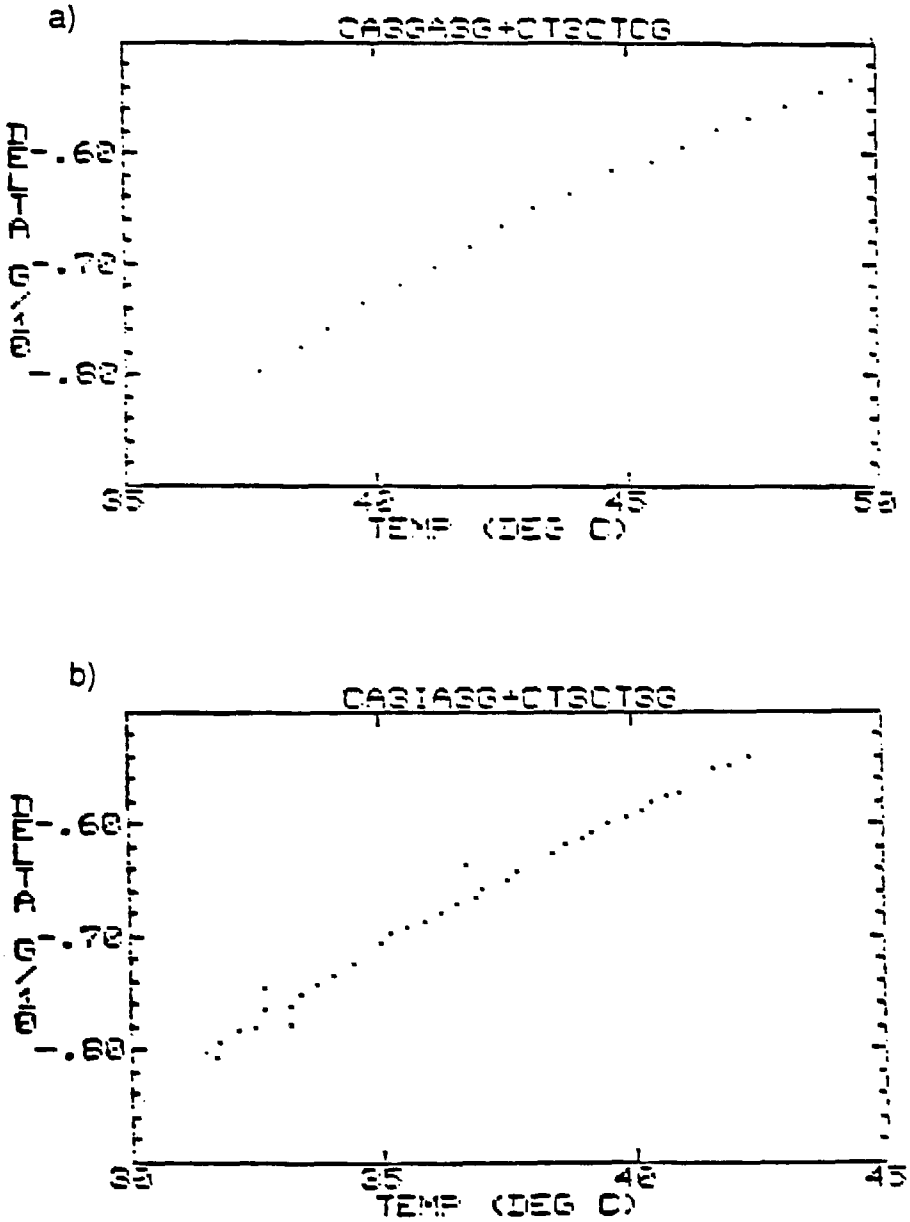


Figure AI-1

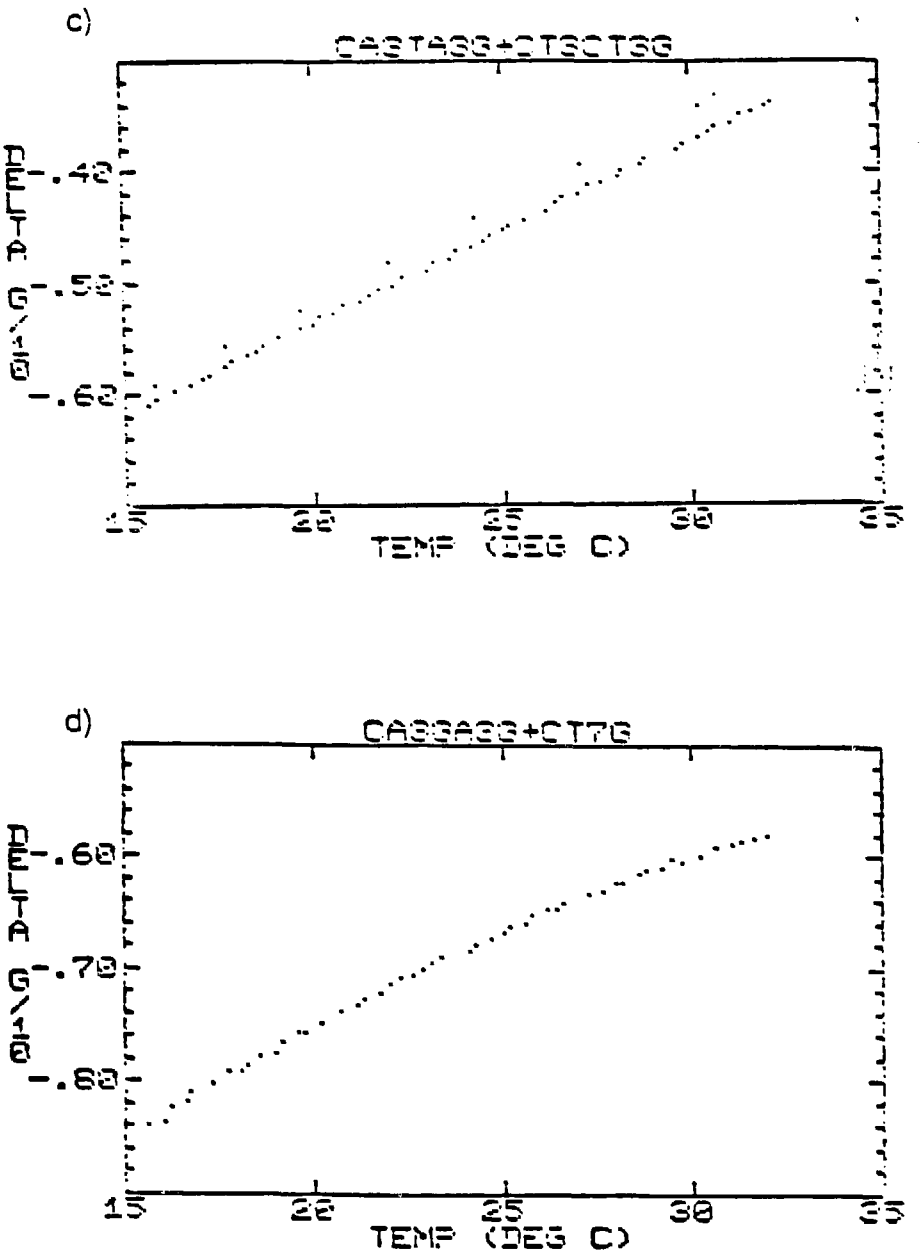
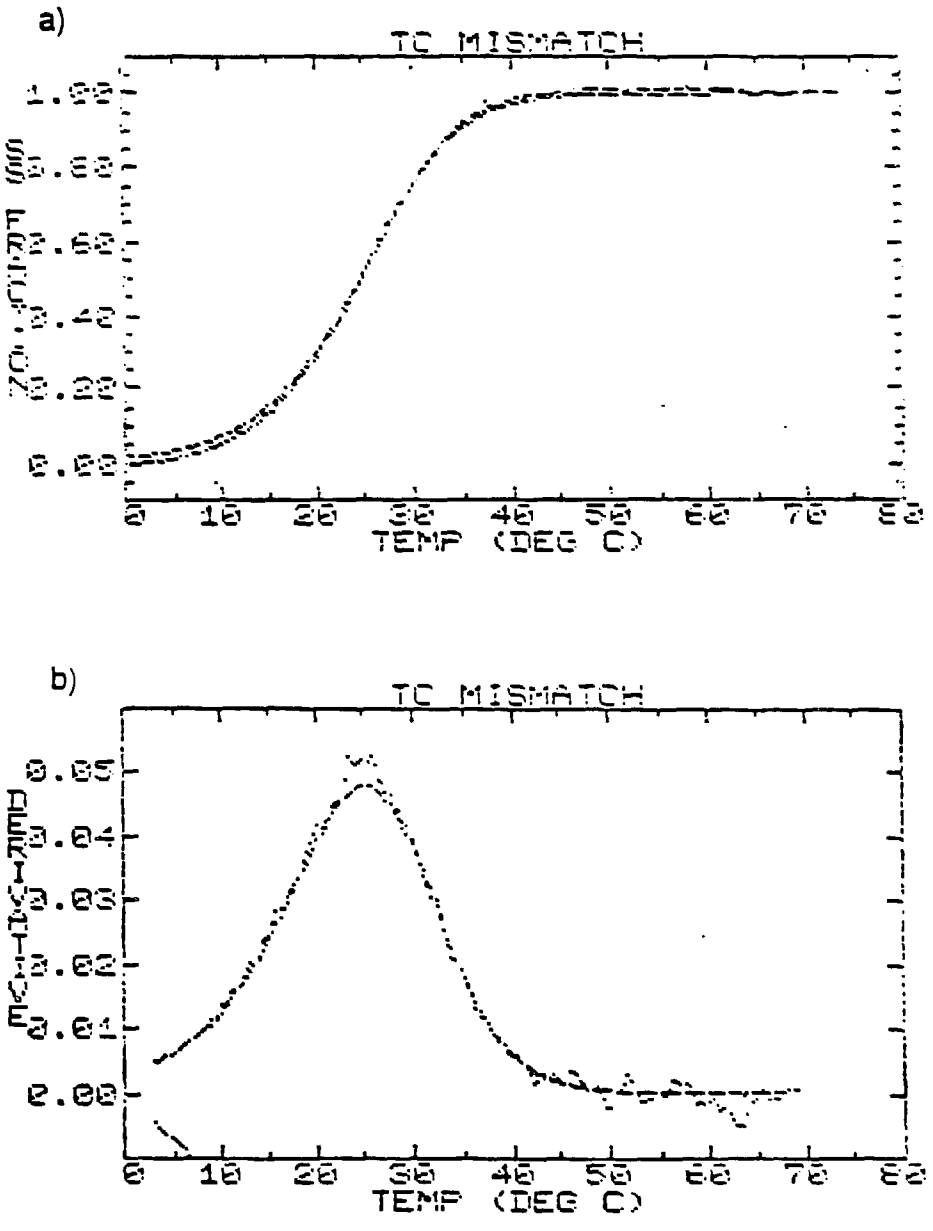

c)
CAGTAGG+CTGCTGG
DELTA G\+0
-.40
-.50
-.60
15
20
25
30
35
TEMP (DEG C)
d)
CAGGAGG+CTTG
DELTA G\+0
-.60
-.70
-.80
15
20
25
30
35
TEMP (DEG C)

Figure AI-2

Experimental and simulated single stranded fraction (a) and derivative (b) versus temperature curves are superimposed. The simulations assumed a two state model. The molecule is $dCA_3TA_3G+dCT_3CT_3G$.



linearity of this dataset is an indication that ΔH^0 and ΔS^0 are constant over the range of the transition and that the transition is therefore two state. Fig. $AI-1$ shows examples of sets of plots of ΔG^0 versus T for several sequences from the series $dCA_3XA_3G+dCT_3YT_3G$. For every sequence examined in this manner, the plots were clearly linear and values calculated for ΔH^0 and ΔS^0 are similar at different concentrations. It is interesting that what little curvature appears in Fig. $II-1$ is most apparent for duplexes that involve Watson-Crick pairing- but even that curvature does not appear to be significant. Thus there is no evidence for the set of sequences used in this study that end fraying, mismatch loop premelting or any other intermediate forms are making a detectable contribution to the equilibrium. Figure $AI-2$ shows an example of a simulated bimolecular two state transition curve using the ΔH^0 and ΔS^0 values reported in chapter II for $dCA_3TA_3G+dCT_3CT_3G$ along with an experimental curve for the same molecule. The excellent agreement between the two curves contrasts with the findings of Werntges et al, who were unable to simulate melting curves for 18-mers using a two state model.

Some additional ideas on testing the two state approximation

Another straightforward method of testing the all or none approximation is to measure the enthalpy calorimetrically and compare it to the van't Hoff enthalpy. For an all or none transition, $N_0 = N$ = the number of residues. Then $\Delta H^0 = N\Delta H_{bp}$, where ΔH_{bp} is the calorimetrically measured enthalpy of the transition for a single residue. If the cooperative unit is less than N, or if intermediate states are involved (2) $\Delta H_{vh} < N\Delta H_{bp}$. However, calorimetry requires far more material than optical studies. Additional measurements using other optical probes, such as CD or absorbance at a different wavelength may detect intermediates not clearly seen at $260nm$. Yet trying to use additional measurements to prove or disprove two-state behavior is not always practical, as an isosbestic point may not be simple to identify. Moreover, conditions

may be difficult to duplicate, so that even for a two state transition one may measure a slightly different T_m using absorbance and CD or absorbance at two wavelengths.

If enough material is available, NMR is potentially the most sensitive probe for intermediates, and offers the best hope for identifying the specific structures involved in the transition. NMR can be useful provided that 1) NMR conditions (i. e. mM or higher strand concentrations) do not change the nature of the transition 2) A significant temperature range exists for each conformer or intermediate involved in which that conformer is well populated and not in intermediate exchange with another form. If conformers are in slow exchange, distinct peaks corresponding to each form will be visible. This is often the case for a duplex to hairpin or B to Z equilibrium. In fast exchange, one will see a single peak at a chemical shift corresponding to an average between that seen for the two forms. 3) The total number of peaks does not prohibit resolution of individual peaks in either a 1-d or a 2-d experiment. In other words, NMR is not always the answer. On the other hand, often NMR is the most powerful method for revealing molecular details of a transition (3) and for doing thermodynamic measurements as well (4, 5).

If none of the above suggestions seem practical, and there is some reason to believe that a transition curve may not be two state, one should consider trying to fit the curve using simulations. This approach is probably most effective when used in combination with sieving column experiments or non denaturing gel electrophoresis in order to establish the (monomer-dimer-trimer) nature of the sample at selected temperature and concentration conditions. The conclusions drawn from these experiments should be confirmed by testing the concentration dependence of the melts (i. e. monomer-monomer or n-mer to n-mer transitions should be concentration independent, other types of transitions should have a concentration dependence). Then programs on the diskette Simdat or modifications thereof can be used to simulate transition curves

involving suspected intermediates for given input values of ΔH_I, ΔS_I (the difference in standard enthalpy and entropy between the initial and the intermediate state) ΔH_D, ΔS_D (differences in parameters for the initial and final state), and d (the difference between the absorbances of the intermediate and initial states divided by the difference between the absorbances of the final and initial states). For example, the program *threest* simulates curves for the reaction

$$A \rightleftharpoons B \rightleftharpoons C$$

Here five parameters are fit to a single curve, so the more information obtained from other sources, such as ΔH^0 values from nearest neighbor data bases (6, 7, 8), melts under different salt or concentration conditions, or melts using analog molecules, the better the chance of obtaining a believable fit to the data.

Statistical mechanical models

If one suspects multiple intermediates, it may be necessary to resort to statistical mechanical models such as Ising models described in chapter III and in (9). Werntges et al. (1) use a model based on an algorithm by Poland (10) which can account for various degrees of domain melting in a helix to coil transition. Again, many parameters are involved, so reliance on supplementary sources of data will probably be required using this approach.

C. The equilibrium assumption - heating rates

If the heating rate in a melting experiment is slow enough to maintain equilibrium throughout the transition curve, then 1) The T_m, shape of the curve, and measured thermodynamic parameters should be unchanged at a significantly slower heating rate and 2) after the experiment, when the temperature is returned to a point near the middle of the transition and maintained indefinitely, the absorbance will settle to whatever value it had at that temperature during the original experiment. Both tests were applied to several of the transition curves in chapter II. Since the transition

curves presented in chapter III were obtained at a heating rate of 0.025^0C per minute, it was impractical to apply the first criteria. However, the following experiment was performed with the shortest polymer length sample, which was shown in chapter III to have the slowest kinetics. The sample was heated to a temperature in the lower middle range of its transition and allowed to incubate while absorbance at 295 nm was recorded overnight. In the morning the absorbance had been constant for a few hours, ignoring slight instrument drift. The temperature was then increased by one degree in approximately one minute, and the sample incubated for three hours. After the temperature jump, the absorbance settled to a constant value after about 40 minutes, indicating that for the sample with the slowest kinetics the heating rate used was marginally slow enough to maintain equilibrium.

The effect of heating too fast on measured thermodynamic parameters

As mentioned above, the use of a two state model in the presence of intermediates will result in an erroneously low measured van't Hoff enthalpy difference between the initial and final states. No such simple generalization can be made for measurements on systems that are not at equilibrium. Consider, for example, the simple case of a first order process

$$A \rightleftharpoons B$$

which can be described by a forward rate constant $K_b = K_0 \times exp^{-H^*/RT}$ where H^* is the activation enthalpy and K_0 is a temperature independent pre-exponential factor. Kinetics will be slow if K_0 is of very small magnitude and/or H^* is of large positive magnitude. If both K_0 and H^* are small, the reaction rate will be slow and will increase relatively little with increasing temperature. Then the apparent or measured fraction of B will lag behind the equilibrium fraction by an increasing amount over the range of the transition, leading to a broader transition with a smaller df/dT. Since the $\Delta H_{vh} \propto -T_m^2 df/dT$ and T_m as measured in Kelvin is not likely to increase dramatically,

the measured ΔH_{vh} will be of smaller magnitude than that measured at a heating rate slow enough to maintain equilibrium. On the other hand, if K_0 is small and H^* is large, then a lag will be observed in f at the beginning of the transition, but the kinetics will increase dramatically as the temperature is increased. Eventually the transition rate will surpass the heating rate and the observed fraction will rapidly reach equilibrium at the end of the transition. An anomously large slope and measured ΔH_{vh} will result. The abruptness of the completion of the transition may also produce a 'pointed' shape at the top of the curve.

An example of a system in which a fast heating rate causes an anomolously low measured ΔH_{vh} is the $B \rightleftharpoons Z$ transition in poly d(^{5me}CG) described in chapter III. Van't Hoff enthalpies obtained from experiments run at 0.25 ^0C per minute are approximately 75% of those obtained for identical samples heated at a rate of 0.025 ^0C per minute (table AI-1). Van't Hoff enthalpies reported for a similar system at a heating rate of 0.1 ^0C per minute (11) are also lower than those reported in chapter II. In contrast to this behavior, van't Hoff enthalpies obtained from UV melting experiments on a duplex to hairpin transition at 1^0C per minute were surprisingly large (12). Evidence is now accumulating that duplex to hairpin transitions are very slow (Joseph Puglisi, Deborah Kallick, personal communications). It is probable that the apparent van't Hoff enthalpy reported for this system by the authors of (12) is higher than the actual enthalpy for the duplex to hairpin transition. The melting curves presented in (12) show signs of the 'pointed hump' expected to be characteristic of systems with a very high activation enthalpy.

References

1. Werntges, H., Steger, G., Riesner, D. and Fritz, H-J. (1986) Nucleic Acids Res. 14, 3773 – 3790.

2. Cantor, C. R. and Schimmel, P. R. (1980) Biophysical Chemistry, W. H. Freeman

and Co., San Francisco, section 21 – 4.

3. Davis, P. W., Hall, K., Cruz, P., Tinoco, I., Jr. and Neilson, T. (1986) Nuc. Acids Res. 14, 1279 – 1291.

4. Wemmer, D. E. and Benight, A. S. (1985) Nuc. Acids Res. 13, 8611 – 8621.

5. Hartel, A. J., Lankhorst, P. P. and Altona, C. (1982) Eur. J. Biochem. 129, 343 – 357.

6. Ref. 4, chapter II.

7. Ref. 12, chapter II.

8. Ref. 14, chapter II.

9. Poland, D. and Scheraga, H. A. (1970) Theory of Helix-Coil Transitions in Biopolymers, Academic Press, New York.

10. Poland, D. (1974) Biopolymers 13, 1859 – 1871.

11. Chaires, J. B. and Sturtevant, J. M. (1986) Proc. Natl. Acad. Sci., USA 83, 5479 – 5483.

12. Marky, L. A., Blumenfeld, K. S., Kozlowski, S. and Breslauer, K. J. (1983) Biopolymers 22, 1247 – 1257.

Appendix II: Introduction to and Instructions for Use of Melt Programs

A. Introduction

Options and limitations

This appendix presents a description of the software currently used in the lab for recording and analyzing melt data along with step by step instructions for its use. The data analysis programs described here are set up for use on the Apple IIE. Chaejoon Cheong has set up a similar system on the Vax. The Apple IIE system as presently constituted can be used to convert absorbance versus temperature data to fraction versus temperature-then convert the fraction versus temperature data to 1) smoothed fraction, 2) K_{eq}, 3) $RTlnK_{eq}$, or 4) derivative of fraction versus temperature, and finally to calculate van't Hoff thermodynamic parameters by at least three different methods. The various steps in this process are diagrammed in figure AII-1. The different methods for calculating van't Hoff parameters are explained in section C of this appendix.

The main defects of this system are 1) The fact that the data are recorded in BASIC while the analysis programs are written in PASCAL. Thus it is necessary to copy the original file from a diskette formatted in BASIC onto a diskette formatted in PASCAL. 2) The Apple IIE has limited memory capacity, which limits the precision attainable. 3) The Apple IIE has limited buffer capacity. This limits the size of an individual program or datafile. The result is a proliferation of small programs for performing specific functions. Though it is desirable for purposes of flexibility to have such small programs it would also be convenient to have one megaprogram which can perform most of the steps required. That would spare the necessity of writing a new file to disk after each operation.

Booting the Apple

Place the boot diskette in drive #4 and either 1) turn the computer on, or 2)

simultaneously hit the ctrl and reset keys.

Formatting Diskettes

At least two floppy disks are required for the recording and analysis of melt data. One diskette, formatted in BASIC, is used for recording data. The second diskette, formatted in PASCAL, will be used for data analysis.

Formatting diskettes in BASIC

Place the diskette labeled DOS 3.3 System.Master (it has a white label and sits in a white box) into disk drive number 1 (the boot drive, also known as #4 to Apple PASCAL), and boot the computer. After the computer gives the welcome message and the prompt appears, remove the System.Master disk and replace it with your own blank diskette. Now type

INIT HELLO

then return. When the formatting is done, the drive will stop making a funny noise, the red light on the drive will go off, and the prompt will return to the screen.

Formatting diskettes in PASCAL

Place the diskette Apple 1 in drive #4, Apple 3 in drive #5 and boot. A menu will appear at the top of the screen. Type "X" for execute. You will be asked which program you want to execute. Type

APPLE3:FORMATTER

then return. You will be asked which diskette you want formatted. Remove Apple 3 from drive 5 and replace it with your blank disk then type 5. The drive will make a bizarre noise until the formatting is complete.

B. Data Collection

The program for data acquisition was written by Dr. Phillip Cruz. Boot the Apple with the diskette labeled Melt Data Acquisition in drive #4 and your data disk in drive 5. A menu will appear in which the first entry will be 1) Gilford Melt. Type

1 (DO NOT HIT RETURN). You will be asked for a disk filename. Choose a name (preferably short - you will be asked for comments later) that starts with a letter and contains no spaces, then end it with a comma followed by D2 (no spaces).

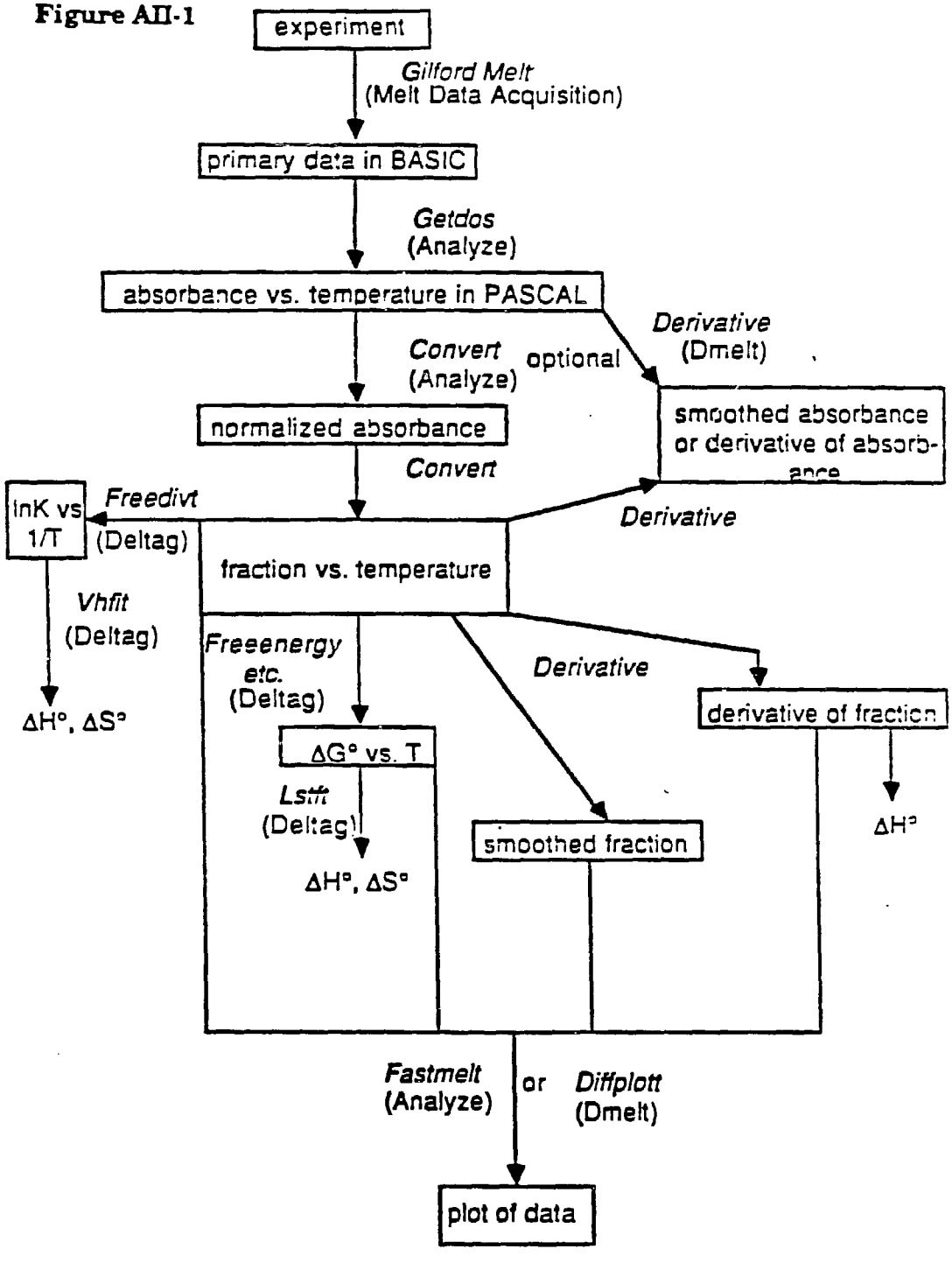
filename,D2

The D2 tells the program to write the file to drive #2. Hit < *CR* >. Now you will be asked for comments. Type whatever you like but avoid commas, colons or periods. The next question is the number of cuvettes. This will include the number of samples plus one (the reference). In other words, if you have two samples, one in position two and one in position three, you should type 3, < *CR* >. Now you will be asked whether or not the printer is on line. Data acquisition will begin when 1) you have answered 'y' to this question 2) the cuvette positioner is on 'auto', and 3) the buttons next to the cuvette positions to be used have been pressed. During the first cycle, the program will set up two plots for each cuvette position (not including position 1). After the first cycle, you can toggle the display between the graphic and the numerical display of the data with the esc key.

It is important to keep in mind that there is an upper limit to the capacity of the data analysis programs of 250 data points. This means that if you want to avoid post-editing your data, you must adjust the temperature range, heating rate, and dwell time accordingly.

You can stop data collection at any time by hitting ctrl A **BUT**, if you do this while data is being written to the diskette (this happens whenever the number of points is a multiple of ten-at this time the red light will come on on drive 5) the printer will not plot the graphic display of the melts which you have just seen on the screen. The data in this case will still be stored in the file and can still be plotted by using the program *getdos* (section C) then *fastmelt* or *diffplot* (section E).

Figure AII-1


experiment
Gilford Melt
(Melt Data Acquisition)
primary data in BASIC
Getdos
(Analyze)
absorbance vs. temperature in PASCAL
Derivative
(Dmelt)
Convert
(Analyze)
optional
normalized absorbance
smoothed absorbance or derivative of absorbance
Convert
Derivative
lnK vs 1/T
Freedivt
(Deltag)
fraction vs. temperature
Vhfit
(Deltag)
ΔH°, ΔS°
Freeenergy etc.
(Deltag)
Derivative
derivative of fraction
ΔG° vs. T
Lsfit
(Deltag)
ΔH°, ΔS°
smoothed fraction
ΔH°
Fastmelt
(Analyze)
or
Diffplott
(Dmelt)
plot of data

C. Converting the data to fraction versus temperature

Converting a BASIC file to a PASCAL file using getdos

These programs have been compiled on the Analyze diskette by Steve Wolk, who wrote the *convert* program. Place the Analyze diskette in drive 4 and your data diskette in drive 5, then boot. You will see the Apple Pascal menu appear at the top of the screen. Type 'x' for execute. When you are asked for a filename type *getdos*. This program will read a file from a diskette that is formatted in BASIC and write the same file to a diskette that is formatted in PASCAL. As mentioned in the introduction, this step is necessary since the data recording program is written in BASIC whereas the analysis programs are in PASCAL. The first questions *getdos* asks you are the source and destination units. To both these questions answer '5'. The program then asks you for a filename. You type the filename followed by < *CR* > (#5 is assumed this time). After the catalog and file are read you are told to insert the destination disk. Remove your BASIC formatted data diskette and replace it with your PASCAL disk, then hit < *CR* >. Probably the screen will show the message "One Moment Please". Finally you are asked whether you want to repeat the process with another file.

Converting data to fraction versus temperature using convert

You now have your data in the form of a PASCAL textfile, and you are ready to use the program *convert* to convert 1) your PASCAL textfile to a datafile. 2) your absorbance versus temperature data file to a normalized absorbance versus temperature file (this step is optional). 3) your absorbance or normalized absorbance versus temperature file to a fraction versus temperature file. All of these operations must be performed in the order listed above.

a. converting a textfile to a datafile and normalizing

Type 'x' for execute and *convert* for the file to be executed. You will be presented with a menu with 8 options. Type '2', then < *CR* >, to input your PASCAL textfile.

When you type the filename, *be sure* to prefix it with #5:

#5:filename

then < *CR* >. For a typical textfile it will take about three to five minutes to input. The slowness of working with textfiles makes it desirable to convert to a datafile. When inputting is completed and the menu reappears, type 6 and < *CR* > to output the array as a datafile. Again, when you type the output filename, *prefix it* with #5:.

To normalize the data, you should have a printout of your raw data available for reference. Beforehand you should decide 1) At what temperature you want to normalize the data to one and 2) Whether and what you want to subtract as a background absorbance due to the cell, buffer, etc. When you have decided all that, and you have the menu on the screen and your data in the array, type 4 and < *CR* >. Respond as requested with your normalization temperature and background absorbance for channel 2. After each response you will be asked whether or not your input is correct, so if you screw up just hit < *CR* > and type 'n' when asked about correctness to repeat the process. After the background absorbance you will be asked for the absorbance at the normalization temperature. All of the absorbance data in channel 2 will be divided by the number that you input here (after both the original data point and the normalized absorbance have been corrected for the background absorbance). As soon as you type 'y' to say that the absorbance at the normalization temperature is correct the data in channel 2 are normalized. The process is repeated for channels 3 and 4, if present. After the last channel is normalized the menu returns. Again outputting the data to a file for safekeeping is recommended, using option 6 as above (remember #5: before the filename!).

b. choosing baselines

In order to convert the absorbance or normalized absorbance to fraction versus temperature one must input baselines for the single stranded and double stranded

absorbance. There is no standard formula for choosing baselines. The best method depends on the transition curve in question and, unfortunately, the judgment of the person doing the analysis. One method that often proves useful is to obtain the lower baseline from a linear least squares fit to about 10 data points in the (hopefully) linear low temperature region and the upper baseline from a similarly linear high temperature region. Of course, in addition to linearity, the other requirement for a baseline is that you can be certain that the system is completely duplex (or whatever your low temperature structure is) for all of the lower baseline points and completely single stranded (or Z-form, etc.) at each upper baseline point. When several transition curves are available with, for example, different strand concentrations or different polymer lengths, it may be advantageous to use the same upper baseline for each normalized absorbance versus temperature curve. The slope for all lower baselines in this case can be obtained by averaging the slopes obtained from several curves unless there is evidence of an exceptional degree of aggregation (manifested in a dramatically larger slope) for certain curves. However, the $0°C$ intercept should be chosen separately for individual curves, based on the absorbance at the highest temperature at which the sample can be said to be greater than 99% in the low temperature form.

Low melting transition curves which lack a lower baseline simply cannot be used to calculate parameters with the precision normally desired. At best, one may obtain an estimate of the fraction versus temperature by assuming a flat lower baseline, using a baseline obtained from another curve (i.e. at higher concentration), or doing some sort of curve fitting. Data simulation programs are already available (see section F of this appendix), and it would be a simple matter to expand these programs to iteratively fit parameters, though the lower baseline will have to be one of the parameters fit.

Choosing baselines is actually the most troublesome aspect of this type of experiment. For all but the most straightforward data, there is probably no way to avoid

repeating the analysis with several choices of baselines in order to determine which method yields the most consistent results. This does not imply that the process is arbitrary, since whatever method is chosen should be used as consistently as possible for any given set of data.

c. converting to fraction versus temperature

When you choose option 5 to convert the data to fraction versus temperature you will be asked 'which channel' and whether you want to change any of the baseline parameters. If you type 'y' you will be asked which parameter you want to change. The choices are ; A, the lower baseline $0°C$ intercept; B, the lower baseline slope; C, the upper baseline intercept and; D, the upper baseline slope. Note that these baselines may correspond to states other than single and double strands for various types of transitions. Type your choice of parameter; $< CR >$; your new value; then $< CR >$ again. If you make a mistake you can repeat the process. As soon as you type 'x' for no changes the conversion is executed and the menu returns. Again it is advised that you output the data to the screen using option 8 in order to check for mistakes, and that you save the data in a file using option 6.

D. Obtaining thermodynamic parameters

Once you have the data on disk in the form of a fraction versus temperature file you can obtain thermodynamic parameters by one of at least three methods for a bimolecular transition and either of at least two methods for a unimolecular transition.

Bimolecular transitions

The three methods available are 1) construct a van't Hoff plot from the concentration dependence of the T_m, extract ΔH^0 from the slope, ΔS^0 from the $T = 0K$ intercept and ΔG^0 from $\Delta G^0 = RT\ln C_t/a$ where C_t is the total strand concentration at which $T_m = T_0$, (T_0 is the temperature of the reference state) and $a = 4$ for a non self complementary duplex to single strand transition or $a = 1$ for a self complementary duplex

to single strand transition (see chapter II). Note that, in principle, you don't need to choose points at the T_m. You could calculate lnK_{eq} for any value of the fraction, but signal to noise ordinarily is maximal at the T_m. 2) Construct a van't Hoff plot from a single transition curve by converting the fraction at several temperatures to lnK_{eq}(see chapter Appendix I). 3) Calculate the derivative of the fraction versus temperature. Then ΔH^0 can be obtained from

$$\Delta H^0 = -6RT^2 df/dT$$

(Gralla and Crothers (1973) J. Mol. Biol. 78, 301-319.) where T is the temperature and df/dT the derivative at the transition midpoint. Note that using this method one can only obtain ΔH^0, whereas ΔS^0 and ΔG^0 cannot be measured from the derivative curve.

Option 1 (concentration dependence)

Method 1 can be performed simply by reading off from the fraction versus temperature files the temperatures at which the fraction=1/2. Concentrations are calculated using the absorbance read at some temperature at which the system is known to be all single stranded, using the upper baseline(see section II-C) to extrapolate the absorbance to 25^0C, then dividing by the extinction coefficient (calculated using option 5 on the Melt Data Acquisition disk), the path length, and appropriate correction factors. The linear least squares fit of the concentration points can be performed using a pocket calculator or any one of the least squares fitting programs available on the Vax and the Apple.

Option 2 (lnKeq from fraction)

The programs on the diskette named Deltag are available for calculations using method 2. The program *freeenergy* (or *freeesc* for self complementary duplexes) takes a fraction versus temperature file as input, converts it to ΔG^0 versus temperature using equations $AI-1, AI-2$ then outputs the data to a file. After booting the Apple with the

Deltag disk and Xecuting the file *freeenergy*, you are asked for a filename. Answer with

#5:filename

then < CR > as before. The concentration should then be input in moles/liter followed by return. When the program asks you whether or not the value of the concentration is correct, it may present you with your number followed by a random letter or digit for some mysterious reason. Don't be alarmed, it doesn't seem to matter. Just hit < CR >. The next question is whether or not you want the data echoed to the screen. If you type 'y' you will see three columns of data. The first column is the temperature, the second column lists the equilibrium constant, K_{eq}, and the third column lists $\Delta G^0/10^4 = -RTlnK_{eq}/10^4$. ΔG^0 is divided in order to make plotting convenient. If the fraction is less than 0.15 or greater than 0.85, then K_{eq} is set to a predetermined value and will be ignored later. After K_{eq} and ΔG^0 have been calculated the program asks whether you want to output the data. You must output it to a file in order to obtain ΔH^0 and ΔS^0. When you type the filename remember #5:. The output file will have three channels, channel two is K_{eq} versus temperature and channel three is $\Delta G^0/10^4$ versus temperature. Only those data points for which $0.15 < f < 0.85$ are output.

You now have a file which contains $\Delta G^0/10^4$ versus temperature, which, if you have an all or none transition, should be linear according to $\Delta G^0 = \Delta H^0 - T\Delta S^0$. The program *lstft* on the Deltag diskette is available to do a linear least squares fit to the output of *freeenergy*. Type 'X' to execute *lstft*. Answer the first question, whether or not you want to input data from a file 'y' then < CR >. Making sure that the disk containing your ΔG^0 versus T file is in drive #5, type #5:filename. The program will tell you how many data points are in the file and will ask you whether or not you want to add data from an additional file to the fit. If you type 'y' the process is repeated, otherwise the next time you press < CR > a linear least squares fit is made to channel

3 and you will be given the slope and the intercept. ΔH^0 is 10^4 times the intercept (in calories/mole) and ΔS^0 is 10^4 times the slope (in eu).

The procedure for using *freeesc* is identical. The only difference is in the manner in which K_{eq} is calculated from f for self complementary duplexes.

Also available is the program *freedivt*, which outputs lnK_{eq} versus $1/T$ and which operates exactly like *freeenergy* and *freeesc*. Then when *vhfit* (another version of *lstft* on Deltag) operates on the output of *freedivt*, the slope corresponds to $\Delta H^0/10$ (in cal/mole) and the intercept to ΔS^0 times 10^3 (in eu). The advantage of *freedivt* is that one does not need to extrapolate to a temperature of absolute zero in order to obtain ΔH^0, while the output of *freeenergy* has the advantage that when plotted it covers a standard temperature range.

Option 3 (taking the derivative)

For bimolecular transitions, only ΔH^0, not ΔS^0 or ΔG^0 can be obtained in this way. Boot the computer with the Dmelt disk in drive 4 and the disk containing your f versus T file in drive 5. Type 'F' for filer when the Apple Pascal menu appears at the top of the screen then 'G' for Get. You are asked 'which file?' and you should type 'derivative' followed by $< CR >$, then 'Q' to quit the filer. For some reason the program aborts if you do not go through this procedure before executing. Now type 'R' to run *derivative* (make sure you have quit the filer!). You are asked the name of the file to be loaded-you should type

#5:filename

The next question is the number of points to be used for a fit. This program chooses a local set of points at each specified temperature (you will decide at what temperatures to take the derivative below) and uses this set of points to do a least squares fit to a second order polynomial. The number of points chosen for each least squares fit is determined by your answer to this question, followed by $< CR >$. You are then

asked for a temperature interval between smoothed points. This will determine how many data points you will output, and so you might consider that this program will take some time to compute (maybe about 20 minutes for 200 data points and using a 15 point fit for each data point). Next you are asked for a channel number (2, 3, or 4 for absorbance or normalized absorbance, 1 for fraction versus temperature as input) and then a starting temperature. The program will search your dataset until it finds the first data point for which the temperature is greater than or equal to your specified starting temperature. If it cannot find enough data points in the range of your starting temperature to do a fit, it will simply repeat the question-'starting temperature=?'. When the program is satisfied with the starting temperature it asks for a final temperature, and finally whether or not you want the data echoed to the screen. If you answer 'y' you will be presented with three columns of data. The first column is the temperature, the second column is the smoothed data; $AT^2 + BT + C$, where A, B, and C are the coefficients of the second order polynomial obtained from the least squares fit. The third column corresponds to the derivative; $2AT + B$. When the final temperature is reached you are asked whether you want to output the *smoothed* data, that is, what you have just seen in column two. Remember #5: when you output the filename and likewise if you decide to output the derivative data (the next option).

Unimolecular transitions

In this case using method (1) of section C is of course not an option, since there should be no concentration dependence. However methods 2 and 3 are still available.

option 2

Here the procedure is the same as it is for bimolecular transitions using this method except that the program *freeuni* is now used to obtain ΔG^0 versus T. Freeuni operates like Freeenergy except that there is no need to input the concentration and the program now calculates K_{eq} based on the equation

$$K_{eq} = f/(1-f)$$

for a unimolecular transition.

option 3

Again use the program *derivative* as in part D of this appendix, but ΔH^0 is now obtained from

$$\Delta H^0 = 4RT^2 df/dT$$

at the transition midpoint. Also it is now possible to use this method to obtain a value for ΔG^0 and ΔS^0 using the fact that at the transition midpoint

$$\Delta G^0 = -RTlnK_{eq} = 0$$

and

$$\Delta G^0 = \Delta H^0 - T\Delta S^0$$

thus

$$\Delta S^0 = \Delta H^0/T_m.$$

E. Plotting the data

The program *fastmelt* on the disk Analyze was written by Steve Wolk. It will plot any of the files output by the programs *getdos* (absorbance versus temperature), *convert* (absorbance, normalized absorbance, or fraction versus temperature) or *freeenergy*, *freeuni*, and *freeesc* (ΔG^0 versus temperature). The program *diffplot* on the Dmelt diskette is a modification of *fastmelt* which should perform all of the functions of *fastmelt* along with plotting the output of the program *derivative*.

Using fastmelt

Boot the computer with the Analyze diskette in drive 4 and your data diskette in drive 5. When the Apple Pascal menu appears at the top of the screen type 's' for swapping. The computer will say

Swapping is off. Toggle swapping?

Then type 'x' for execute, and Fastmelt when you are asked for a filename. If you

have toggled swapping you can answer 'n' to the first question then input the filename with the following format

#5:filename (return)

You are given a menu with 4 options. You want '(3) Plot the present array'. Then type < *CR* >, 'n' for a new plot and < *CR* >. You are asked whether you want to reload default values for the plot parameters. Since you have not yet entered any parameters, it does not matter how you answer this question at this time, the default parameters will be loaded anyway. Just hit < *CR* > and you will be given a menu of plot parameters to change. The parameter titles are semi-self explanatory, but in order to avoid confusion I suggest that you ignore parameters 11-14, 32, 33, and 35. Type in the number of the parameter you want to change followed by < *CR* >, then type in the new value followed by < *CR* >. When you are satisfied with all the parameters, type 0 followed by < *CR* >. Then you are asked for a channel number. Enter 2, 3, or 4 followed by < *CR* >. You will be told how many points are being plotted and given the opportunity to reduce that number if you choose. If you answer 'n' to 'change it?' the plot will appear on the screen, including your data points if you have chosen XMIN, XMAX, YMIN, and YMAX in such a way that your data is within the range of the plot parameters. When you want to go to the next step just hit < *CR* >. You are asked whether you want to start a new plot. If you want to plot the same data over again because you weren't satisfied with the plot parameters, type 'y'. Otherwise hit < *CR* >. Then you can plot one of the other channels on the old plot or you can enter 0 if you want to load new data or send your plot to the printer. The former can be done using option 2, and the latter using option 4 on the next menu.

Using diffplot

diffplot functions exactly like *fastmelt* except that you can plot a fraction, derivative, or smoothed data file by typing '1' when asked for the channel number.

F. Other options

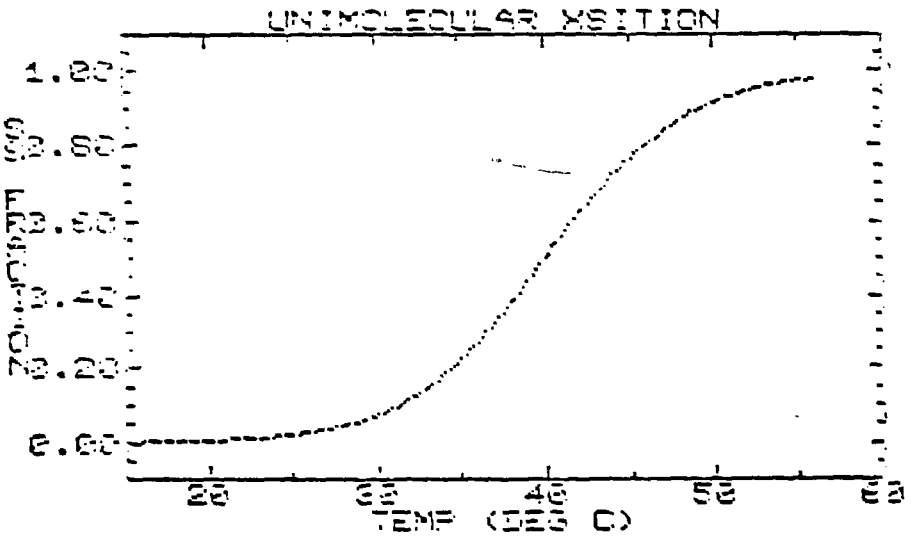
Creative use of the programs described above could lead to all sorts of applications. For example biphasic melts could be analyzed by producing two fraction versus temperature files, one for each transition. The derivative program can actually be used simply to smooth any dataset which is input with the right format, including its own output. Thus one could smooth a dataset several times over. In fact, with respect to signal to noise, two six point smooths in theory should be equivalent to one twelve point smooth, but resolution should be enhanced in the former.

If there is one more program or set of programs that could make the system complete it might be a fitting program which iteratively calculates parameters based on the shape of the curve. Such a program could be helpful in refining thermodynamic parameters which have been calculated by other methods for well behaved transition curves, though it would be difficult to obtain convergence to reasonable values for a random input due to the number of variables involved.

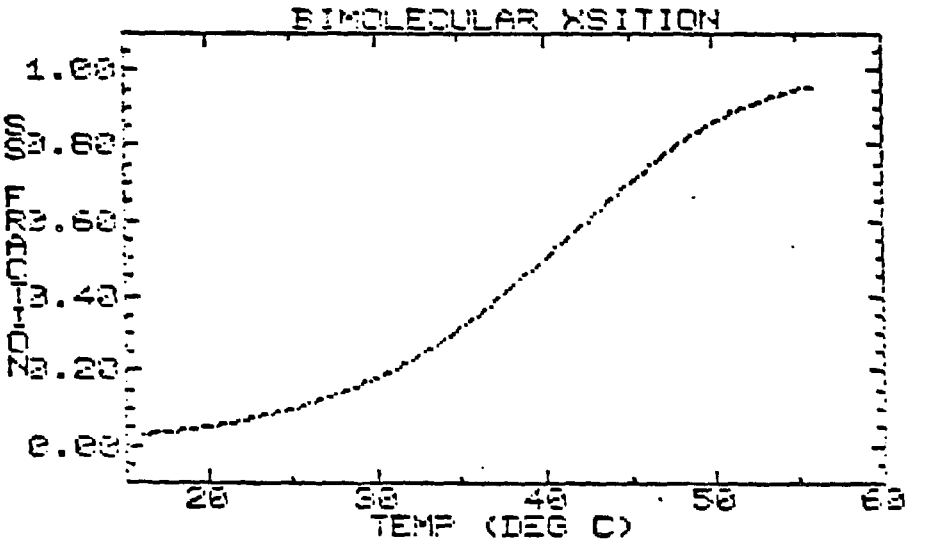
Already it is possible to simulate a transition curve for a given concentration, ΔH^0, and ΔS^0 using programs on the disk Simdat. Figure AII-F shows simulated transition curves and their derivatives for a duplex to single strand and for a unimolecular transition. Each has been simulated using the same T_m, ΔH^0, and ΔS^0. It is clear that, in theory, one can determine the type of transition present from the shape of the curve, though in practice baseline problems may make this impossible for the general case. There is also a program for simulating biphasic curves consisting of two intramolecular transitions. However, if the first transition is bimolecular, the Apple Pascal lacks the precision to do the necessary calculations.

Figure AII-2 Simulated single strand fraction versus temperature curves and their derivatives for unimolecular, self complementary bimolecular, and non self complementary bimolecular transitions. Each simulation assumes the same values for the van't Hoff enthalpy (-47 kcal mol^{-1}) and for the T_m (313 °C) (see text and Gralla and Crothers, (1973) J. Mol. Biol. 78, 301-319).

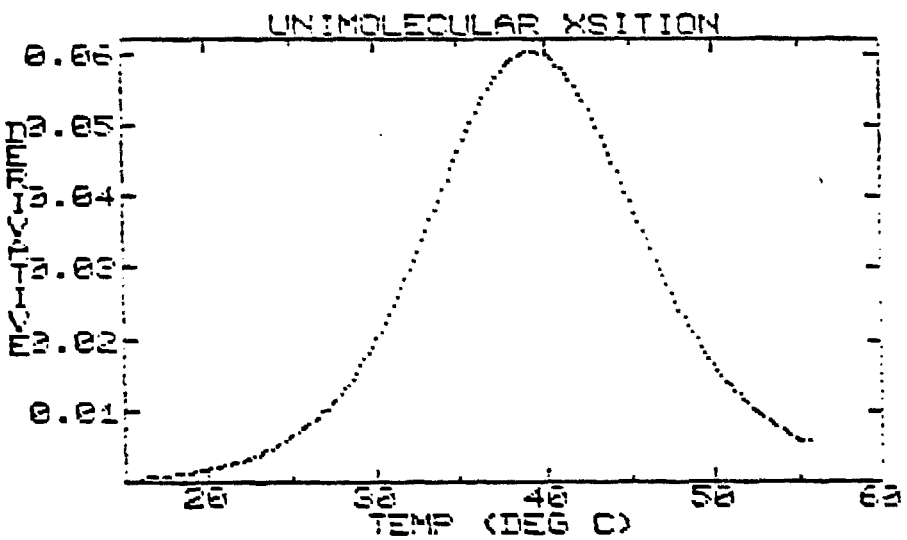
a) Simulated fraction vs. temperature for a unimolecular two state transition.



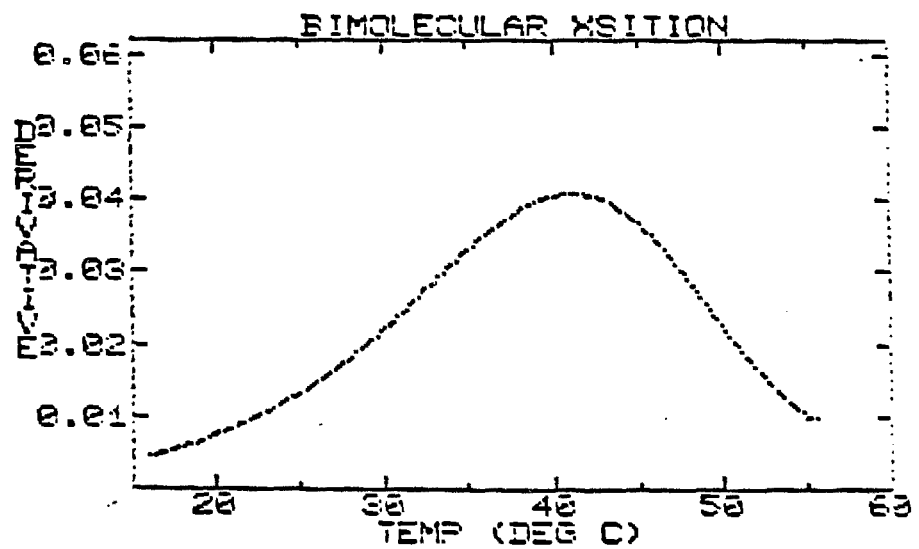
b). Simulated ss fraction vs. temperature for a self complementary bimolecular transition (A+A -->B).



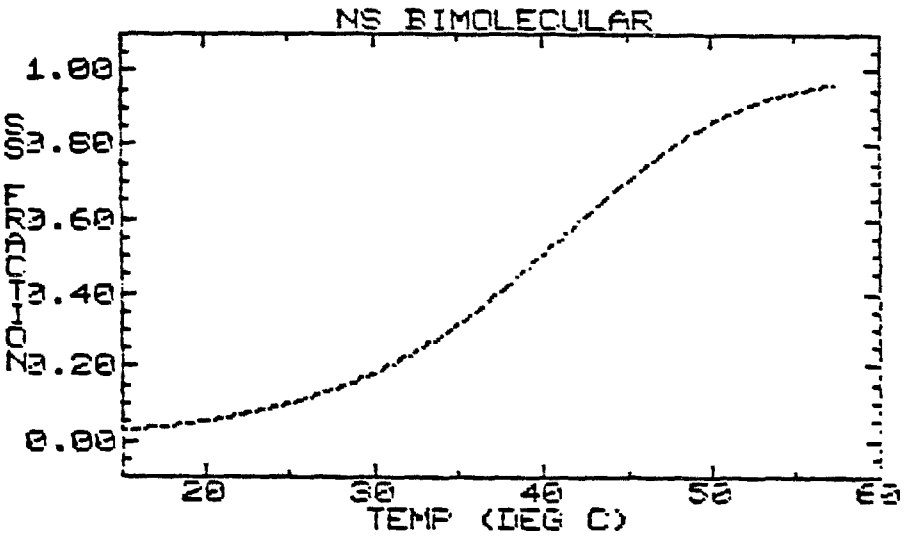
c). Derivative of simulation in a).



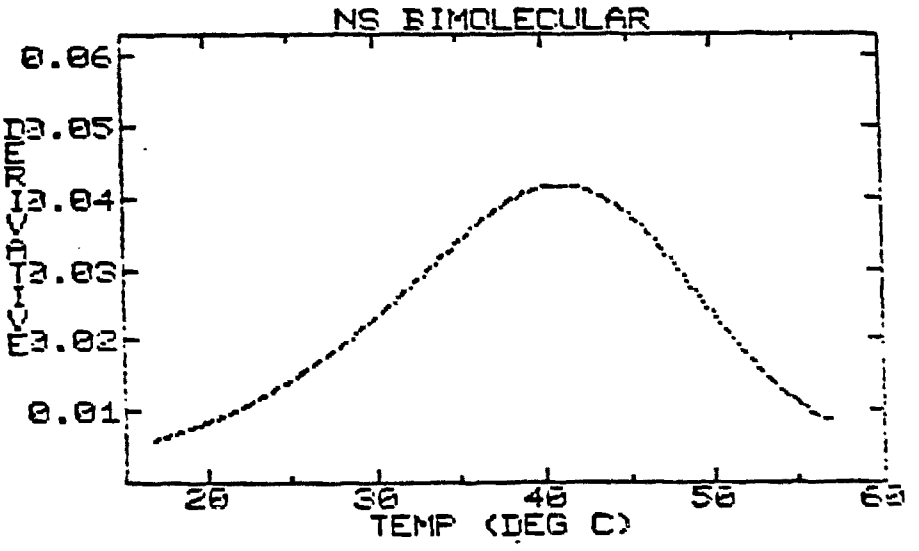
d). Derivative of simulation in b).



e). Simulated ss fraction vs. temperature for a non self complementary bimolecular transition.



f). Derivative of simulation in e).



G. List of programs

Program	Location	Function
convert	Analyze	*-converts PASCAL textfile to PASCAL datafile* **OR** *converts PASCAL formatted absorbance versus temperature data into a. normalized absorbance versus temperature* **OR** *b. fraction versus temperature*
derivative	Dmelt	*-smooths then takes the derivative of output from any of the PASCAL programs listed here*
derdivt	Dmelt	*-same as derivative but outputs* $df/d(1/T)$
diffplot	Dmelt	*-plots output of the other PASCAL programs, including derivative(a modification of fastmelt)*
fastmelt	Analyze	*-plots the output of convert or the programs on the Deltag disk*
freedivt	Deltag	*-converts fraction versus*

		temperature to lnK_{eq} versus inverse of temperature for bimolecular, non self complementary transition
freeenergy	Deltag	*-converts f versus T to ΔG^0 versus T for bimolecular, non self complementary transitions*
freeesc	Deltag	*-same as freeenergy for self complementary duplex*
freeuni	Deltag	*-same as freeenergy for unimolecular transitions*
getdos	Analyze	*-transfers data from a BASIC formatted disk to a PASCAL formatted disk*
Gilford Melt	Melt Data Acq.	*-collects data in BASIC*
lstft	Deltag	*-does a linear least squares fit to the output of freeenergy, freeuni, or freeesc to attain ΔH^0 and ΔS^0 from the slope and intercept*
vhfit	Deltag	*-same as lstft but operates on output of freedivt*
Simulation programs	Simdat	*-simulate unimolecular, bimolecular, or biphasic*

transition curves

(see Appendix III)

Appendix III; Computer programs

```
PROGRAM DERIVATIVE;
(*THIS PROGRAM TAKES A WOLK ABSORBANCE VS TEMPERATURE FILE OR A
FRACTION VS TEMPERATURE FILE , OR A FILE TAKEN FROM ITS OWN OUTPUT,
AT A TEMPERATURE INTERVAL CHOSEN BY THE USER, IT DOES A LOCAL LEAST
SQUARES FIT TO A SECOND ORDER POLYNOMIAL USING ANYWHERE FROM 5 TO 15
LOCAL DATA POINTS. THE PROCEDURES SMOOTH AND DETERM ARE TAKEN FROM
"DATA REDUCTION AND ERROR ANALYSIS ..." BY BEVINGTON. THE COEFFICIENTS
OF THE FITTED POLYNOMIAL ARE THEN USED TO CALCULATE AN APPROXIMATE
DERIVATIVE AT THAT TEMPERATURE. THE SMOOTHED DATA OR THE DERIVATIVE DATA
CAN THEN BE OUTPUT TO ANOTHER DISKFILE, WHICH CAN THEN BE PLOTTED USING
THE PROGRAM DIFFPLOT.*)
USES TRANSCEND;
TYPE LONG=INTEGER[36];
     BLOCKK=ARRAY[1..3,1..3] OF LONG;
VAR MATRIX:BLOCKK;
    DER:ARRAY[1..3,1..250] OF REAL;
    ABSTEMPS:ARRAY[1..2,1..250] OF REAL;
    A:ARRAY[1..3] OF REAL;
    SUMX,SUMY:ARRAY[1..5] OF LONG;
    SSUMX,SSUMY:ARRAY[1..5] OF REAL;
    I,NUMBER,U,MARKER,COUNTER,R,Q,NPTS,NMAX,LEN,NTERMS,N,J,I,K,L,M,L1:INTEGER;
    X,Y:ARRAY[1..250] OF REAL;
    IX,IY:ARRAY[1..250] OF INTEGER;
    LX,LY:ARRAY[1..250] OF LONG;
    NDATASET,CHANO,NUMCHAN,SSBLA,DSBLA,SSBLB,DSBLB,COR,S,CHISQ,FREE,DELTA:REAL;
    LASTPT,COUNT,INTERV,FSTPT:REAL;
    W,P,D,XI,YI,XTERM,YTERM:LONG;
    ANS7,ANS6,ANS2,ANS3,RESPONSE,ANSWER:CHAR;
    DATACHECK,T:STRING;
    SOURCE:STRING[20];
    CHECK:BOOLEAN;
    F,XF,X2F:FILE OF REAL;
    OFILE,O2FILE:STRING[20];


  FUNCTION DETERM(MATRIX:BLOCKK;R:INTEGER):REAL;
  (*CALCULATES THE DETERMINANT OF A THREE BY THREE MATRIX*)

  VAR L,K,J,I,L1,SVALUE:INTEGER;
      P,SAVE:LONG;
      VALUE:LONG;
      Z:REAL;
```

```
BEGIN
IF(R=1) THEN W:=D;
D:=1;
M:=1;
(*VALUE IS A DUMMY VARIABLE USED TO CALCULATE THE DETERMINANT.
AT THE END DETERM IS SET EQUAL TO VALUE.*)
VALUE:=1;
L:=1;
FOR L:=1 TO 3 DO
BEGIN
K:=L;
(*IF A DIAGONAL ELEMENT OF THE MATRIX IS 0, REARRANGE ROWS AND
COLUMNS SO THAT THE VALUE IS NONZERO*)
IF(MATRIX[K,K]=0) THEN
  BEGIN
  M:=0;
  FOR J:=K TO 3 DO
    (*CHECK TO MAKE SURE THAT NONDIAGONAL ELEMENTS ARE NONZERO.
    IF THEY ARE ALL ZERO,THE DETERMINANT IS ZERO*)
    IF(MATRIX[K,J]<>0) THEN
    BEGIN
      M:=1;
      (*REARRANGE ROWS AND COLUMNS*)
      FOR I:=K TO 3 DO
        BEGIN
        SAVE:=MATRIX[I,J];
        MATRIX[I,J]:=MATRIX[I,K];
        MATRIX[I,K]:=SAVE;
        VALUE:=-VALUE;
        END;
    END;
  (*SET VALUE TO ZERO IF ALL ELEMENTS ARE 0*)
  IF (M=0) THEN VALUE:=0;
  END;
IF (VALUE<>0)  THEN
(*REARRANGE ROWS AND COLUMNS UNLESS THE ABOVE LOOP FOUND ALL
ELEMENTS EQUAL TO ZERO*)
BEGIN
   (*THE ALGORITHM REARRANGES ROWS AND COLUMNS UNTIL THE SECTION
   OF THE MATRIX BELOW THE DIAGONAL IS ZERO.  THEN THE DETERMINANT
   WILL BE EQUAL TO THE PRODUCT OF THE DIAGONAL TERMS*)
  VALUE:=VALUE*(MATRIX[L,L]);
    L1:=L+1;
    FOR I:=L1 TO 3 DO
    BEGIN
      FOR J:=L1 TO 3 DO
        BEGIN
        MATRIX[I,J]:=MATRIX[I,J]-
        ((MATRIX[I,L]*MATRIX[L,J]) DIV MATRIX[L,L]);
        END;
    END;
  END;
END;
```

```
   (*CORRECT FOR THE FACT THAT INPUT DATA HAS BEEN MULTIPLIED BY
   VARIOUS FACTORS. EARLIER THE INPUT DATA WAS CONVERTED TO THE
   TYPE LONG INTEGER, NOW THE OUTPUT OF THE FUNCTION DETERMINANT
   IS CONVERTED BACK TO THE TYPE REAL.*)
   IF(VALUE)MAXINT) OR (VALUE(-MAXINT)  THEN
     BEGIN
       P:=(VALUE DIV MAXINT);
       REPEAT
         D:=D*10;
       UNTIL(D)P) AND (D)-P);
     END;
        VALUE:=VALUE DIV D;
   SVALUE:=TRUNC(VALUE);
   IF (D)W) THEN P:=(D DIV W)
   ELSE          P:=W DIV D;
   STR(P,T);
   IF (D)W) THEN LEN:=LENGTH(T)-1
   ELSE LEN:=-(LENGTH(T)-1);
   COR:=EXP(LEN*LN(10));
   IF (R()0) THEN
   BEGIN
   Z:=SVALUE;
   DETERM:=Z*COR;
   END
   ELSE
   BEGIN
   DETERM:=SVALUE;
   END;
   END(*DETERM*);

CEDURE INPUTEMP;(*ASKS FOR STARTING TEMPERATURE AND SCANS THE ARRAY
TEMPS TO FIND THE APPROPRIATE DATA POINT TO INITIATE SMOOTHING*)
IN
:=0;
EPEAT
 WRITELN('STARTING TEMPERATURE=?');
 READLN(FSTPT);
 REPEAT
   G:=G+1;
 UNTIL(ABSTEMPS[1,G])FSTPT) OR (ABSTEMPS[1,G]=FSTPT);
NTIL(G)(NPTS DIV 2)) OR (G=(NPTS DIV 2));
RITELN('FINAL TEMPERATURE=?');
EADLN(LASTPT);
;
```

```
PROCEDURE INPUTFILE;(*ASKS FOR AN INPUT FILENAME AND CHECKS TO MAKE
SURE THAT THE FILE IS PRESENT.  IF THE FILE IS NOT AVAILABLE, A MESSAGE
TO THAT EFFECT IS SENT TO THE SCREEN*)
  BEGIN
  WRITELN('NAME OF DATA FILE TO BE LOADED?(.DATA ASSUMED IF NOT GIVEN)');
  READLN(SOURCE);
  DATACHECK:=COPY(SOURCE,LENGTH(SOURCE)-4,5);
  IF (DATACHECK()'.DATA') THEN SOURCE:=CONCAT(SOURCE,'.DATA');
  (*$I-*)
  RESET(F,SOURCE);
  IF (IORESULT=0) THEN CHECK:=TRUE;
  IF CHECK=TRUE THEN
  ELSE
  BEGIN
    WRITELN('THAT FILE IS NOT AVAILABLE! TRY AGAIN.');
  END;
  (*$I+*)
END;

PROCEDURE LOAD;(*LOADS DATA FROM THE DISKFILE INTO THE ARRAY ABSTEMPS.
ABSTEMPS[1,P] IS THE TEMPERATURE CORRESPONDING TO THE PTH DATA POINT,
AND ABSTEMPS[2,P] IS THE CORRESPONDING ABSORBANCE POINT*)
  BEGIN
     MARKER:=1;
     GET(F);
     NUMCHAN:=F^;
     GET(F);
     NDATASET:=F^;
     IF (NUMCHAN()1) THEN
     BEGIN
     GET(F);
     CHANO:=F^;
     GET(F);
     DSBLA:=F^;
     GET(F);
     DSBLB:=F^;
     GET(F);
     SSBLA:=F^;
     GET(F);
     SSBLB:=F^;
     END;
     GET(F);
      COUNTER:=0; K:=1;
      WRITELN('FILE IS NOW BEING LOADED...');
      IF MARKER()0 THEN
         BEGIN
         WRITELN('FILE HAS',NUMCHAN,'CHANNELS.  SMOOTH WHICH CHANNEL?');
         WRITELN('(IF INPUT IS FRACTIONS, SMOOTHED DATA, OR DERIVATIVES');
         WRITELN('THEN TYPE 1)');
         READLN(NUMBER);
         FOR I:=1 TO (2*NUMBER-2) DO
           BEGIN
```

```
            J:=0;
            REPEAT
              J:=J+1;
              GET(F);
            UNTIL(J=NDATASET);
          END;
        J:=0;
  REPEAT
        COUNTER:=0;
        BEGIN
            J:=J+1;
            COUNTER:=COUNTER+1;
            GET(F);
            ABSTEMPS[1,J]:=F^;
        END;
  UNTIL(J=NDATASET);
  K:=2; J:=0;
  REPEAT
       COUNTER:=0;
       BEGIN
           J:=J+1;
           COUNTER:=COUNTER+1;
           GET(F);
           ABSTEMPS[2,J]:=F^;
       END;
  UNTIL(J=NDATASET);
       END;
  CLOSE(F);
  END;

PROCEDURE CHOOSEPT;(*FOR EACH OUTPUT POINT, THIS PROCEDURE CYCLES THROUGH
THE INPUT DATA POINTS UNTIL IT FINDS THE FIRST INPUT POINT FOR WHICH
THE TEMPERATURE IS GREATER THAN OR EQUAL TO THE OUTPUT TEMPERATURE.
NPTS DATA POINTS ARE THEN INPUT INTO THE ARRAYS X AND Y AND CONVERTED
INTO LONG INTEGERS AFTER BEING MULTIPLIED BY THE APPROPRIATE FACTORS.
WARNING: IF INPUT TEMPERATURE IS GREATER THAN 390 OR THE INPUT
ABSORBANCE IS GREATER THAN 3.9 THE PROGRAM WILL CRASH.*)
  BEGIN
    COUNT:=COUNT+INTERV;
    L:=Q-TRUNC(NPTS DIV 2);
    FOR N:=1 TO NPTS DO
    BEGIN
      X[N]:=ABSTEMPS[1,L+N];
      IX[N]:=ROUND((10)*X[N]);
      LX[N]:=IX[N];
      LX[N]:=LX[N]*100;
      Y[N]:=ABSTEMPS[2,L+N];
      IY[N]:=ROUND((10000)*Y[N]);
      LY[N]:=IY[N];
      LY[N]:=LY[N]*100;
    END;
    IF (ABSTEMPS[1,Q]<COUNT) AND (Q<(NDATASET-4)) THEN
    BEGIN
    REPEAT
      Q:=Q+1;
    UNTIL(ABSTEMPS[1,Q]>COUNT) OR (ABSTEMPS[1,Q]=COUNT) OR (Q>NDATASET-4);
    END;
  END;
```

```
PROCEDURE SMOOTH;(*CALCULATES COEFFICIENTS FOR A SECOND ORDER POLY-
NOMIAL LEAST SQUARES FIT TO NPTS INPUT DATA POINTS FOR EACH OUTPUT
POINT. THE ALGORITHM IS BASED ON A THE FORTRAN PROGRAM POLFIT IN
"DATA REDUCTION AND ERROR ANALYSIS" BY BEVINGTON, PAGES 140-142*)

  BEGIN
  P:=1;
  D:=1;
  NMAX:=2*NTERMS-1;
  N:=0;
  FOR I:=1 TO 3 DO
    BEGIN
      A[I]:=0;
    END;
  REPEAT
    N:=N+1;
    SUMX[N]:=0;
    UNTIL(N=NMAX);
    J:=0;
  REPEAT
    J:=J+1;
    SUMY[J]:=0;
  UNTIL(J=NTERMS);
  I:=0;
  REPEAT
    I:=I+1;
    XI:=LX[I];
    YI:=LY[I];
    XTERM:=1;
    N:=0;
    REPEAT
      N:=N+1;
      SUMX[N]:=SUMX[N]+XTERM;
      XTERM:=XTERM*XI;
      YTERM:=YI;
    UNTIL(N=NMAX);
    N:=0;
    REPEAT
      N:=N+1;
      SUMY[N]:=SUMY[N]+YTERM;
      YTERM:=YTERM*XI;
    UNTIL(N=NTERMS);
  UNTIL(I=NPTS);
  J:=0;
  REPEAT
    J:=J+1;
    K:=0;
    REPEAT
      K:=K+1;
      N:=J+K-1;
      MATRIX[J,K]:=SUMX[N];
    UNTIL(K=NTERMS);
  UNTIL(J=NTERMS);
  R:=0;
  DELTA:=DETERM(MATRIX,R);
  L:=0;
  R:=1;
```

```
  REPEAT
    L:=L+1;
    J:=0;
    REPEAT
      J:=J+1;
      K:=0;
      REPEAT
        K:=K+1;
        N:=J+K-1;
        MATRIX[J,K]:=SUMX[N];
      UNTIL(K=NTERMS);
      MATRIX[J,L]:=SUMY[J];
    UNTIL(J=NTERMS);
    R:=L;
    IF (DELTA()0) THEN
    BEGIN
    A[L]:=DETERM(MATRIX,R)/(DELTA);
    FOR I:=(L+1) TO 3 DO
      BEGIN
      A[L]:=A[L]/1000;
      END;
    END;
 UNTIL(L=3);

END;

PROCEDURE DERIV;(*INPUTS MASSAGED DATA INTO THE ARRAY DER, WHICH WILL
BE THE SOURCE OF OUTPUT.  DER[1,P] IS THE TEMPERATURE*FOR THE PTH
OUTPUT, DER[2,P] IS THE SMOOTHED ABSORBANCE, AND DER[3,P] IS THE
DERIVATIVE(OBTAINED BY TAKING THE DERIVATIVE OF THE FITTED POLYNOMIAL
AT THE TEMPERATURE DER[1,P])*)
BEGIN
  Z:=Z+1;
  DER[1,Z]:=COUNT;
  DER[2,Z]:=(COUNT*COUNT*A[3])+(COUNT*A[2])+A[1];
  DER[3,Z]:=2*A[3]*COUNT+A[2];
  IF(RESPONSE='Y') THEN
  WRITELN('TEMP=',DER[1,Z],':A=',DER[2,Z],':DERIV=',DER[3,Z]);
END;

PROCEDURE OUTPUT;(*OUTPUTS TO DATA FILE*)
BEGIN
WRITELN('OUTPUT SMOOTHED DATA TO A FILE?');
READLN(ANS2);
IF ANS2='Y' THEN
  BEGIN
    WRITELN('NAME OF OUTPUT FILE(.DATA ASSUMED IF NOT GIVEN)');
    READLN(OFILE);
    DATACHECK:=COPY(OFILE,LENGTH(OFILE)-4,5);
    IF DATACHECK()'.DATA' THEN OFILE:=CONCAT(OFILE,'.DATA');
    REWRITE(XF,OFILE);
    WRITELN('NOW WRITING...');
    XF^:=1;
    PUT(XF);
    XF^:=I;
    PUT(XF);
    XF^:=U;
    J:=1;
```

```
        REPEAT
             PUT(XF);
             XF^:=DER[1,J];
             J:=J+1;
        UNTIL(J=U+1);
        J:=1;
        REPEAT
             PUT(XF);
             XF^:=DER[2,J];
             J:=J+1;
        UNTIL(J=U+1);
        PUT(XF);
        XF^:=@@@;
        CLOSE(XF,LOCK);
        WRITELN('...DONE WRITING');
   END;
      WRITELN('OUTPUT DERIVATIVE DATA TO A FILE?');
      READLN(ANS3);
      IF ANS3='Y' THEN
      BEGIN
        WRITE('NAME OF OUTPUT FILE(.DATA ASSUMED IF NOT GIVEN)');
        READLN(O2FILE);
        DATACHECK:=COPY(O2FILE,LENGTH(O2FILE)-4,5);
        IF DATACHECK <> '.DATA' THEN O2FILE:=CONCAT(O2FILE,'.DATA');
        REWRITE(X2F,O2FILE);
        WRITELN('NOW WRITING...');
        X2F^:=1;
        PUT(X2F);
        X2F^:=1;
        PUT(X2F);
        X2F^:=U;
        J:=1;
        REPEAT
             PUT(X2F);
             X2F^:=DER[1,J];
             J:=J+1;
        UNTIL(J=U-1);
        J:=1;
        REPEAT
             PUT(X2F);
             X2F^:=DER[3,J];
             J:=J+1;
        UNTIL(J=U-1);
      PUT(X2F);
      X2F^:=@@@;
      CLOSE(X2F,LOCK);
      WRITELN('...DONE WRITING');
 END;
 END;
```

```
  PROCEDURE PRECIS;(*NOT USED AND WON'T WORK. THE PURPOSE IS TO
  CALCULATE CHI SQUARED. AND THIS PROCEDURE IS LEFT FOR POSSIBLE
  FUTURE USE*)
    BEGIN
      CHISQ:=0;
      FOR I:=1 TO NPTS DO
      BEGIN
        CHISQ:=CHISQ+(Y[I]*Y[I]);
      END;
      J:=1;
      REPEAT
         SSUMY[J]:=TRUNC(SUMY[J]);
        CHISQ:=CHISQ-(2*A[J]*SSUMY[J]);
        FOR K:=1 TO 3 DO
          BEGIN
          N:=J+K-1;
          SSUMX[N]:=TRUNC(SUMX[N]);
          CHISQ:=CHISQ+(A[J]*A[K]*SSUMX[N]);
          END;
        J:=J+1;
      UNTIL(J=4);
      FREE:=NPTS-3;
      CHISQ:=CHISQ/FREE;
      END;

EGIN(*MAIN PROGRAM*)

  ANS6:='Y';
  WRITELN('IF YOU HAVE RECENTLY REBOOTED, YOU MUST TYPE F, THEN G AT');
  WRITELN('THE COMMAND LEVEL OR YOU WILL CRASH. DO YOU WANT TO EXIT?');
  READLN(ANS7);
  IF (ANS7='Y') THEN EXIT(DERIVATIVE);
  REPEAT
  NTERMS:=3;
  I:=1;
  CHECK:=FALSE;
  REPEAT
    INPUTFILE;
  UNTIL(CHECK=TRUE);
  WRITELN('NUMBER OF POINTS FOR FIT=(INPUT A VALUE BETWEEN 5 AND 15)');
  WRITELN('NOTE:  FITTING WITH MORE POINTS WILL TAKE MORE TIME, BUT');
  WRITELN('       PRODUCE A SMOOTHER FIT.');
  READLN(NPTS);
  WRITELN('TEMPERATURE INTERVAL BETWEEN SMOOTHED POINTS=?');
  READLN(INTERV);
  LOAD;
  WRITELN('...FILE IS LOADED');
  INPUTEMP;
  COUNT:=FSTPT-INTERV;
  WRITELN('ECHO DATA TO SCREEN?');
  READLN(RESPONSE);
  U:=TRUNC((LASTPT-FSTPT)/INTERV)+1;
  Z:=0;
  WRITELN('NOW SMOOTHING DATA...');
```

```
    REPEAT
      CHOOSEPT;
      SMOOTH;
      DERIV;
    UNTIL(ABSTEMPS[1,Q]>LASTPT) OR (Q>NDATASET-4);
    OUTPUT;
    WRITELN('INPUT ANOTHER FILE?');
    READLN(ANS6);
    UNTIL (ANS6='N');
    WRITELN('GOODBYE THEN');
    END.



PROGRAM FREEENERGY;
  (*TAKES SSFRACTION VS. TEMPERATURE DATA FROM WOLK FILE AND OUTPUTS VA
  LUES OF K AND OF FREE ENERGY AS A FUNCTION OF TEMPERATURE*)
  USES TRANSCEND;
  VAR ABSTEMPS,DELTAG:ARRAY[1..2,1..250] OF REAL;
      XF,DF:FILE OF REAL;
    DATCHECK,OUTFILE,ANSWER,QUESTION:STRING;
    NDATASET,CHANO:REAL;
    NUMCHAN:REAL;
    SSBLA:REAL;
    SSBLB:REAL;
    DSBLA:REAL;
    DSBLB:REAL;
    OFILE,SCFILE:STRING[20];
    ANS3,ANS2,ANS:CHAR;
    P,M,K,MARKER,I,J,COUNTER:INTEGER;
    X,C:REAL;
    CHECK,REALCHECK:BOOLEAN;



    PROCEDURE INPUTCHECK(QUESTION:STRING);
    (*INPUTS VALUE FROM THE SCREEN AND CHECKS TO MAKE SURE THAT THE
    INPUT IS OF THE CORRECT VARIABLE TYPE*)
    BEGIN
      WRITE(QUESTION);
      (*$I-*);
      READLN(C);
      WRITE('VALUE=',C:11:3,'     IS THIS CORRECT?:');
      READLN(ANSWER);
      IF (LENGTH(ANSWER)=0) THEN REALCHECK:=TRUE
                            ELSE
                              BEGIN
                                IF (ANSWER[1]='Y') THEN REALCHECK:=
                                TRUE;
                              END;
      (*$I+*)
    END;
```

```
PROCEDURE INPUTFILE;
(*INPUTS THE FILENAME,CHECKS TO MAKE SURE THAT THE FILE IS PRESENT
AND OF THE CORRECT TYPE AND ADDS '.DATA' TO THE FILENAME IF NOT
GIVEN*)
BEGIN
    WRITE('NAME OF INPUT DATA FILE(.DATA ASSUMED IF NOT GIVEN)');
    READLN(SCFILE);
    DATCHECK:=COPY(SCFILE,LENGTH(SCFILE)-4,5);
    IF DATCHECK <> '.DATA' THEN SCFILE:=CONCAT(SCFILE,'.DATA');
    (*$I-*)
    RESET(DF,SCFILE);
    IF (IORESULT=0) THEN CHECK:=TRUE;
    IF CHECK=TRUE THEN
    WRITELN('THAT FILE IS AVAILABLE.')
    ELSE
    BEGIN
      WRITELN('THAT FILE IS NOT AVAILABLE!  TRY AGAIN.');
    END;
    (*$I+*)
END;


PROCEDURE OUTPUTFILE;
(*WRITES FINAL DATA TO AN OUTPUTFILE*)
BEGIN
    WRITE('NAME OF OUTPUT FILE(.DATA ASSUMED IF NOT GIVEN)');
    READLN(OFILE);
    DATCHECK:=COPY(OFILE,LENGTH(OFILE)-4,5);
    IF DATCHECK <> '.DATA' THEN OFILE:=CONCAT(OFILE,'.DATA');
    REWRITE(XF,OFILE);
    XF^:=2;
    PUT(XF);
    XF^:=3;
    PUT(XF);
    XF^:=P;
    PUT(XF);
    XF^:=0;
    PUT(XF);
    XF^:=1;
    PUT(XF);
    XF^:=0;
    PUT(XF);
    XF^:=1;
    (*INSERT DUMMY COLUMN SINCE PLOT PROGRAM WILLNOT PLOT CHANNEL 1*)
    REPEAT
        I:=I+1;
        IF (ABSTEMPS[2,I]>0.15) AND (ABSTEMPS[2,I]<0.85) THEN
        PUT(XF);
        XF^:=2.0;
        UNTIL(I=NDATASET);
    I:=1;
    REPEAT
        I:=I-1;
        IF (ABSTEMPS[2,I]>0.15) AND (ABSTEMPS[2,I]<0.85) THEN
        PUT(XF);
        XF^:=2.0;
        UNTIL (I=NDATASET);
```

```
    J:=1;
    (*INSERT TEMPERATURE FOR CHANNEL 2:THE K CHANNEL*)
    REPEAT
        J:=J+1;
        IF (ABSTEMPS[2,JJ]>0.15) AND (ABSTEMPS[2,JJ]<0.85) THEN
        PUT(XF);
        XF^:=ABSTEMPS[1,JJ];
        UNTIL (J=NDATASET);
    J:=1;
    (*INSERT THE VALUES OF K:TO BE PLOTTED AS CHANNEL 2*)
    REPEAT
        J:=J+1;
        IF (ABSTEMPS[2,JJ]>0.15) AND (ABSTEMPS[2,JJ]<0.85) THEN
        PUT(XF);
        XF^:=DELTAG[1,JJ];
        UNTIL (J=NDATASET);
    J:=1;
    (*INSERT TEMPERATURE FOR CHANNEL 3 WHICH WILL CONTAIN DELTAG*)
    REPEAT
        J:=J+1;
        IF (ABSTEMPS[2,JJ]>0.15) AND (ABSTEMPS[2,JJ]<0.85) THEN
        PUT(XF);
        IF (ABSTEMPS[2,JJ]>0.15) AND (ABSTEMPS[2,JJ]<0.85) THEN
        XF^:=ABSTEMPS[1,JJ];
        UNTIL (J=NDATASET);
    J:=1;
    (*INSERT THE VALUES OF DELTAG:TO BE PLOTTED AS CHANNEL 3*)
    REPEAT
        J:=J+1;
        IF (ABSTEMPS[2,JJ]>0.15) AND (ABSTEMPS[2,JJ]<0.85) THEN
        PUT(XF);
        IF (ABSTEMPS[2,JJ]>0.15) AND (ABSTEMPS[2,JJ]<0.85) THEN
        XF^:=DELTAG[2,JJ];
        UNTIL (J=NDATASET);
     PUT(XF);
     XF^:=200;
     CLOSE(XF,LOCK);
END;
```

```
PROCEDURE LOAD;(*READS DATA FROM THE INPUTFILE INTO THE ARRAY ABSTEMPS
AND USES THE INPUT DATA TO SET PARAMETERS*)
BEGIN
    MARKER:=DF^;
    GET(DF);
    NUMDATASET:=DF^;
    GET(DF);
    NDATSET:=DF^;
    GET(DF);
    CHAN2:=DF^;
    GET(DF);
    DSBL2:=DF^;
    GET(DF);
    DSBL3:=DF^;
    GET(DF);
    SSBL2:=DF^;
    GET(DF);
    SSBL3:=DF^;
    GET(DF);
     COUNTER:=0; K:=1;
     IF MARKER()0 THEN
        FOR K:=1 TO 2 DO
                    BEGIN
                    J:=0;
     REPEAT
                   COUNTER:=0;
                   BEGIN
                       J:=J+1;
                      COUNTER:=COUNTER+1;
                      GET(DF);
                      ABSTEMPS[K,J]:=DF^;
                      END;
    UNTIL (J=NDATASET);
                    END;
   CLOSE(DF);
   END;



PROCEDURE DESCRIBE;
BEGIN
WRITELN('THIS PROGRAM TAKES INPUT FROM A MOLE FRACTION VS TEMPERATURE');
WRITELN('FILE, CALCULATES AN EQUILIBRIUM CONSTANT AND FREE ENERGY AT');
WRITELN('EACH TEMPERATURE FOR WHICH THE FRACTION IS GREATER THAN 0.15');
WRITELN('AND LESS THAN 0.85 AND OUTPUTS THE DATA TO A FILE.  K IS OUT-');
WRITELN('PUT AS CHANNEL 2 AND G AS CHANNEL 3 IN A MELT TYPE FILE.  THE');
WRITELN('PROGRAM LSTFT CAN THEN DO A LINEAR LEAST SQUARES FIT OF THE');
WRITELN('DATA IN ORDER TO CALCULATE DELTA H AND DELTA S ASSUMING A');
WRITELN('ALL OR NONE NON SELF COMPLEMENTARY DUPLEX TO SINGLE STRAND');
WRITELN('TRANSITION.');
WRITELN();
END;
```

```
BEGIN(*MAIN PROGRAM*);
DESCRIBE;
ANS:='Y';
REPEAT
CHECK:=FALSE;
REPEAT
  INPUTFILE;
UNTIL (CHECK=TRUE);
LOAD;
QUESTION:='CONCENTRATION=?(ANSWER IN MOLES):';
REALCHECK:=FALSE;
REPEAT
  INPUTCHECK(QUESTION);
UNTIL (REALCHECK=TRUE);
I:=1;
(*CALCULATE THE VALUES OF K AND DELTAG. IF F(0.15 F)0.85 THEN F IS
RESET TO AN ARBITRARY VALUE*)
WRITELN('ECHO DATA TO SCREEN?');
READLN(ANS9);
P:=0;
REPEAT
   I:=I+1;
   IF (ABSTEMPS[2,I](0.15) THEN ABSTEMPS[2,I]:=0.1;
   IF (ABSTEMPS[2,I])0.15) AND (ABSTEMPS[2,I](0.85) THEN
   P:=P+1;
   IF NOT (ABSTEMPS[2,I](0.85) THEN
      ABSTEMPS[2,I]:=0.99999;
      DELTAG[1,I]:=2*(1-ABSTEMPS[2,I])/(SQR(ABSTEMPS[2,I])*C);
    DELTAG[2,I]:=-1.9878*(273+ABSTEMPS[1,I])*2.303*LOG(DELTAG[1,I])/10000;
  IF ANS9='Y' THEN
  WRITELN('TEMP=',ABSTEMPS[1,I],':K=',DELTAG[1,I],':G=',DELTAG[2,I]);
  UNTIL (I=NDATASET);
I:=1;
WRITELN('WRITE TO AN OUTPUT FILE?');
READLN(ANS2);
IF (ANS2='Y') THEN OUTPUTFILE;
WRITELN('ANOTHER DATA SET?(TYPE Y IF YES N IF NO)');
READLN(ANS);
UNTIL (ANS()'Y');
WRITELN('THEN GOODBYE');
END.
```

```
PROGRAM FREEUNI;
(*TAKES FRACTION VS. TEMPERATURE DATA FROM WORK FILE AND OUTPUTS VA
LUES OF K AND OF FREE ENERGY AS A FUNCTION OF TEMPERATURE. ASSUMING AN
ALL OR NONE, UNIMOLECULAR TRANSITION*)
```

```
PROCEDURE DESCRIBE;
BEGIN
WRITELN('THIS PROGRAM INPUTS DATA FROM A WOLK FRACTION VS TEMPERATURE');
WRITELN('FILE AND AT EACH TEMPERATURE CALCULATES A VALUE FOR THE ');
WRITELN('EQUILIBRIUM CONSTANT, K, AND FOR THE FREE ENERGY DIFFERENCE');
WRITELN('ASSUMING AN ALL OR NONE UNIMOLECULAR TRANSITION.  IF YOU');
WRITELN('CHOOSE TO OUTPUT TO A FILE THEN K WILL BE OUTPUT AS "CHANNEL');
WRITELN('2" AND DELTA G WILL BE OUTPUT AS "CHANNEL 3". THE PROGRAM');
WRITELN('LSTFT CAN DO A LINEAR LEAST SQUARES FIT TO THE OUTPUT OF');
WRITELN('THIS PROGRAM IN ORDER TO OBTAIN DELTA H AND DELTA G.');
WRITELN();
WRITELN('NOTE:  ONLY THOSE DATA POINTS FOR WHICH THE FRACTION IS');
WRITELN('       GREATER THAN 0.15 AND LESS THAN 0.85 WILL BE OUTPUT.');
END;
```

Program Selfc

```
*THIS PROGRAM SIMULATES A SS FRACTION VS. TEMPERATURE CURVE FOR AN ALL OR
E, SELF COMPLEMENTARY DUPLEX TO SINGLE STRAND TRANSITION*)
S APPLESTUFF,TRANSCEND;

  ABSTEMPS:ARRAY[1..2,1..250] OF REAL;
  XF:FILE OF REAL;
  OFILE:STRING[20];
  U,I,J:INTEGER;    ANS2:CHAR;
  DATACHECK: STRING[5];
  C,H,K,S,F,CONC,FSTPT,LASTPT,COUNT,PERIOD:REAL;

ROCEDURE OUTPUT;
EGIN
RITELN('OUTPUT SIMULATED DATA TO A FILE?');
EADLN(ANS2);
F ANS2='Y' THEN
 BEGIN
   WRITELN('NAME OF OUTPUT FILE(.DATA ASSUMED IF NOT GIVEN)');
   READLN(OFILE);
   DATACHECK:=COPY(OFILE,LENGTH(OFILE)-4,5);
   IF DATACHECK()'.DATA' THEN OFILE:=CONCAT(OFILE,'.DATA');
   REWRITE(XF,OFILE);
   WRITELN('NOW WRITING...');
   XF^:=1;
   PUT(XF);
   XF^:=1;
   PUT(XF);
   XF^:=250;
   J:=1;
   REPEAT
        PUT(XF);
        XF^:=ABSTEMPS[1,J];
        J:=J+1;
   UNTIL(J=251);
```

```
      J:=1;
      REPEAT
            PUT(XF);
            XF^:=ABSTEMPS[2,J];
            J:=J+1;
      UNTIL(J=251);
      PUT(XF);
      XF^:=000;
      CLOSE(XF,LOCK);
      WRITELN('...DONE WRITING');
   END;
   END;


BEGIN(*MAIN PROGRAM*);

    WRITELN('CONCENTRATION=?');
    READLN(CONC);
    WRITELN('STARTING TEMPERATURE=?');
    READLN(FSTPT);
    WRITELN('FINAL TEMPERATURE=?');
    READLN(LASTPT);
    WRITELN('INTERVAL BETWEEN POINTS=?');
    READLN(PERIOD);
    WRITELN('DELTA H=?');
    READLN(H);
    WRITELN('DELTA S=?');
    READLN(S);
    U:=ROUND((LASTPT-FSTPT)/PERIOD);
    COUNT:=FSTPT;
    FOR I:=1 TO U   DO
       BEGIN
          ABSTEMPS[1,I]:=COUNT;
          K:=EXP((-H/(1.98717*(ABSTEMPS[1,I]+273))) + (S/1.98717));
          IF(K)0.01) THEN
          ABSTEMPS[2,I]:=((1+4*K*C)-SQRT(1+8*K*C))/(4*K*C)
          ELSE ABSTEMPS[2,I]:=0.001;
          COUNT:=COUNT+PERIOD;
       END;
    OUTPUT;
END.
```

Program Threeft

```
(*THIS PROGRAM SIMULATES THEN OUTPUTS APPARENT FRACTION VERSUS TEMPERATURE
DATA FOR A BIMOLECULAR THREE STATE TRANSITION. INPUTS EXPERIMENTAL
DATA AND CALCULATES DHI*)
```

```
:S TRANSCEND;


?  ABSTEMP:ARRAY[1..7,1..250] OF REAL;
   DATPT:ARRAY[1..2,1..250] OF REAL;
   F,XF:FILE OF REAL;
   SOURCE,OFILE:STRING[20];
   CHAN,L,NUMBER,I,J,U:INTEGER;    ANS3,ANS2:CHAR;
   BSPAR:CHAR;
   DATACHECK: STRING[5];
   NEWN,W,R,Y,Z,HI,HD,SI,SD,DI,S,K,COUNT,FSTPT,LASTPT,PERIOD:REAL;
   SUM,NDATA,LBLN,UBLN:REAL;
   CHECK:BOOLEAN;


?ROCEDURE OUTPUT;
(*OUTPUTS TO A DATA FILE. ALLOWS SEVERAL OPTIONS FOR PARAMETERS TO OUTPUT*)
EGIN
    WRITELN('OUTPUT WHICH PARAMETER?');
    WRITELN('TYPE A NUMBER BETWEEN 2 AND 7:');
    WRITELN('2=KI');
    WRITELN('3=KD');
    WRITELN('4=FAPP');
    WRITELN('5=KAPP');
    WRITELN('6=GAPP');
    WRITELN('7=ABS');
    READLN(I);
    WRITELN('NAME OF OUTPUT FILE(.DATA ASSUMED IF NOT GIVEN)');
    READLN(OFILE);
    DATACHECK:=COPY(OFILE,LENGTH(OFILE)-4,5);
    IF DATACHECK()'.DATA' THEN OFILE:=CONCAT(OFILE,'.DATA');
    REWRITE(XF,OFILE);
    WRITELN('NOW WRITING...');
    XF^:=1;
    PUT(XF);
    XF^:=1;
    PUT(XF);
    XF^:=U;
    J:=1;
    REPEAT
          PUT(XF);
          XF^:=ABSTEMP[1,J];
          J:=J+1;
    UNTIL(J=U);
    J:=1;
    REPEAT
          PUT(XF);
          XF^:=ABSTEMP[I,J];
          J:=J+1;
    UNTIL(J=U);
    PUT(XF);
    XF^:=000;
    CLOSE(XF,LOCK);
    WRITELN('...DONE WRITING');
 END;
```

```
PROCEDURE LOAD;
  BEGIN
  WRITELN('SOURCE=',SOURCE);
  GET(F);
  GET(F);
  NDATA:=F^;
  WRITELN(F^);
  WRITELN('NDATA=',NDATA);
  FOR L:=1 TO 6 DO
      BEGIN
      GET(F);
      END;
  J:=0;
  WRITELN('INPUT WHICH CHANNEL?(INPUT 1 FOR FRACTION OR SMOOTHED FILES)');
  READLN(CHAN);
  WRITELN('NOW LOADING DATA');
  REPEAT
      J:=J+1;
      I:=0;
      REPEAT
          I:=I+1;
          GET(F);
          IF J=(2*CHAN)-1 THEN
          DATPT[1,I]:=F^;
          IF (J=2*CHAN) THEN DATPT[2,I]:=F^;
          UNTIL(I=ROUND(NDATA));
      UNTIL (J=8);
      CLOSE(F,LOCK);
      WRITELN('FINISHED LOADING DATA');
   END;


PROCEDURE IFILE;
  BEGIN
  WRITELN('NAME OF DATA FILE TO BE LOADED?');
  READLN(SOURCE);
  IF (LENGTH(SOURCE) <> 0) THEN
  DATACHECK:=COPY(SOURCE,LENGTH(SOURCE)-4,5);
  IF (DATACHECK <> '.DATA') THEN SOURCE:=CONCAT(SOURCE,'.DATA');
  (*$I-*)
  RESET(F,SOURCE);
  IF (IORESULT=0) THEN CHECK:=TRUE;
  (*$I+*)
  END;

CEDURE PARAMETERS;
  BEGIN
```

```
WRITELN('DELTA HI(CALORIES)=?');
READLN(HI);
WRITELN('DELTA HD=?');
READLN(HD);
WRITELN('DELTA SI(EU)=?');
READLN(SI);
WRITELN('DELTA SD=?');
READLN(SD);
WRITELN('DI=?');
READLN(DI);
W:=0;
R:=0;
Y:=1;
Z:=0;
REPEAT
WRITELN('LOWER BASELINE=A+B*T');
WRITELN('UPPER BASELINE=C+D*T');
WRITELN('PRESENT VALUES:  A=',W,'   ;B=',R);
WRITELN('                 C=',Y,'   ;D=',Z);
WRITELN('CHANGE WHICH PARAMETER?(ENTER X FOR NO CHANGES)');
READLN(BSPAR);
IF (BSPAR()'X') THEN
BEGIN
WRITELN('NEW VALUE=?');
READLN(NEWN);
END;
   IF (BSPAR='A') THEN BEGIN W:=NEWN; END;
   IF (BSPAR='B') THEN BEGIN R:=NEWN; END;
   IF (BSPAR='C') THEN BEGIN Y:=NEWN; END;
   IF (BSPAR='D') THEN BEGIN Z:=NEWN; END;
UNTIL(BSPAR='X');
END;

PROCEDURE GENERATE;
BEGIN
      ABSTEMP[1,I]:=COUNT;
      ABSTEMP[2,I]:=EXP((HI/(1.98717*(273+COUNT)))+(-SI/1.98717));
      ABSTEMP[3,I]:=EXP((HD/(1.98717*(273+COUNT)))+(-SD/1.98717));
      IF ABSTEMP[3,I]0 THEN
      ABSTEMP[4,I]:=ABSTEMP[3,I]*(1+((DI*ABSTEMP[2,I])/ABSTEMP[3,I]))
      ELSE ABSTEMP[4,I]:=0.00001;
      ABSTEMP[5,I]:=ABSTEMP[4,I]/(1+((1-DI)*ABSTEMP[2,I]));
      ABSTEMP[6,I]:=-(1.98717)*(ABSTEMP[1,I])*2.303*(LOG(ABSTEMP[5,I]));
      ABSTEMP[4,I]:=ABSTEMP[5,I]/(1+ABSTEMP[5,I]);
      LBLN:=W+R*ABSTEMP[1,I];
      UBLN:=Y+Z*ABSTEMP[1,I];
      ABSTEMP[7,I]:=LBLN+ABSTEMP[4,I]*(UBLN-LBLN);
 END;
```

```
    PROCEDURE DIFFERENCE;
    BEGIN
         SUM:=0;
         WRITELN('NOW COMPUTING ERROR');
         FOR I:=1 TO ROUND(NDATA) DO
              BEGIN
              COUNT:=DATPT[1,I];
              GENERATE;
              SUM:=SUM+((DATPT[2,I]-ABSTEMP[7,I])*(DATPT[2,I]-ABSTEMP[7,I]));
              END;
         SUM:=SQRT(SUM)/NDATA;
    END;


    PROCEDURE SIMULATE;
       BEGIN
          COUNT:=0;
          WRITELN('STARTING TEMPERATURE=?');
          READLN(FSTPT);
          WRITELN('FINAL TEMPERATURE=?');
          READLN(LASTPT);
          WRITELN('INTERVAL BETWEEN POINTS(DEGREES C)=?');
          READLN(PERIOD);
          U:=ROUND((LASTPT-FSTPT)/PERIOD);
          WRITELN('ECHO DATA TO SCREEN?');
          READLN(ANS3);
          FOR I:=1 TO U DO
             BEGIN
             GENERATE;
             IF(ANS3='Y') THEN
             WRITELN('TEMP=',ABSTEMP[1,I],'   :FAPP=',ABSTEMP[7,I]);
             COUNT:=COUNT+PERIOD;
          END;
          OUTPUT;
          END;


BEGIN(*MAIN PROGRAM*)

    PARAMETERS;
    IFILE;
    LOAD;
    DIFFERENCE;
    WRITELN('CHISQUARED=',SUM);
    WRITELN('OUTPUT A SIMULATED DATA FILE?');
    READLN(ANS2);
    IF (ANS2='Y') THEN
    SIMULATE;
END.
```

Program Threest

```
(*THIS PROGRAM SIMULATES THEN OUTPUTS SS FRACTION VERSUS TEMPERATURE
DATA FOR A BIPHASIC (PAIR) OF UNIMOLECULAR TRANSITION(S)*)
```

Program Unimol

```
(*THIS PROGRAM SIMULATES THEN OUTPUTS SS FRACTION VERSUS TEMPERATURE
DATA FOR A UNIMOLECULAR ALL OR NONE TRANSITION*)
USES TRANSCEND;
```

Basic Programs

```
LOAD BICOOP
LIST

 10  HOME
 20  INPUT "SIGMA=";S
 30  INPUT "NUMBER OF TERMS=?";I
 40  INPUT "N=?";N
 50 A = 0
 60 B = 0
 70 C = 0
 80 NBAR = (1 - (S ^ (0.5))) ^ N
 90 NBAR = NBAR * N
100 J = 1
110 A = S ^ (0.5)
115 A = A ^ J
120 A = A * (N / (J + 1))
130 D = 1 - (S ^ (0.5))
140 D = D ^ (N - J)
150 A = A * D
160 B = N
170  PRINT J
180 K = 1
190  IF J = 1 THEN  GOTO 1320
200 B = B * ((N - K) / (K + 1))
250 K = K + 1
300  IF K < J THEN  GOTO 1000
400 C = (J * N) - ((4 * J) - 6)
450 C = C / J
520 NBAR = NBAR + (A * B)
530 NBAR = NBAR - (A * C)
540  PRINT NBAR
550 J = J + 1
600  IF J <  = I THEN  GOTO 700
650  PRINT "N=",N
650  PRINT "NBAR=",NBAR
700  INPUT "ANOTHER DATA PT?";A$
720  IF A$ = "Y" THEN  GOTO 500
```

(this program overestimates the effect of constraining B-Z junction size to be 10 bps.

```
]LOAD COST10
]LIST

200  HOME
300  INPUT "SIGMA=";S
400  INPUT "NUMBER OF TERMS=?";I
500  INPUT "N=?";N
520 A = 0
530 B = 0
540 C = 0
550 NBAR = (1 - (S ^ (0.5))) ^ N
560 NBAR = NBAR * N
600 J = 1
700 A = S ^ (0.5)
705 A = A ^ J
710 A = A * (N / (J + 1))
720 D = 1 - (S ^ (0.5))
730 D = D ^ (N - J)
740 A = A * D
800 B = N
840  PRINT J
900 K = 1
950  IF J = 1 THEN  GOTO 1320
1000 F = (N - (10 * K))
1050 B = B * (F / (K + 1))
1060 K = K + 1
1100  IF K ( J THEN  GOTO 1000
1320 NBAR = NBAR + (A * B)
1340  PRINT NBAR
1350 J = J + 1
1400  IF J (  = I THEN  GOTO 700
1500  PRINT "N=",N
1550  PRINT "NBAR=",NBAR
1600  INPUT "ANOTHER DATA PT?";A$
1700  IF A$ = "Y" THEN  GOTO 500
```