Protein-DNA Interactions:
Molecular Modeling of Structure and Energetics

Thesis by

Kevin W. Plaxco

In partial fulfillment of the requirements
for the degree of Doctor of Philosophy

California Institute of Technology
Pasadena, California
1994
(submitted July 1, 1993)

To Lisa

Acknowledgments

My tenure at Caltech has been simultaneously the most difficult and the most rewarding period of my life. The caliber of the people here, both intellectually and spiritually, has never ceased to amaze and delight me, as well as push me to expand my own limits and aspirations. I cannot possibly acknowledge all the people who have had an impact on me over the years, both as coworkers and as friends. I have had, as a side effect of the many collaborative projects that I have entertained, the pleasure of working with and, I hope, befriending a frightful number of people here at 'Tech.

First, I must record the great pleasure to be derived from working under someone for whom one has absolute respect. That, I think, will be one of the greatest gifts that I will take away from my time at Caltech. Thank you Bill.

It would take several more pages to list all the people who have been good friends over the years, and who have made the time here more enjoyable, but there are, of course, a few of special note. The boys of 007 have been a constant source of inspiration, insight and were even down right fun from time to time; I'm not sure if they, in the end, helped get me through my research or held me up, but it was a pleasure whichever way it went. There are more than a few folks who actually did push me through my research and out the door, and for whom I have special thanks. Alan Mathiowetz, K.T. Lim, Naoki Karasawa, Murco Ringnalda, Siddharth Dasgupta, and Adel Naylor: thank you one more time for all that MD/Unix/CMM/PS-GVD guidance you've given me over the years, I hope it wasn't too much of a bother.

Special thanks are in order too to my dear friends Faiz Kayyem, for the countless inspirations our friendship has fostered, and Lindsay Hansen, who took the time to remind me that there is life beyond and above science. I can not imagine what graduate school would have been like without them.

And, of course, the people who have been there the longest are my family, who have been so silently supportive for the many years this adventure has required. Along with them I've saved the best for last I feel so fortunate to have had a home away from Caltech, a little cottage refuge made secure by my loving and supportive wife, who, as I have completed my graduate studies and beyond, remains my love and salvation.

Abstract

The thesis deals with the structural elements involved in and the energetics of sequence specific recognition of DNA.

Chapter 1 of the thesis provides a brief overview on the mechanics and applicability of molecular dynamics based methods for studying the structure and function of molecules of proteins, DNA and other macromolecules of biological relevance.

Chapter 2 presents a constrained molecular dynamics derived model we have developed for the DNA binding domain of the protein Hin Recombinase. Based on a combination of homology modeling and experimentally derived placement constraints we used molecular dynamics to conduct a search of conformation space constrained to remain consistent with the then known experimental characterizations of the protein. The model generated by this approach allowed us to correctly predict the sequence selectivity of the Hin, and lead to a number of insights into the nature of its sequence selectivity.

Chapter 3 discusses a variety of experimental results that have been obtained on the structure of the Hin-*hix* complex. While these results primarily conform with model based predictions, others have pointed towards further refinements that are possible.

Chapter 4 of the thesis provides a brief overview of the methods and applicability of perturbation thermodynamic analysis as applied to the molecular dynamics

based simulation of proteins and DNA in general and some specific issues of concern with regard to the simulations reported in this thesis.

In chapter 5 we report on our perturbation thermodynamic molecular dynamics analysis of the relative free energy of solvation of thymine and uricil. This work provides important insights into the role solvation plays in the formation of sequence specific protein-DNA complexes.

Finally, chapter 6 is a report on our investigations into the mechanisms of sequence specific binding for the minor groove binding peptide Netropsin. Steric, electrostatic and solvation effects are all investigated using a perturbation thermodynamic approach to elucidate the mechanisms involved in complex formation for this important class of DNA binding ligands.

Table of Contents

Acknowledgments ..... *iv*

Abstract ..... *vi*

Overview of the Thesis 1

Chapter 1. The Molecular Dynamics Description of Molecular Motion

Abstract 4

I Introduction 5

II The Force field Description 6

III Energy Minimization 13

IV The Equations of Motion 14

V Monte Carlo Techniques 16

VI Characteristics of the Ensemble 18

VII Accuracy of the Force field-Molecular Dynamics Approach 20

VIII Periodic Boundary Conditions and Nonbond Cutoffs 22

IX The Limitations of High Frequency Internal Modes 26

X Conclusions 28

XI References 29

Chapter 2. Structural Predictions of Protein-DNA Interactions

Abstract 33

I Introduction 34

II Generating the Constraints 40

III Modeling Techniques 41

IV Accuracy in Homology Modeling 46

V Conclusions 47

VI Predictions of Structural Elements for the Binding of
Hin Recombinase with the *Hix* Site of DNA 48
VII References 67

Chapter 3. Experimental Investigations of the Hin-*hix* Complex
Abstract 71
I Introduction 72
II Structural Tests 73
III Identification of Residues on the DNA Binding Surface 76
IV Contributions to the Free Energy of Binding by Thymine Methyl Groups 78
Materials and Methods 81
V Consistency with the Literature 85
Sequence Specificity 85
Increased Affinity Mutants 88
Chemical Modification of Hin and *hix* 89
Comparison with the Homeodomain Structure 91
VI Structural Refinements 93
VII Conclusions 93
VIII References 95

Chapter 4. Perturbation Thermodynamics
Abstract 98
I Introduction 99
II Methodologies 105
Generating the Ensemble 105
The F. F. Description 106

Equilibrium and Convergence 108

Estimation of Error 110

III Biological Perturbation Thermodynamics 113

IV Conclusions 115

V References 116

Chapter 5. Perturbation Thermodynamic Analysis of the Free Energy of Solvation of the Thymine Methyl Group: Solvent Driven Contributions to Sequence Specific Binding

Abstract 119

I Introduction 120

II Methods 124

The F. F. Description 124

Simulation Methods 128

Convergence and Equilibrium 133

Internal Energy Contributions 134

III Results 137

IV Conclusions 138

V References 142

Chapter 6. Perturbation Thermodynamic Analysis of Netropsin: Sequence Specificity of DNA Binding Driven by Electrostatic Interactions

Abstract 145

I Introduction 146

II Methods 154

The F.F. Description 154

Simulation Methods 159

Convergence and Equilibrium 163

III Results 163

IV Conclusions 165

V References 169

Appendicies

I The Derivation of Perturbation Thermodynamics 172

II A Review of Water Potentials 174

III Molecular Dynamics, an Outline 182

IV Some Issues in Computational Efficiency 189

Overview of the Thesis

The overall goal of our research has been to develop insights into the structural an energetic underpinnings of an important class of molecular recognition systems, namely, protein-DNA and peptide-DNA interactions. In the long term, one would envision the development of DNA binding ligands tailored to any specificity for use in research and, more importantly, in a theraputic context for the vast number of diseases that could be ameliorated by fine scale manipulation of the genome. Consequently, we have focused our research on two aspects of peptide-DNA interactions: (i) the development of a simulation technique for the atomic level prediction of protein-DNA interactions based on a molecular dynamics conformational search constrained by homology, substrate specificity and biochemical characterization, and (ii) the use of perturbation thermodynamics simulation techniques to further our insight into the mechanisms and energetics of sequence specific binding.

The first chapter of this thesis details the fundamental principles of molecular dynamics and reviews the issues pertinent to the simulation work reported in later chapters.

The second chapter of this thesis reports our work on modeling the structure of the protein Hin recombinase in complex with the *hix* element of DNA. Typical protein folding rates are on the order of milliseconds to tens of seconds, more than 12 orders of magnitude slower than the longest reasonable molecular dynamics simulation times. Thus, the applicability of molecular dynamics simulations to the *ab initio* protein folding "problem" is limited. However, the exploration of

protein structure is not conducted in an intellectual vacuum. More than 500 protein structures have been determined and a large fraction of all proteins belong to a homologous family with at least one member of known structure. Biochemical characterization techniques such as affinity cleavage, substrate footprinting, and the determination of cysteine bonds can be used to generate further constraints on conformational searches. Realizing the applicability of molecular dynamics routines to constrained conformational searches, we have attempted the characterization of the DNA binding domain Hin, a 52 amino acid polypeptide consisting of the amino-terminal third of the *Salmonella* enzyme Hin recombinase. The model generated exhibits atomic level detail, and allows for the prediction of the primary structural elements of the DNA binding domain as well as prediction of its sequence specificity.

Because molecular modeling is of no value unless the model generated has predictive utility, the third chapter of this thesis consists of biochemical and genetic investigations into model based structural predictions on the Hin-*hix* complex. To date, our model of the Hin structure has largely withstood the experimental onslaught by correctly predicting the sequence selectivity of Hin, the characteristics of a variety of Hin mutants, and the biochemistry of *hix* site recognition. The minor discrepancies noted between experiment and theory can be accommodated and rather than discrediting the model have lead to refinements of it. We believe that this approach of biasing molecular simulations with experimentally derived knowledge and using the results of these simulations to design new, more precise experimental tests will be of general utility.

The second aspect of our studies, the energetics of sequence specific protein-

DNA interactions, are reported in chapters four through six of this thesis. Chapter four consists of an overview of simulation based methods for the estimation of free energies and addresses specific issues that we have explored regarding the accuracy of these techniques. In it we report that molecular dynamics simulations can be used to generate ensembles for molecular systems over which dynamical averages can be taken to determine "ensemble average properties" such as free energy. The true utility of molecular simulations is that the individual components of an energetic system can be separated in a way that often experiment can not achieve which can often lead to further insights into the mechanisms involved.

The fifth chapter of this these reports simulation work we have conducted as a follow up to our work on the major groove recognition elements of Hin recombinase. We have used perturbation thermodynamic analysis to determine the relative free energy of solvation of the thymine methyl group in order to estimate the contribution that solvation effects make to this important mechanism of sequence selective binding.

The conclusion of this thesis details simulations we have conducted on the binding of sequence specific DNA binding polypeptides, the minor groove binding poly-pyrrole antibiotics, that have lead to insights into the mechanisms of their selectivity. By using perturbation thermodynamic analysis, we have been able to ascertain the role of solvation, steric interactions and electrostatics on complex formation and the generation of sequence specificity–issues which are difficult to address experimentally.

Chapter 1

The Molecular Dynamics Description of Molecular Motion

Abstract

Molecular dynamics techniques for the elucidation of the structure and energetics of biological macromolecules have come of age. Advances in both methodology and computational power have created a set of computational tools that are beginning to achieve a high degree of utility in chemistry and biochemistry. In particular, molecular dynamics based techniques for the determination of protein structure and the energetics of solvation, ligand interactions, and molecular recognition have yielded a plethora of theoretical insights into the nature of protein structure and function. While these techniques are founded on a variety of assumptions that must be carefully assessed when judging the validity of computational results, they show great potential as an aid in solving a variety of problems in biochemistry and biophysics.

Introduction

Molecular dynamics–the science of simulation of the motions of a system of atoms–has been applied to biological macromolecules to provide insights into the structure and thermodynamics of a variety of biologically relevant molecular systems. As applied to biological macromolecules, the technique is predicated on a classical consideration of atomic motion and on a simple but demonstrably accurate method of describing molecular energies and forces. It has been used successfully to study a wide variety of problems in biology (reviewed in Karplus and Petsko, 1990), including the dynamical aspects of protein-ligand interactions (Czerminski and Elber, 1991), to derive structural models of proteins (Plaxco *et al.*, 1989), and for the determination of relative free energies of solvation (Lybrand *et al.*, 1985a), protein-ligand interactions (Bash *et al.*, 1987), and even catalysis (Arad *et al.*, 1990).

The essential elements of a molecular dynamics simulation are a knowledge of the interaction potential and masses of the particles involved, from which the forces and resultant accelerations that drive the time evolution of the system can be derived. Two aspects of molecular dynamics are the source of the technique's utility in the investigation of biological macromolecules. First, molecular dynamics realistically describes molecular motions and the energetics of molecular conformations. Because of this, molecular dynamics can be used to accurately sample conformation space and generate realistic ensembles for the determination of thermodynamic averages and for the search for low energy conformers. Second, although the potential used is at best an approximation of the actual energetics of the system, it is completely under the user's control, so that the specific contribution of individual

components of the potential in determining a given property can be exhaustively examined.

In the remainder of this chapter the detailed methods and technical limitations of the general molecular dynamics approach are described. In following chapters, applications will be illustrated, and a more detailed analysis of some of these techniques will be discussed.

The Forcefield Description

Molecular dynamics simulations are based on the precept that intra- and intermolecular forces can be well approximated by a limited set of nonbond and valence characteristics, termed a forcefield, and that molecular motions are limited to classical responses to these forces.

A variety of forcefield methods have evolved in the last decade (Brooks *et al.*, 1983; Weiner *et al.*, 1984; Mayo *et al.*, 1990) that predominantly differ from one another only in the potential functions and parameter sets they utilize for their description of molecular forces. In the work reported here, three different forcefields, AMBER, DREIDING and TIP3P, each optimized to characterize a different class of molecules, were used. The following discussion focuses on the specific functional descriptions used in the AMBER forcefield (Weiner *et al.*, 1984; Weiner *et al.*, 1986), with the relevant DREIDING (Mayo *et al.*, 1990) and TIP3P (Jorgensen *et al.*, 1983) differences noted. The TIP3P forcefield and other published water forcefields are discussed in more detail in appendix II.

The underlying premise of the forcefield description is that an accurate description of the total energy of a molecular system is given by:

$$E_{tot} = E_B + E_\theta + E_\omega + E_\phi + E_{vdW} + E_{el} + E_{Hb}$$

where E_{tot}, E_B , E_θ , E_ω , E_ϕ , represent the total, bond, angle, torsion, and inversion energies of the system, and E_{vdW}, E_{el} and E_{Hb} represent van der Waals, electrostatic and hydrogen bonding nonbond components of the total energy.

For many systems, simple harmonic terms can be used to adequately describe bonding and angle deformation energies so we have for bonds:

$$E_B = \frac{1}{2}\sum_{i=1}^{N} k_i^b (r_i - r_i^{eq})^2$$

and angles:

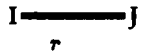
$$E_\theta = \frac{1}{2}\sum_{i=1}^{M} k_i^\theta (\theta_i - \theta_i^{eq})^2$$

where N and M are the total number of bonds and bond angles respectively, k^b and k^θ are empirically or theoretically derived force constants, and r^{eq} and θ^{eq} are the equilibrium bond lengths and angles. These and other valence characterization formulations are diagrammed in figure 1.

Torsional barriers (also called 1-4 bonding interactions), which are barriers to bond rotation, are also described in the forcefields, usually as a Fourier series expansion about the torsion angle ϕ:

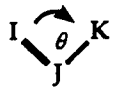
$$E_\phi = \sum_{i=1}^{T} |k_{m\phi}| - k_{m\phi} cos(n\phi), n = 1, 2, 3, 4, 5, 6,$$

Bonded interactions



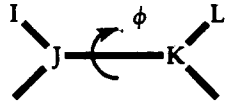
$$E_B = \frac{1}{2}k^b(r - r^{eq})^2$$

Angle interactions



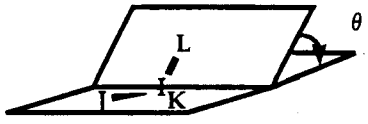
$$E_\theta = \frac{1}{2}k^\theta(\theta - \theta^{eq})^2$$

Torsion interactions



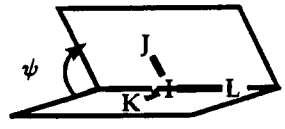
$$E_\phi = |k_{m\phi}| - k_{m\phi}cos(n\phi), n = 1, 2, 3, 4, 5, 6$$

DREIDING type inversion



$$E_\theta = \frac{1}{2}k^\theta(cos(\theta) - cos(\theta^{eq}))^2$$

AMBER type inversion



$$E_\psi = \frac{1}{2}k^\psi cos(n(\psi - \psi^{eq}))$$

Figure 1: A schematic diagram of the various valence characteristics described by common biological macromolecule forcefields.

where T is the total number of torsions, n, is the periodicity of the torsional constraint ($1 \leq n \leq 6$), $k_{m\phi}$ the force constant and ϕ the torsional angle.

Inversion barriers are used to describe the geometry around trigonally bonded atoms. Inversion barriers maintain planarity or, for trigonally bonded centers with a lone pair, prevent inversion around the center through a planar intermediate. For three atoms, l, j, and k bonded to a central atom i, AMBER describes inversions as improper torsions, with the angle ψ between the ijk and ijl planes determining the energy:

$$E_\psi = \sum_{i=1}^{P} \frac{1}{2} k_i^\psi cos(n(\psi_i - \psi_i^{eq}))$$

where P is the number of inversion centers, ψ^{eq} is the equilibrium inversion angle, and n is the periodicity. Here, $n = 2$ represents planar angles and $n = 3$ for tetrahedral angles. In AMBER, $n = 3$ is used for tetrahedral carbons having one implicit hydrogen. For the DREIDING forcefield, a different inversion formulation is used which relates the angle θ between the il bond and the ijk plane with an umbrella term:

$$E_\theta = \sum_{i=1}^{P} \frac{1}{2} k_i^\theta (cos(\theta_i) - cos(\theta_i^{eq}))^2.$$

These conventions are illustrated in figure 1.

Nonbonded terms are also a major contributor to the energy of molecular conformations. The AMBER forcefield describes these nonbond interactions with three terms: van der Waals, electrostatic and hydrogen bonding. For 1-2 (covalent)

bonded and 1-3 (angle) bonded atoms, the nonbond contribution to the intramolecular force is included in the valence term and therefore not explicitly calculated. For AMBER, 1-4 (torsion) nonbond interactions are scaled by 0.5, while DREIDING forcefield parameters have been optimized without such scaling.

The van der Waals term describes both the London dispersion attractive interaction and a higher order term reflecting electron exchange repulsions.

$$E_{vdW} = \sum_i \sum_{j>i} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}$$

where A_{ij} and B_{ij} are constants derived from empirically and theoretically derived well depths and equilibrium interaction distances in small molecule crystals and r_{ij} is the distance between atoms i and j. For off-diagonal interactions (*i.e.*, nonbond interactions between unlike atoms), the geometric (AMBER) or arithmetic (DREIDING) mean of the two parameters are used. For the mixed forcefield type calculations that appear in this work, the more widely accepted geometric mean was used.

Electrostatic interactions are calculated from the Coulomb potential function:

$$E_{el} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

where q_i represents the charge on atom i, $4\pi\epsilon_0$ represents geometric and dielectric contributions, and r_{ij} is the distance between atoms i and j. Early work in molecular mechanics and dynamics simulations was conducted in the absence of solvent, which led to unacceptably high Coulombic interaction terms. This was commonly alleviated by scaling the coulombic term by r^{-1}, a technique, referred to as a "scaled

Coulombic," which ostensibly represents the electrostatic screening effect of solvent atoms when they are not explicitly present and has been carefully validated by simulation studies conducted on the Trp repressor-operator complex (Guenot and Kollman, 1992). For the protein structural work reported here, the modeling of the Hin-*hix* complex, simulations were conducted in the absence of explicit water molecules and a dielectric shielding factor of r^{-1} was included. For the simulations reported in this work that were conducted with explicit solvent, this was replaced with the vacuum dielectric constant ($\epsilon = 1$).

The charges publised for the AMBER and DREIDING forcefields were generated from quantum mechanical calculations on small molecules rather than on the molecules that are described by the forcefields, and are not, in general, as accurate a characterization of the electrostatic properties of molecules as the van der Waals parameters in the forcefield are of the Lenard-Jones potential. For the high precision thermodynamic analysis reported here, more accurate charges, derived from pseudospectral-generalized valence bond (ps-GVB) (Ringnalda *et al.*, 1993) calculations conducted on the whole molecule under investigation were used. Ps-GVB is a novel numerical techniqued that uses both a basis set and a Cartesian grid to reduced Hartree-Fock calculations from order n^4 to n^3 which results in a significant speedup for calculations on the large (by *ab initio* standards) molecules simulated here, without a noticable reduction in accuracy. A second advantage of using ps-GVB calculated charges is that full atom charges can be derived from these calculations; the charges published for DREIDING and AMBER are reported as the united atom charges on aliphatic carbons (*i.e.*, hydrogen charges and van der Waals characteristics not explicitly included, rather they modify a single hybrid

atom located at the carbon center). For the molecular modeling work, which took advantage of the united atom approximation, AMBER charges were used.

The contribution to the energy from hydrogen bonds is explicitly included in the AMBER and DREIDING forcefields (it is not included in most water forcefields). This term modifies the interaction between heteroatom hydrogen bond acceptors and hydrogen atoms bonded to heteroatom hydrogen bond donors. In the AMBER forcefield, hydrogen atoms are assigned charges but do not have a van der Waals interaction term. AMBER uses a Lennard-Jones 10-12 potential to describe the van der Waals like character of this class of interaction, which "fine tunes" the much larger electrostatic component of hydrogen bonding (Ferguson *et al.*, 1992):

$$E_{Hb} = \sum_{Hbonds} \left(\frac{5 \cdot C_{ij}}{r_{HA}^{12}} - \frac{6 \cdot D_{ij}}{r_{HA}^{10}}\right).$$

The parameters C_{ij} and D_{ij} are determined, based on the minimum of the interaction potential and r_{HA} is the hydrogen-acceptor distance. The DREIDING forcefield includes neither charges nor van der Waals terms for hydrogens. Rather it characterizes heteroatom bonded hydrogen-heteroatom interactions as follows (where D is the hydrogen bond donor, H is the heteroatom bonded hydrogen and A is the heteroatom acceptor):

$$E_{Hb} = D_o \sum_{Hbonds} \left(\frac{5 \cdot C_{ij}}{r_{DA}^{12}} - \frac{6 \cdot D_{ij}}{r_{DA}^{10}}\right) cos^2\theta_{D-H-A},$$

where θ_{A-H-D} and θ_{AA-A-H} represent the acceptor-hydrogen-donor angle and acceptor+1-acceptor-hydrogen angles. For the simulations simultaneously using the TIP3P forcefield and AMBER or DREIDING, explicit hydrogen charges were

included and the explicit hydrogen bond term was excluded for solvent-solute and solvent-solvent interactions because the TIP3P forcefield was developed specifically to characterize water molecule interactions without using an explicit hydrogen bonding term. For this forcefield, the dipole moment set up by the charge distribution in the water molecule accurately describes the hydrogen bonding potential of water.

Energy Minimization

The potential energy calculated by summing the energies of various interactions is a numerical value that is characteristic only of that particular conformation. While this number is of utility in evaluating a particular conformation, it is seldom of general use, because it tells very little about the true nature of the conformations accessible to the system, and is often dominated by a few very unfavorable interactions. Energy minimization is a technique which can be used to explore conformation space to search for local energy minima. Energy minimization is usually performed by gradient optimization: atoms are moved iteratively so as to reduce the net forces acting on them. The minimized structure has small forces on each atom and therefore serves as an excellent starting point for further searches of conformation space.

Energy minimization is usually performed in Cartesian coordinates by optimizing along pathways in $3n$ -dimensional space, where n is the number of particles in the system. The pathway along which optimization occurs is usually the gradient ∇ E, where each of the components of the gradient is described as:

$$\nabla_x E = \frac{\partial E}{\partial x}.$$

A variety of techniques can be used to interpolate a path along the gradient to a set of atomic coordinates at which ∇E is closer to zero and the process repeated iteratively until some predetermined convergence criterion is met (Press *et al.*, 1989).

It is also possible to minimize the energy of a conformation by optimizing the torsional rather than Cartesian degrees of freedom. Here the minimization occurs in m-dimensional space, where m is the number of dihedral angles. Torques, which are the derivatives of the potential with respect to torsion angles, are minimized instead of Cartesian forces. It has been shown that torque minimization, when followed by a Cartesian minimization, can produce an overall lower energy conformation than Cartesian minimization alone because torque minimization is better suited for avoiding local minima (Mathiowetz, 1993).

The Equations of Motion

The potential energy of a single static conformation generated by applying a forcefield based calculation to the conformation is a measure of the enthalpy of that conformation, which is a static characteristic of that conformation and is not, in and of itself, a measure of any of the "dynamic" properties of the system. But if the forcefield is an adequate description of the energy of the system, it will also allow us to calculate the intra-molecular forces inherent in the system, and from that we can follow the time dependent evolution of a system using Newton's equations of motion and calculate dynamic properties. Because the Newtonian integrals cannot

be solved in closed form for systems of any realistic size, numerical solutions must be calculated.

It is this ability to calculate the time dependent evolution of molecular systems that is the heart of the utility of molecular dynamics. First, since each molecule in a simulation has an associated, fluctuating kinetic energy, local energy minima can be overcome and a thorough search of conformation space can be completed. Second, the conformations produced during a simulation can form a thermodynamic ensemble, from which thermodynamic properties such as heat capacities, diffusion constants and free energies can be determined.

Molecular dynamics calculations evaluate the forces acting on each particle and use these to determine the accelerations these particles undergo. The x component of force associated with a molecular conformational energy, E, is derived as:

$$F_x = \frac{\partial E}{\partial x}$$

which can be coupled with Newton's laws of motion to calculate accelerations (a_x) and time dependent velocities ($v(t)$):

$$a_x = \frac{F_x}{m_i}$$

$$v(t) = v(0) + \int_0^t a_x(t)dt.$$

The integral derived here can almost never be solved exactly, so we must resort to one of a number of numerical solutions (Press et al., 1989). The most popular of these methods is the Verlet algorithm (Rahman, 1964; Verlet, 1967) which has

many formulations. The simulations reported here use the leapfrog Verlet algorithm (Heermann, 1986), in which the velocities at timestep $n + \frac{1}{2}$ are determined from the previous velocities, the timestep, h, and the acceleration:

$$v_x^{n+\frac{1}{2}} = v^{n-\frac{1}{2}} + ha_x^n \ .$$

The new velocities, $v_x^{n+\frac{1}{2}}$, are then used to update the coordinates for timestep $n + 1$:

$$x^{n+1} = x^n + hv_x^{n+\frac{1}{2}}.$$

The new coordinates are used to calculate new energies and forces and the next iteration of the dynamics routine is started.

The timestep of the integration must be smaller than the highest frequency modes of the system for the numerical solution to accurately calculate this integral. For most molecular simulations, this timestep is on the order of 0.5 to 4 femtoseconds.

Torques are calculated as the derivative of the potential with respect to the torsion angles of a molecule rather than the Cartesian coordinates of its atoms. By coupling torques with angular moments of inertia, a torsion-only form of dynamics is generated which can be used to search torsion space quite rapidly. This technique is used in the work reported here as an initial search of conformation space for homology-based modeling techniques (Mathiowetz, 1993).

Monte Carlo Techniques

Molecular dynamics techniques are not the only method of conducting conformational searches and generating ensembles. A second set of techniques, termed Monte Carlo (MC) techniques, can be used to search conformation space for low energy conformations or to generate an ensemble of states for the determination of thermodynamic averages.

MC techniques generate an ensemble of states by randomly altering starting conformations and, in the Metropolis implementation (Metropolis *et al.*, 1953), adding to the growing collection of conformations those conformations that fit specific criteria for inclusion in the ensemble. The criteria used are as follows: For a given conformation i, a new conformation, $i+1$, is generated by randomly translating a subset of the atoms or molecules in the system. This new conformation will always be accepted as part of the ensemble (and as a starting point for the next conformation) if it is lower in energy than i. If conformation $i+1$ is higher in energy than conformation i, then conformation $i+1$ will be included in the ensemble some fraction of the time, with the probability, ρ, of acceptance dependent on the energy difference between the conformations and the relationship:

$$\rho = e^{\frac{E_{i+1}-E_i}{k_b T}}.$$

This procedure can be used to generate a canonical ensemble of states over which thermodynamic averages may be taken.

There are several advantages to MC techniques. From a computational standpoint, MC simulations are simpler because new conformations are generated without considering the force acting on the previous conformation and therefore the first

derivative of the potential does not need to be calculated. For the type of simulations included in this work, the calculation of derivatives does not significantly slow simulations, so this advantage is reduced.

Another advantage of MC techniques is that they often search conformation space more rapidly than molecular dynamics techniques because the random nature of the coordinate shifts used to generate new conformations can often overcome energetic barriers that molecular dynamics simulations would cross only slowly. This can be a significant advantage for some systems, such as searching torsion space for a low energy torsion. MC techniques are used for that purpose in some of the work reported here.

The primary disadvantage of MC simulation techniques is that the protocol used to generate new conformations often leads to a significant rejection rate. Thus, many of the conformations generated are analyzed but not included in the ensemble, thus slowing the generation of the ensemble. For some investigations, such as a search for low energy torsional conformations, this problem is easily circumvented and high efficiency can be achieved (Mathiowetz, 1993).

Characteristics of the Ensemble

A variety of molecular dynamics methods have been developed to generate ensemble of states for the determination of average thermodynamic properties. All of these techniques differ in their approach to scaling the total kinetic energy of the system, and thus how they regulate the temperature of the simulation.

The simplest way to regulate the temperature of a molecular dynamics simulation is to scale the velocities of every atom at some point during the simulation. The velocities, and thus the temperature of the system, can be scaled at every timestep, but this eliminates the random thermal fluctuations that characterize real systems. These fluctuations arise from the fact that energy, while conserved, does partition between the kinetic energy associated with atomic motion and the potential energy of bonds and nonbond interactions in the system. These fluctuations are apparent as the heat capacity of the system. Because the relative magnitude of such fluctuations is inversely proportional to the total number of atoms in the system, for large systems the fluctuations are small and not completely damped by frequent scaling so realistic heat capacities and other thermodynamic averages can be calculated.

Another approach to temperature regulation in simulations is to simply let the simulation run without removing potential or kinetic energy. Such simulations are termed microcanonical and are often used in molecular dynamics, but these too often exhibit an unrealistically low level of energy fluctuations.

Scaled velocity and microcanonical simulations will, eventually, sample all of the conformation space available to truly canonical methods, but should tend to slow convergence because they sample conformation space more slow. This is due to the lack of realistic kinetic and potential energy fluctuations which prevents these simulations from rapidly sampling less favorable conformations that may well make important contributions to the average under investigation.

In the last decade, molecular dynamics simulations have been extended from the microcanonical to the canonical by coupling systems to an external heat bath

using a pair of conjugate variables that are exponentially linked to the velocity scaling factor via a virtual time factor (Nosé, 1984). Hoover has extended this result into the real time domain (Hoover, 1985). Both formualtions have seen widespread use. Other, less rigorous techniques to simulate a molecular system in equilibrium with an external heat bath include the scaling of the velocities of a limited subset of atoms at the physical edge of the system, and through them the entire system becomes coupled to a heat bath, without the difficulties associated with constant scaling of all atoms (*e.g.*, Hard and Nilsson, 1992).

The thermodynamic simulations reported use the scaled velocity technique (see appendicies III and IV). The simulations are of sufficiently large size (typically 4500 atoms) that thermal averages converge in a reasonable simulation length despite this scaling (see appendix II).

The simulations reported here were all conducted using a constant temperature, constant volume ensemble so the free energies determined would rigourously correspond to Helmholtz free energies, not the more commonly reported Gibbs (constant temperature and pressure). Because the change in molar volume seen in these simulations is small (in all cases less than 0.4%; usually substantially less), the two types of free energy are expected to be essentially equivalent at the accuracy levels typically seen with this type of simulation.

The Accuracy of the Forcefield-Molecular Dynamics Approach

Despite the approximations inherent in the forcefield description of molecular

energies and forces, numerous studies have indicated that it is sufficently accurate to realistically describe the structure and energetics of proteins and solutions.

For homogeneous systems, such as a periodic box of water molecules, detailed verification of the structural and dynamic behavior predicted by simulations is possible, and has been achieved for such diverse physical parameters as heat capacities, density, average molecular potential, and diffusivity (Stillinger and Rahman, 1974, Jorgensen *et al.,* 1983). For the heterogeneous sytems that characterize biology, experimental verification is somewhat more problematic, but it has also been achieved through a variety of simulation approaches used to describe diverse experimental approaches.

The accuracy of the forcefield description of proteins was first assessed by conducting long scale simulations of a protein of known structure, bovine pancreatic trypsin inhibitor, which demonstrated remarkable stability to such simulations (McCammon *et al.*, 1977). Simulations of more than 200 picoseconds have been conducted on the protein *in vacuo* and 80 picoseconds in solution (Levitt and Sharon, 1988). The root mean square (rms) deviation between the latter time-averaged structure and the starting crystal structure was 1.13 Å for all atoms. This compares exceptionally well to the 1.10 Å rms deviation observed between the different crystal forms of the protein indicating that no major systematic errors were introduced during even these relatively long simulations.

The fast motions sampled by molecular dynamics are also seen in X-ray diffraction data, in which they cause atoms to appear to occupy more space (in a time averaged set of diffraction data) than they would in a rigid molecule. Comparisons

of the atomic mobility observed in molecular dynamics simulations of proteins with this experimental determination of molecular motion demonstrates a high degree of correlation (Aqvist *et al.*, 1985) and is further evidence of the general accuracy of the forcefield description.

Free energy simulations, using a perturbation thermodynamic approach (see chapter 4) offer a very demanding test of the accuracy of the forcefield description. Examples in the literature include the determination of the relative free energy of solvation of bromide and chloride, (Lybrand, 1985a), the solvation of alkanes, amino acids and dipeptides (Sun *et al.*, 1992); and the energetics of inhibitor binding (Bash *et al.*, 1987). The correlation of these demanding simulations with experiment are illustrated in table 1. This technique and its accuracy will be addressed more fully in chapter four.

Periodic Boundary Conditions and Nonbond Cutoffs

In order to simulate a non-finite sample and avoid surface and end effects, a technique called periodic boundary conditions (PBC) can be used in which molecules at the edge of a defined PBC box be affected by forces generated from interactions with molecules from the opposite face of the box. Simulations run in this way mimic the effects of an infinite array of periodic cells with identical molecular configurations in each cell. Obviously, calculating an infinite number of nonbond interactions is impossible, so some form of long range cutoff must be used beyond which nonbond interactions are not calculated.

Table 1: A collection of molecular dynamic perturbation thermodyamics results for a variety of systems, indicating the type of accuracies achievable with the force-field/molecular dynamics description of molecular systems (from Lybrand *et al.*, 1985a; Lybrand *et al.*, 1985b; Jorgensen *et al.*, 1988; Rao *et al.*, 1987).

Transformation	Simulation R.F.E.(kcal/mol)	Experimental R.F.E.(kcal/mol)
$Cl^- \rightarrow Br^-$	3.2 ± 0.2	3.15
SC-24:$Cl^- \rightarrow$ SC-24:Br^-	4.15 ± 0.35	4.0
CH_4(aq) $\rightarrow$ Cl^-(aq)	-79.1 ± 2.0	-77.0
Subtilisin 155Asn$\rightarrow$Ala Δ G_{cat}	3.4 ± 1.1	3.67

Even for simulations of finite systems of size n, the number of nonbond interactions scales roughly with n^2 and quickly comprises the bulk of computation time for systems of any reasonable size. A number of nonbond cutoff schemes have been developed to minimize the total computational power needed to accurately calculate the intermolecular forces for a given size system, including some which scale with n. For the simulations reported in this work, n^2 type nonbond techniques were used.

One approach to the problem of nonbond cutoffs is called minimal image in which each atom in the cell sees the minimum image necessary for all intracell nonbond interactions to occur. This is done by translating the apparent position of interacting atoms such that all interaction distances are less than one half of the

unit cell dimensions (ucd). The effect of this form of simulation is that each atom has acting upon it the forces generated as if it were at the center of the unit cell.

$$r_x = x(i) - x(j)$$

$$\text{if } (r_x > ucd_x) \text{ then } r_x = r_x - \frac{ucd_x}{2}$$

$$\text{if } (r_x < ucd_x) \text{ then } r_x = r_x + \frac{ucd_x}{2}$$

Minimal image simulations, in our experience, lead to excessively deep local minima, causing slow convergence of equilibrium properties and inappropriately low diffusion constants. They are also computationally intensive because the interaction of every atom with every other atom in the cell must be explicitly calculated.

A less computationally intense method for nonbond cutoffs is the straight cuttoff in which interactions of greater than a given distance are ignored. Unfortunately, for highly polar molecules like water, this leads to charge-charge interactions when only part of a water molecule is within the nonbond cutoff distance. For any reasonable cutoff distance (typically on the order of 8-12 Å) this leads to very large fluctuations in the energy and force on a given molecule and leads to an inaccurate representation of the equilibrium ensemble of the system. For simulations of proteins, which are neither as highly polar as water nor as demanding in terms of an accurately determined ensemble of states, the 8 Å straight cutoff was used without periodic boundary conditions.

For simulations of solvent, the neutral cell nonbond cutoff method avoids problems with these artificial long range charge interactions by cutting off calculations

at the whole water molecule level. If the oxygen atom of a given water molecule is within the cutoff distance then the interaction between the observed water molecule and all three atoms in the distant molecule is used. This technique eliminates some of the larger energy and force fluctuations seen in the straight cutoff method, but still exhibits unacceptably high fluctuations for any reasonable size cutoff distance due to the very high dipole moment of water molecules.

To reduce these artificial energy and force fluctuations and the excessive structure (and computational complexity) seen with a minimal image approach we have used the spline technique in which energy and force are scaled in a distance dependent fashion to reduce long range nonbond interactions. This ramping is commonly conducted only over a limited region of space. That is, short range interactions are typically calculated without the spline scaling, interactions of intermediate range are scaled, and interactions at greater than some defined cutoff distance are ignored. For the thermodynamic work reported here, a cubic spline (BIOGRAF, 1992) was used:

$$E_{ij} = E_{ij}\frac{(r_{off}-r_{ij})^2(r_{off}+2r-3r_{on})}{(r_{off}-r_{on})^3},$$

where E_{ij} equal zero for $r_{ij} \geq r_{off}$ and not scaled for distances less than r_{on}. Our simulation work has indicated that a smooth cubic scaling from r=0 to r=12 Å: gives the best fit of phonon dispersion curves in solids (Goddard, 1993) and leads to good agreement with the observed heat capacity in simulations of bulk TIP3P water, as reported in appendix II.

One order n method is the Cell Multipole Method (CMM), which approximates (to a very high degree of accuracy) long range van der Waals and electro-

static interactions by dividing the volume of a simulation into a number of cells and summing over these cells to determine monopole, dipole and higher order terms to describe the field generated by that region in space. The calculation of the interaction energy of a given atom with this multipolar expansion of the far field can be substantially more computationally accessable than the explicit computation of the interaction of every pair of atoms, and scales as $n\log n$. If volume is divided in a geometrically expansive fashion, such that the cells over which the expansion is taken are get larger as the distance between the atom under consideration and the cell increases, then this problem scales with n without a significant dimunition of accuracy (Ding *et al.*, 1992). Although the accuracies available by using this technique are generally high (on the order of 0.01% in total energy and 0.2% in rms force), CMM tends, in our experience, to lead to non-convergent thermodynamic averages for highly polar systems like water, and while used for some early equilibration work, was not used for determining the thermodynamic results reported here.

The Limitations of High Frequency Internal Modes

The integration timestep of a molecular dynamics simulation is limited to a small fraction of the period of the fastest internal vibrations of the system, typically C-H or O-H bond vibrations. As the period of this movement is characteristically on the order of a few femtoseconds, typical dynamics time steps are limited to approximately 0.5 to 1.0 femtoseconds for systems containing explicit hydrogens.

Over the years, a variety of approaches have been developed to overcome this

inherent limitation, with varying degrees of success. We have explored and exploited several in this work.

According to the harmonic oscillator description of bond vibrations, the characteristic time of a bond vibration is proportional to the square root of the mass of the atoms vibrating:

$$\text{period} = \omega^{-1} = (\frac{m}{k_b})^{\frac{1}{2}}.$$

From this, we see that a simple approach to increasing integration step size is to increase the mass of the hydrogen atom, which will increase the characteristic vibration time of the bond, but not effect equilibrium bond lengths or ensemble averages for the system (Lybrand, 1985a). The typical improvement possible using this technique for solvation studies is about a factor of 2 to 4 in timestep length. One drawback of this approach is that while the increase in molecular mass does not affect the equilibrium characteristics of the system, it does slow system diffusion rates and thus slows the convergence of equilibrium based averages.

The SHAKE algorithm (Ryckaert and Berendsen, 1977) actively removes selected high frequency internal modes by iteratively reallocating the force and velocity of each atom (net force and total momentum remain constant) so that no net force or velocity remains along the bond axis. The algorithm can also be used to eliminate forces and residual velocities that would contribute to high frequency angle modes. The iterative process is easily parallelizable and scales with n. Using this technique, we are routinely able to conduct solvent simulations with integration timesteps of up to 4 femtoseconds, and collect high quality thermodynamic averages with timesteps of 1 femtosecond.

Conclusions

Advances in both the methodology and computational power of molecular dynamics have created a set of computational tools that are begining to achieve a high degree of utility in chemistry and biochemistry. The forcefield description of molecular energies and forces, and the application of Newton's laws of motion and numerical integration to harness this description to calculate realistic atomic and molecular motions has been repeatably demonstrated to be both accurate and useful in the analysis of macromolecular structure, function, and energetics. In the remaining chapters of this thesis we will report on the application of the techniques described above to such diverse problems as modeling the structure of a protein-DNA complex and the estimation of the relative free energy of the sequence specific binding of proteins and drugs to their cognate sites on DNA.

References

Aqvist, J., van Gunsteren, W. F., Leijonmarck, M., and Tapia, O., (1985), *J. Molec. Biol.*, **183**, 461-477

bigskipArad, D., Langridge, R., and Kollman, P. A., *J. Am. Chem. Soc.*, (1990) **112**, 491-502

Barnes, J. E., *Nature*, (1989), **338**, 123-126

Bash, P. A., Chandra Singh, U., Brown, F. K., Langridge, R., and Kollman, P. A., (1987), *Science*, **235**, 574-576

Beveridge, D. L. and DiCapua, F. M., (1989), *Annu. Rev. Biophys. and Biophys. Chem.*, **18**, 431-493

BIOGRAF/POLYGRAF (1992) copyright Molecular Simulations, Inc. (Pasadena CA)

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathat, S., and Karplus, M., (1983), *J. Comp. Chem.*, 4, 187- 217

Czerminski, R., and Elber, R., (1991), *Proteins*, **10**, 70-80

Ding, H. Q., Karasawa, N., and Goddard, W. A., III, (1992), *J. Chem. Phys.*, **97**, 4309-4315

Ferguson, D. M., Pearlman, D. A., Swope, W. C., and Kollman, P. A., (1992), *J. Comp. Chem.*, **13**, 362-370

Goddard, W. A., III, (1993), *M. S. C. Technical Note*, **119**, (Materials Simulation Center, California Institute of Technology, Pasadena CA)

Guenot, J., and Kollman, P. A., (1992), *Protein Sci.*, **1**, 1185-1205

Hard, T., and Nilsson, L., (1992), *Nucleosides and Nucleotides*, **11**, 167-173

Heermann, D. W., (1986), *Computer Simulation Methods in Theoretical Physics*, (Springer-Verlag, Berlin)

Hoover, W. G., (1985), *Phys. Rev. A*, **31**, 1695-1697

Jorgensen, W. L., Chandrasekhar, J., Madurea, J. D., Impey, R. W., and Klein, M. L., (1983), *J. Chem Phys.*, **79**, 926-935

Jorgensen, W. L., Blake, J. F., and Buckner, J. K., (1988), *Chem. Phys.*, **129**, 193-200

Karplus, M., and Petsko, G. A., *Nature*, (1990), **347**, 631-634

Levitt, M. L., and Sharon, J. A., (1988), *Proc. Nat. Acad. Sci. U.S.A.*, **85**, 7557-7561

Lybrand, T. P., Ghosh, I., and McCammon, J. A., (1985a), *J. Am. Chem. Soc.*, **107**, 985-986

Lybrand, T. P., McCammon, J. A., and Wipff, G., (1985b), *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 833-836

Mathiowetz, A.M., (1993), *Dynamic and Stochastic Protein Simulations: From Peptides to Viruses*, Thesis, (California Institute of Technology, Pasadena Ca)

Mayo, S. L., Olafson, B. D., and Goddard, W. A. III, (1990), *J. Phys. Chem.*, **94**, 8897-8909

McCammon, J. A., Gelin, B. R., and Karplus, M., (1977), *Nature*, **267**, 585-590

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H., (1953), *J. Chem. Phys*, **21**, 1087-1092

Nosé, S., (1984), *J. Chem. Phys.*, **81**, 511

Plaxco, K. W., Mathiowetz, A. M., and Goddard, W. A., III, (1989), *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 9841-9845

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T., (1989), *Numerical Recipes*, (Cambridge University Press, Cambridge, MA)

Rahman, A., (1964), *Phys. Rev.*, **136**, 405

Rao, S. N., Singh, U. C., Bash, P. A., and Kollman, P. A., (1987), *Nature*, **328**, 551-554

Ringnalda, M. N., Langlois, J-M., Greeley, B. H., Russo, T. V., Muller, R. P., Marte, B., Won, Y., Donnelly, R. E. Jr., Pollard, W. T., Miller, G. H., Goddard, W. A. III, and Freisner, R. A., (1993), *PS-GVB*, v0.08, Schroedinger, Inc., (Pasadena, CA)

Ryckaert, G. C. and Berendsen, H. J. C., (1977), *J. Comp. Phys.*, **23**, 327-340

Stillinger, F. H., and Rahman, A., (1974), *J. Chem. Phys*, **60**, 1545-1557

Sun, Y., Spellmeyer, D., Pearlman, D. A., and Kollman, P. A., (1992), *J. Am. Chem. Soc.*, **114**, 6798-6801

Verlet, L., (1967), *Phys. Rev.*, **159**, 98

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. Jr., and Weiner, P., (1984), *J. Am. Chem. Soc.*, **106**, 765-783

Weiner, S. J., Kollman, P. A., Nguyen, D. T., and Case, P. A., (1986), *J. Comp. Chem.*, **7**, 230 - 242

Chapter 2

Structural Predictions of Protein-DNA Interactions

Abstract

The sequence specific interactions between protein and DNA provide the most basic level of regulation in biology and as such, offer insights into the mechanisms of the regulation of genetic expression and a tantalizing site for designing highly specific therapeutic agents. For these reasons, protein- DNA complexes have been the subject of intensive experimental and theoretical investigation. The focus of the work discussed in this chapter is the development of an atomic level model for the structure of the DNA binding domain of the protein Hin recombinase using an experimentally constrained molecular dynamics routine we have developed. The protein has been the subject of numerous genetic and biochemical investigations which have generated a diverse assortment of experimental data that can be used to constrain the volume of conformation space searched by a molecular dynamics based technique for determining protein structure. The model generated is in close agreement not only with experimental data used to constrain the dynamics, but with numerous experimental results not included in the modeling effort. Further, the model makes a variety of predictions about the structure and characteristics of the protein, many of which have already been experimentally verified. It offers an example of a general approach to the prediction of structures involved in this important molecular interaction.

Introduction

To date, no one has succeeded in accurately predicting the native structure of a protein from its amino-acid sequence using basic principles. It is impossible, for example, to take a fully extended polypeptide chain, run a molecular dynamics simulation and have it fold automatically into the correct structure. Why is this? First, the time required to fold up a protein in solution, (usually estimated to be on the order of 10^{-3} to 10^{3} seconds) which is prohibitively long to simulate because typical integration step times are on the order of femtoseconds. With each step requiring at least a few seconds to calculate on today's fastest computers, it would require millenia to simulate the folding path of a single protein. Second, even if there were enough computer time to run a protein folding simulation, because the free energy of denaturation of proteins is very small, (on the order of 0.01 kcal/atom) a frightening level of accuracy would be required to correctly simulate the entire folding pathway (Karplus and Petsko, 1990).

Simulations, however, are not conducted in an intellectual vacuum, and need not be able to solve the folding problem from basic principles to still be of significant applicability. If the basic potential function of molecular dynamics is supplemented by experimentally derived constraints, a simulation may be able to find the correct structure by searching conformation space for as short as a few tens of picoseconds of simulation time. A great deal of work has been undertaken using NMR determined inter-atomic distances to constrain molecular dynamics conformational searches (Schmitz *et al.,* 1992), but suitable constraints can be derived from a variety of other experimental studies, including such properties as the known substrate

interactions of a protein, homology to proteins of known structure, known cysteine bonds, or circular dicroism based analysis of secondary structure. It is the use of homology and substrate interactions to derive constraints for molecular dynamics that we will discuss in this chapter as we report the use of a technique, termed homology modeling, to predict the atomic resolution structure of the DNA binding domain of the protein Hin Recombinase in complex with the *hix* site of DNA.

Hin Recombinase is a 190 amino acid site specific DNA recombinase from the bacteria *Salmonella typhimurium* (Zeig *et al.,* 1977) that functions to switch expression between two antigenically distinct flagellar filament proteins by inverting a 995 bp region of the *Salmonella* chromosome which controls flagellin expression (Johnson *et al.,* 1984). The mechanism for Hin mediated inversion, illustrated in figure 1, requires the presence of two protein cofactors: FIS, which binds to a cis acting recombinational enhancer sequence present on the inversion element and HU, a histone like protein thought to facilitate the DNA bending necessary for the formation of the correct protein-DNA complex geometry. The recombination event occurs between two crossover sites, designated *hix* L and *hix* R, in an inverted repeat configuration on supercoiled substrates (Johnson *et al.,* 1984). Each *hix* site is a 24 base pair long psuedo-twofold symmetric site at which Hin binds as a dimer (Johnson *et al.*, 1984). Hin belongs to a family of recombinases that are cross complimentary and that include Gin from phage Mu, Cin from phage P1 and Pin from the e14 element of *E. coli*, all of which exhibit approximately 80% sequence identity as shown in figure 2 (Plasterk and van de Putte, 1984). The recognition elements of these proteins exhibits a consensus sequence:

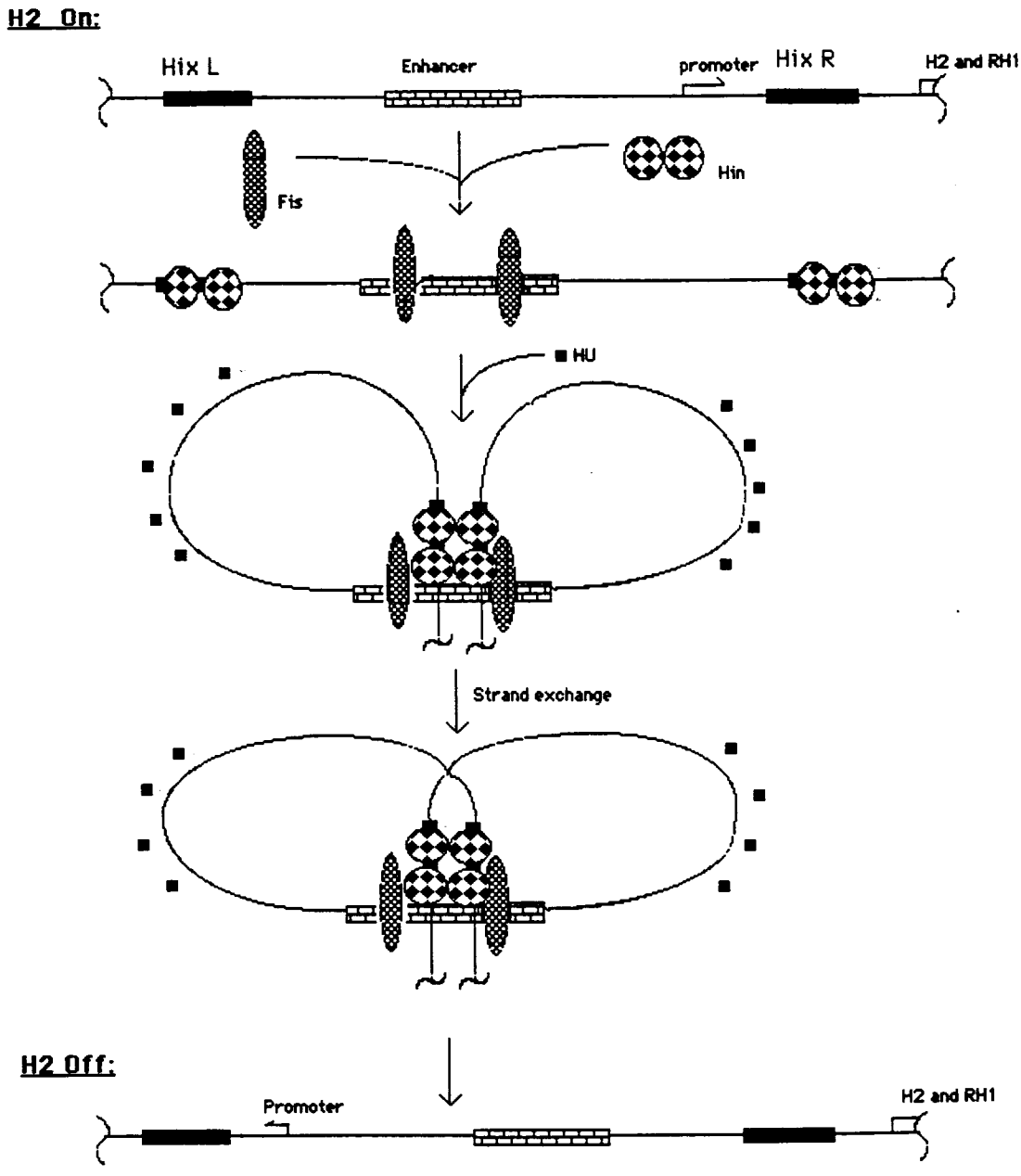


Figure 1: The mechanism of Hin mediated recombination (Johnson *et al.*, 1986).

5'-TTaTC-AAACCAGGTTT-GAtAA-3'

(where lower case letters denote a non-perfect conservation at that site).

The binding of Hin recombinase to the *hix* recombination site has been extensively characterized both *in vivo* and *in vitro*. Hin binds the *hix* site cooperatively as a dimer with an apparent dissociation constant of 40 nM in 100 mM sodium chloride (Glasgow *et al.*, 1989a), corresponding to a $\Delta G_{binding}$ of approximately 10 kcal/mol). DNase footprinting data indicates that the Hin protein covers a 32 base pair region of the *hix* site (Glasgow *et al.*, 1989b). The binding of Hin to the *hix* site has also been investigated by a variety of chemical protection and interference assays, which have defined the primary recognition sites in the major and minor groove of the *hix* element, as indicated in figure 3 (Glasgow *et al.*, 1989a).

The first evidence localizing the DNA binding domain in Hin recombinase came from analogy to the chymotrypsin cleavage of $\gamma\delta$ resolvase from transposon $\gamma\delta$, which produces a 43 amino acid protein that exhibits identical binding selectivity to the uncleaved resolvase (Bruist *et al.*, 1987). The corresponding 52 amino acid segment of Hin, which is closely related to $\gamma\delta$ resolvase, was synthesized and determined to contain the entire DNA binding specificity element of the native protein (Bruist *et al.*, 1987) as measured by DNase footprinting and competitive inhibition of the enzymatic activity of the native Hin protein. Polypeptides with truncated amino-termini exhibited reduced binding until at a length of 31 residues, binding was no longer detected (Bruist *et al.*, 1987). Longer polypeptides did not exhibt significantly enhanced binding. Polypeptides shortened at the carboxy terminal end of the protein exhibit no reduction in binding until at least six residues

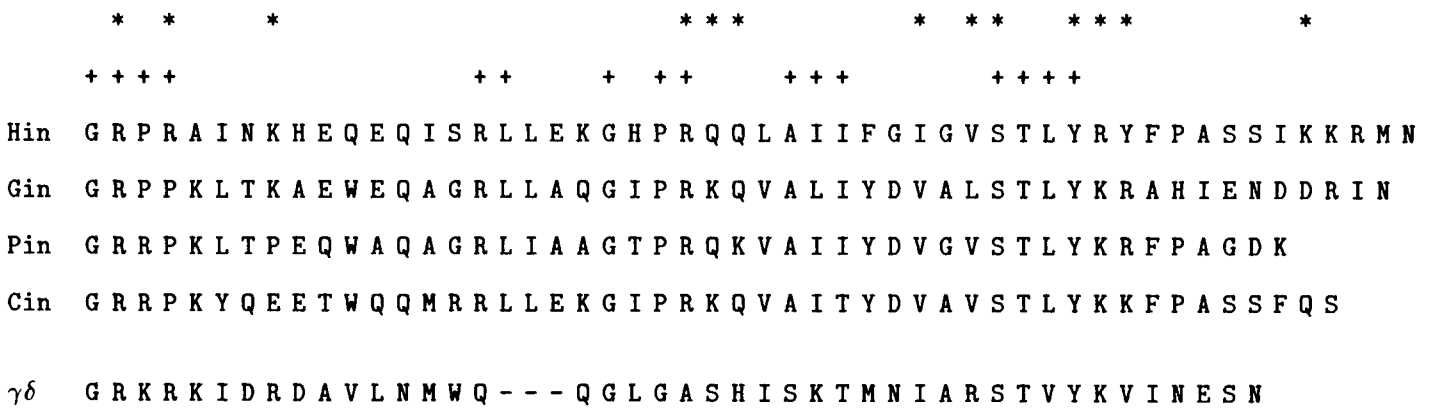


Figure 2: The Hin family of recombinases. Perfect conservation in the family is marked by a +. Note that in addition to the perfectly conserved sites, there are numerous sites that exhibit only chemically conservative mutations. Sites marked with astrisk denote positions that are postulated to form DNA contacts in the Hin-*hix* complex. For comparsion, the C-terminal 43 residue chymotrypsin cleavage product of $\gamma\delta$ is also shown, with a 3 residue gap for optimal sequence match.

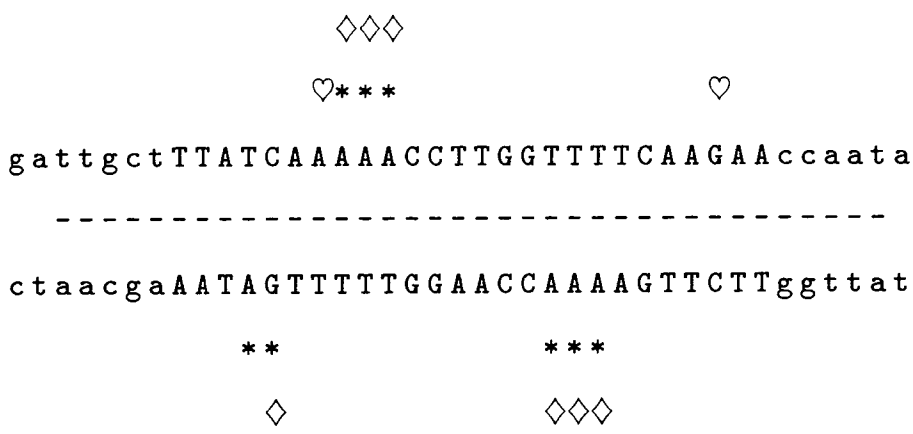


Figure 3: Methylation protection and interference assay results for the *hixL* site. *Dashes* indicate bases protected from DNase footprinting, *asterisks* denote purine sites protected from methylation by Hin binding, *diamonds* represent purine sites at which methylation prevents complex formation, and *hearts* indicate sites at which methylation is enhanced by complex formation. Note that adenine is methylated in the minor groove, while Guanine is methylated in the major groove. (From Glasgow *et al.*, 1989a.)

have been removed (Mack *et al.*, 1990). Fis and Hu are not necessary for DNA binding (Glasgow *et al.*, 1989a)

The 52 amino acid DNA binding domain of Hin exhibits the same sequence specificity, identical chemical protection patterns, and approximately half the free energy of association measured for the dimeric native enzyme (Bruist *et al.*, 1987). This 52 amino acid protein, shown to contain all of the elements necessary for the recognition of the *hix* element of DNA, that was the basis of our efforts to model the structure of the DNA binding domain of Hin recombinase.

Generating the Constraints

The DNA binding domain Hin recombinase has been the focus of a variety of genetic, biochemical and NMR based experimental investigations that allow for the development of structural constraints for molecular dynamics based simulations of its structure. Three types of experimentally derived constraints were used to reduce the volume of conformation space that had to be searched to develop a plausible model for the structure and function of this domain. First, as noted above, the domain exhibits weak sequence similarity to the helix-turn-helix family of DNA binding proteins. Second, the structure of the substrate that this protein binds with, DNA, is large and relatively well characterized, that can be used as a template to guide folding and limit conformational searches. Finally, by using chemical modification data we were able to predict the approximate geometry of the interaction between the *hix* site and the amino terminus of the protein.

The helix-turn-helix domain is found in many DNA binding proteins and several members of the protein family have been solved structurally, including λ Cro (Ohlendorf *et al.,* 1982), λ Repressor (Jordan and Pabo, 1988), and the Trp Repressor (Otwinowski *et al.,* 1988). The Hin DNA binding domain exhibits approximately 20% similarity with Cro and other members of the helix-turn-helix family of DNA binding proteins. While this level of identity can easily occur randomly between unrelated proteins, a high degree of chemical similarity between the residues of Hin and known helix-turn-helix proteins is indicative of its true homology to the family (Dodd and Egan, 1990; Sluka *et al.,* 1990). The sequence lineup and various putatively conserved elements are shown in figure 5.

The Hin protein forms high affinity complexes with several DNA elements whose sequence has been determined. Since the structure of B-form DNA is reasonably constant and well defined, these sequence elements can be used as a template on which to fold the Hin binding domain. A set of semi-quantitative chemical cleavage experiments on the naturally occurring *hix* sites and several fortuitous low affinity sites found in the plasmid pBR322 (Sluka, 1988) allowed for the determination of a putative optimal binding sequence:

5'–TTCTCCAAAA–3'

3'–AAGAGGTTTT–5'

This sequence element was used for the simulation work reported here.

The Hin DNA binding domain has been the subject of intensive chemical modification experiments that have aided in the localization of the domain on the *hix* site. Of primary interest for model building efforts, the addition of the DNA cleaving functionality Fe(II)-EDTA to the α-amino group of the binding domain allowed for the low resolution localization of the amino terminus in the minor groove of the *hix* site at positions 4 and 5 (Sluka *et al.*, 1987; Sluka *et al.*, 1990). This "phasing" information was very helpful in constraining the otherwise highly mobile amino terminus of the protein.

Modeling Techniques

Homology modeling is predicated on the observation that homologous proteins tend to remain structurally similar through long periods of evolutionary separation.

It has been shown that distantly homologous proteins, with sequence similarities of only 20%, tend to exhibit no more than a 2 Å rms deviation in backbone atom location in their common elements (Chothia and Lesk, 1986). Conformational changes of this magnitude are easily sampled with currently available molecular dynamics simulations (Karplus and Petsko, 1990). The question still remains, how do we convert a naive starting guess based on homology to a refined structural prediction? We have developed an iterative molecular mechanics/dynamics simulation technique to sample conformation space biased by such starting guesses. In the remainder of this chapter we will report on the application of this technique to the structure of the DNA binding domain of Hin recombinase.

Several methods for modeling complete protein structures from C_α coordinates have been published in recent years (Purisima and Scheraga, 1984; Correa, 1990; Jones *et al.*, 1991; Rey and Skolnick, 1992). We have developed a technique, based on torsion and Cartesian molecular mechanics and molecular dynamics (mm/md) conformational searching, that is well suited for the magnitude of conformational searches required to successfully model protein structures based on homology. Our technique differs significantly from other published techniques, which range from the purely geometeric (Purisima and Scheraga, 1984; Rey and Skolnick, 1992) to methods based primarily on database searches of several consecutive residues (Reid and Thornton, 1989; Jones *et al.*, 1991; Holm and Sander, 1991) or molecular mechanics (Correa, 1990), in that as a torque/Cartesian mm/md conformational search, a rapid and much more thorough search of conformational space can be performed.

The C_α builder technique was used (Mathiowetz, 1993) as an extension of the BIOGRAF molecular simulations program from Molecular Simulations, Incorporated (BIOGRAF, 1992). The calculations reported here were run on a VAX 8650 and visualized on an Evans and Southerland PS-300 graphics workstation. A schematic outline of the steps of our homology modeling technique is shown in figure 4.

During the first stage of the model building process, the protein is created one residue at a time until the entire protein has been built. As each residue is added to the growing chain, its geometry is initially built from the standard peptide geometries in the BIOGRAF peptide library, then the backbone (ϕ, ψ) and sidechain (χ) dihedral angles are minimized using torsional molecular mechanics followed by Cartesian molecular mechanics. A harmonic constraint is then added between the C_α of the residue and the corresponding C_α coordinates. Torsional dynamics simulation using ϕ and ψ are then used to search a pulse of residues (typically five) to search for conformations that pack well and are consistent with the included C_α constraints. The residues in preceeding pulses are held fixed at this point, but are included in the energy calculations. The process occurs sequentially with the addition of each new residue, with each residue participating in several iterations of optimization until the pulse moves beyond it and it is held in a fixed position.

The initial protein conformation generated by the C_α builder is strongly biased towards the homologous structure input and often contains, in part because of this bias and in part due to the moving pulse method of coordinate optimization,

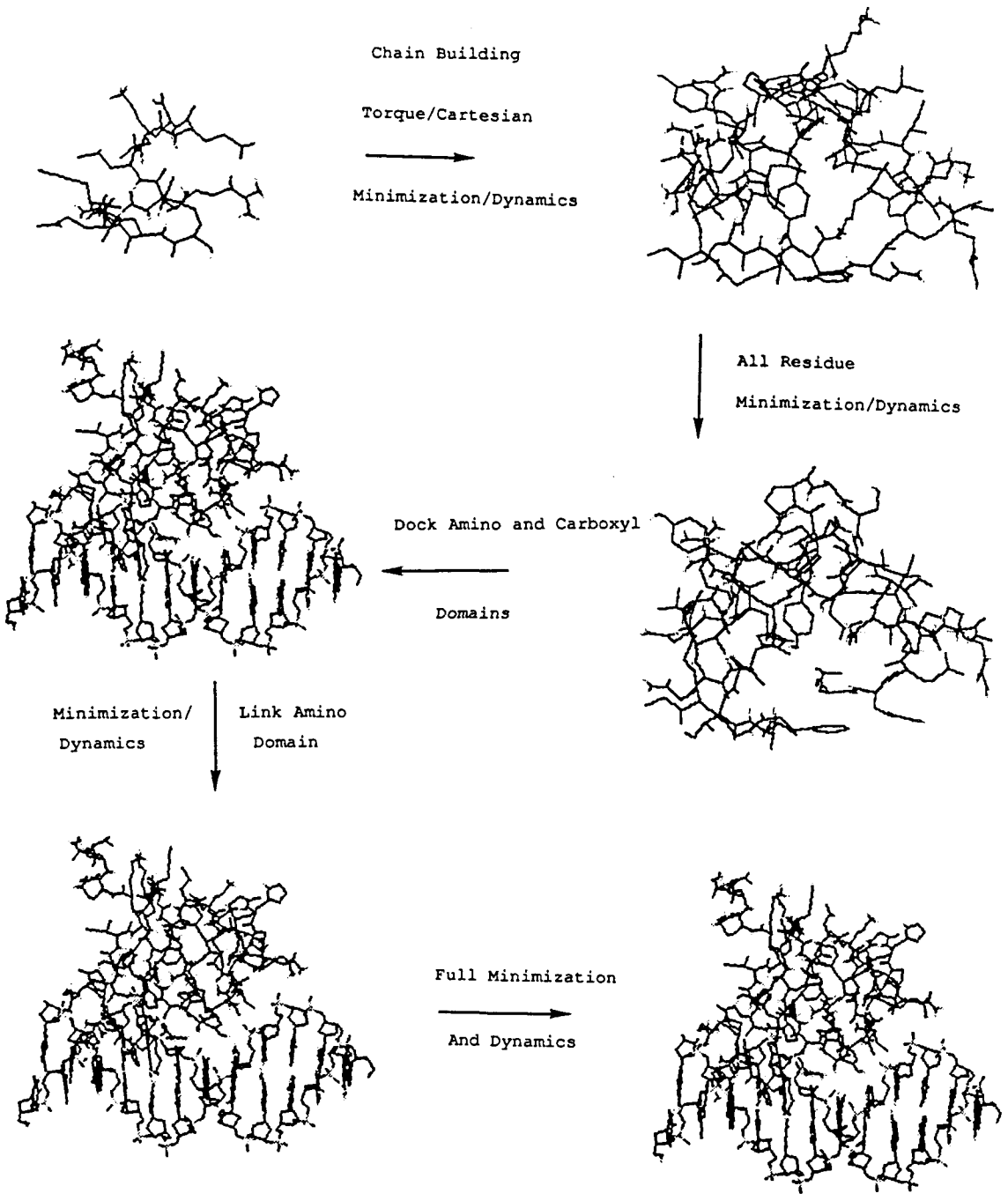


Figure 4: A schematic diagram of the steps of the Cartesian/Torsion mm/md homology modeling technique that was used to generate the Hin-*hix* complex model.

regions of high energy (*i.e.*, inappropriate steric clashes, distorted bonds, angles and torsions, and poor electrostatic interactions). The second stage of this modeling technique is to relax these unfavorable interactions by performing 10 picoseconds of dynamics on the initial conformation in the presence of the C_α constraints (which prevents unfavorable interactions due to the moving pulse technique from moving the conformation very far from the homology biased structure). This step serves primarily to relax artifacts of the iterative optimization, it does not remove model bias, which is then relaxed with 10 picoseconds of dynamics in the absence of C_α constraints.

In order to use the *hix* element as a folding template, the initial Hin model had to be docked to the *hix* site with a reasonable degree of accuracy. Interactive graphics were used to manually dock the putative major groove binding portion of the protein to the area of the *hix* element shown to be recognized in the major groove by protection/interference studies (Glasgow *et al.,* 1989a). In the middle of this region of the *hix* element there is a highly conserved adenine, which corresponds to the adenine in the center of the Cro recognition element. A presumably homologous hydrogen bonding interaction with this site by 174-ser in Hin was used to further refine the docking. For the minor groove element, a long range harmonic term was included to bond the amino terminus of the protein to the minor groove as per the chemical modification data. After this docking, further energy minimization followed by 10 picoseconds of molecular dynamics was conducted with a fixed DNA structure. This served to minimize inappropriate energetic interactions caused by this naive docking protocol and to search local conformation space to find favorable interactions between the protein and its binding site, as well as further refine protein

packing. A final 10 picoseconds of dynamics was conducted with mobile DNA. No significant changes were observed.

Accuracy in Homology Modeling

An estimation of the highest degree of accuracy possible in homology modeling comes from investigations of proteins that are perfectly homologous, *i.e.*, by trying to reconstruct a protein's known structure from a knowledge of its sequence and its C_α coordinates. Studies undertaken with our algorithm on the protein crambin indicate that accuracies on the order of 1.3 Å (rms deviation for all atoms) are achieved when compared to the crystal structure (Mathiowetz, 1993), which is significantly higher than other published attempts on perfectly homologous systems (*e.g.*, Holm and Sander, 1992). Studies undertaken with less perfectly homologous proteins will, of course, be less accurate, with recent published comparisons by a variety of methods resulting in accuracies of 2-5 Å rms deviation for homology based models of ras (Dykes *et al.*, 1993), thrombin (Koymans *et al.*, 1993), and hemoglobin (Srinivasan *et al.*, 1993). To date, no similar study has been undertaken using homology biased molecular dynamics further constrained by substrate interactions and chemical modification data, but we are confident that, given the demonstrably high degree of accuracy our technique achieves on the crambin test case, that the coupling of our technique with these experimentally derived constraints can generate highly accurate models of protein structure.

Conclusions

The work reported here represents an example of an atomic resolution structural prediction based on experimentally constrained molecular mechanics. The techniques used to generate the constraints, and the method used to couple such constraints with a molecular dynamics based search of conformation space have been reported in detail in this chapter. The next chapter of this thesis will discuss the consistency of this model building effort with experimental investigations into the nature of the Hin-*hix* complex that were conducted during or after the model building effort. In addition, some analysis of the consistency of the model with our knowledge of the principles of protein structure will also be reported.

Structural Predictions of Protein-DNA Interactions

The text of the remainder of this chapter consists of an article published in *Proceedings of the National Academy of Sciences U.S.A.*, volume 86, pages 9841-9845, (1989) coauthored by Kevin W. Plaxco, Alan M. Mathiowetz, and William A. Goddard III.

Predictions of structural elements for the binding of Hin recombinase with the *hix* site of DNA

(protein-DNA interactions/ helix-turn-helix motif/molecular modeling structure-function relationships/ *Salmonella typhimurium)*

Kevin W. Plaxco, Alan M. Mathiowetz, and William A. Goddard III

Arthur Amos Noyes Laboratory of Chemical Physics, California Institute of Technology, Pasadena CA 91125

Contributed by William A. Goddard III, September 25, 1989

Abstract Molecular dynamics simulations were coupled with experimental data from biochemistry and genetics to generate a theoretical structure for the binding domain of Hin recombinase complexed with the *hix* site of DNA. The theoretical model explains the observed sequence specificity of Hin recombinase and leads to a number of testable predictions concerning altered sequence selectivity for various mutants of protein and DNA.

A critical problem for fully exploiting the opportunities in protein engineering is to understand the principles determining why a protein binds selectively to a particular base-pair sequence of DNA. Advances in this understanding have been made by a number of indirect studies; however, the difficulties associated with crystallization and analysis of protein-DNA complexes limit the opportunities to obtain structural information directly from crystallography. Our research objectives are to elucidate such interactions by using a combination of molecular mechanics and molecular dynamics simulations constrained by knowledge-based structural predictions. Because of the vast amount of solution-phase experimental data accumulated about the DNA-binding characteristics of Hin recombinase (Bruist *et al.,* 1987; Sluka *et al.,* 1987; Glasgow *et al.,* 1989a; Glasgow *et al.,* 1989b; Hughes *et al.,* 1988; J. Sluka, A.C. Glasgow, M.I. Simon, and P.B. Dervan, personal communication), we selected this system for application of a constrained simulations approach (developed by A.M.M and W.A.G.; to be published at a later date). Utilizing these theoretical techniques in conjunction with information gleaned from various experiments, we have derived a theoretical model of the Hin-DNA binding that is consistent with current experimental data. This model suggests a number of new experiments to test and refine the ideas about the interactions determining site-specific protein-DNA binding.

These studies illustrate what we believe will be an effective mode of elucidating the mechanisms of sequence-specific protein-DNA binding. Experimental techniques such as chemical and enzymatic footprinting, affinity cleavage, and genetics can specify the regions (both sequence and groove location) of DNA sites involved in protein-DNA recognition, and define the structural motifs involved in

protein binding. However, these techniques do not provide detailed atomic-level information about the interactions responsible for site specificity. Theoretical molecular dynamics calculations can provide useful information about interactions at the atomic level, both with current techniques these studies are only practical if the region of protein and DNA involving the interaction is specified. The detailed model can then be used to design experiments that can distinguish subtle differences in the nature of the specific interactions and refine the theoretical model.

Hin recombinase is a 190-amino acid protein from *Salmonella typhimurium* containing both a specific DNA-binding activity and a DNA-recombination activity. Hin recombinase mediates a site-specific recombination between two 26-base-pair elements (*hixL* and *hixR*) separated by 993 base pairs of *Salmonella* chromosomal DNA. Hin binds to the pseudo-dyad symmetric *hix* binding sites as a dimer, with one molecule of the protein at each of the two half-sites comprising the dyad repeat. Purified Hin is able to catalyze a phosphodiester cleavage at the center of symmetry of each *hix* site *in vitro* and in the presence of the proteins Fis and Hu is able to perform strand exchange and religation of the DNA between the two *hix* sites (reviewed in Glasgow *et al.*, 1989b).

Hin is a member of a large family of site-specific recombinases from widely divergent organisms that are homologous members of the helix-turn-helix family of proteins (Bruist *et al.*, 1989). This family includes λ Cro and λ repressors, which have been structurally determined (Ohlendorf *et al.*, 1983; Jordan and Pabo, 1988) as indicated in Fig. 5.

By analogy to the closely related protein $\gamma\delta$ resolvase, Hin was postulated

to contain two domains: a catalytic domain comprising the amino-terminal 138 residues of the protein and a DNA-binding domain consisting of the carboxyl-terminal 52 residues (Bruist *et al.,* 1987) and containing the putative helix-turn-helix motif. Experiments with a chemically synthesized DNA-binding domain [Hin-(139-190)] confirmed that this region is sufficient for binding to the symmetric halves of each *hix* element and that it does so with a binding energy at least half that of dimeric wild-type protein (Bruist *et al.,* 1987; Sluka, 1988; J. Sluka, A.C. Glasgow, M.I. Simon, and P.B. Dervan, personal communication). Because proteins containing 55 and 60 amino acids of Hin bind with an affinity comparable to that of the 52-residue protein, it is reasonable that Hin-(139-190) contains all that is necessary for maximal binding. On the other hand, experiments with the 31-residue polypeptide of Hin-(160-190), thought to correspond to the entire helix-turn-helix domain, demonstrated that it is not sufficient for binding, suggesting that other elements in the DNA-binding domain play an additional role in Hin binding and selectivity (Bruist *et al.,* 1987). Fis and Hu are not necessary for Hin Binding (Glasgow *et al.,* 1989a).

DNA footprinting has delineated the region of DNA involved in recognition (Bruist *et al.,* 1987), while genetic studies have been used to define the sequence requirements for Hin binding (Hughes *et al.,* 1988). DNA methylation interference and protection patterns have been used to elucidate specific contacts between Hin and its binding site (Glasgow *et al.,* 1989a; Sluka, 1988). These data plus comparison of the four naturally occurring *hix* half-sites provide the consensus sequence of the DNA sequence recognized by Hin (Hughes *et al.,* 1988):

$$5'\text{-}(A/T)^1C^2C^3A^4A^5A^6A^7N^8C^9T^{10}(A/C)^{11}T^{12}T^{13}N^{14}N^{15}\text{-}3'$$

For our calculations, we have used the sequence:

$$5'\text{-}A^4A^5A^6A^7C^8C^9T^{10}C^{11}T^{12}T^{13}G^{14}G^{15}\text{-}3'$$

This element corresponds to that portion of the consensus sequence that has been shown to be contacted by Hin as determined by DNase I and methylation protection assays (Bruist *et al.*, 1987; Glasgow *et al.*, 1989).

Affinity-cleavage studies utilizing proteins equipped with nonspecific cleaving moieties [Fe(II)-EDTA] have defined the location of the amino and carboxyl termini of the Hin DNA-binding domain. Sluka *et al.* (Sluka *et al.*, 1989) have put forward a model based on a helix-turn-helix motif where the amino terminus of the protein is located in the minor groove near the symmetry axis of the *hix* site. The residues Gly^{139}, Arg^{140}, Pro^{141}, and Arg^{142}, located in the minor groove, participate in sequence-specific recognition. Additional sequence-specific interactions are provided by the putative recognition of the site (D. Mack and P. Dervan, personal communication). These data serve to define the orientation of the interaction, but do not indicate the detailed atomic interactions responsible for recognition.

CALCULATIONS

The theoretical studies involved torsion-space and traditional molecular mechanics simulations aided by constraints imposed to bias the conformations to fit experimental data and insights. The initial model for the binding domain of Hin was

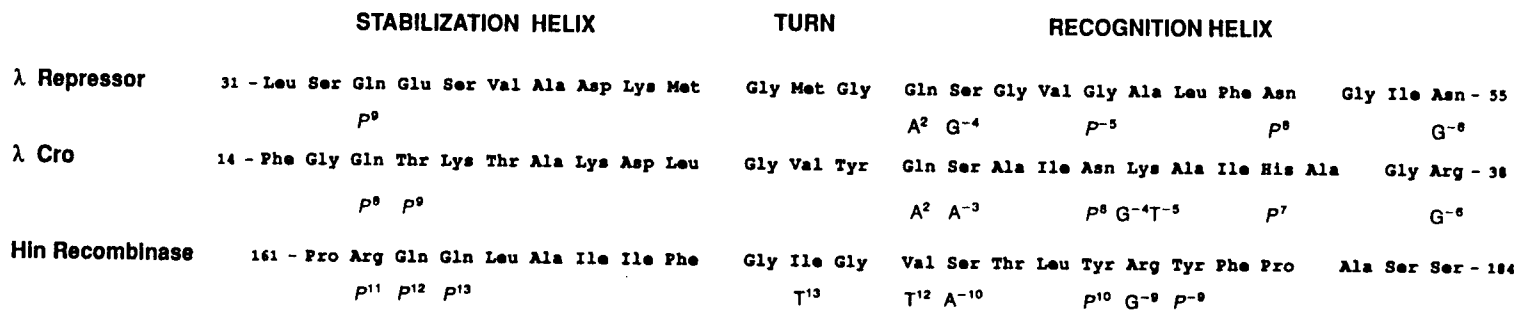


Figure 5: DNA-binding domains of λ repressor (phage λ), λ Cro (phage λ), and Hin recombinase (from *S. typhimurium*). The important interactions with DNA are indicated below each amino acid sequence. The nucleotides and phosphates contacted are numbered from the center of symmetry of the appropriate sites. Data are from protein/DNA cocrystal structure (λ repressor; Jordon and Pabo, 1988) model building from known protein structure (λ Cro; Ohlendorf *et al.*, 1982), and this work (Hin).

constructed by aligning the sequence of the carboxyl-terminal 52 residues of Hin with the sequence of Cro. With this alignment (partially illustrated in Fig. 5), the helix-turn-helix domain of Hin (residues 146-190) was built onto the C-alpha coordinates of Cro (Ohlendorf *et al.,* 1983) by using the C-α-constrained torque mechanics approach to structural prediction. The structure was created one residue at a time starting at residue number 146 and proceeding through residue 190. As each residue was added to the growing chain, the structure was optimized by minimizing normal valence and nonbond potential energy terms, in conjunction with harmonic potentials constraining the Hin C-α atoms to the positions in Cro. The structure thus created was then optimized in the absence of constraints and allowed to equilibrate the molecular dynamics.

Comparison of the sequence selectivity of Hin and Cro provided clues for the initial docking of Hin to DNA. The binding elements of both share a sequence, CTNT, for which in Cro the suggested structure (Ohlendorf *et al.,* 1982) involves Lys-G and Ser-A hydrogen bonding to the first two base pairs. Since the corresponding residues of Hin are Arg^{178} and Ser^{174}, initial docking was performed by allowing Ser^{174} to produce a bridging hydrogen bond with A^{-10} (note: nucleotides in the strand opposite to that shown in the consensus sequence will be designated with negative superscripts corresponding to the numbers of the nucleotides to which they are base-paired), in an orientation analogous to that suggested for Cro (Ohlendorf *et al.,* 1982). This orientation is consistent with the carboxyl-terminal affinity-cleavage experiments of Mack and Dervan (personal communication), which localized the carboxyl end of the domain to a region proximal to the dyad center.

Once the helix-turn-helix element of Hin was docked in this orientation, the structure was optimized in three steps:(*i*) energy minimization of the protein-DNA complex with the use of defined docking constraints (artificial bonds involving Ser^{174}·A^{-10} and Arg^{178}·G^{-9}) to reduce steric interactions, (*ii*) equilibration of the minimized complex by molecular dynamics, and (*iii*) unconstrained dynamics and minimization to produce a low-energy conformation. During the first two levels of calculation, the *hix* DNA was held fixed in a standard B-form DNA structure.

By covalently attaching the cleavage reagent Fe(II)-EDTA to the peptide, Sluka *et al.* (Sluka *et al.,* 1987) showed that the amino-terminal end of the binding domain (residues 139-146) is located in the minor groove, a conclusion confirmed by methylation studies (Sluka, 1988; J. Sluka, A. C. Glasgow, M. I. Simon, and P. B. Dervan, personal communication). Thus, initially we treated the amino-terminal region (residues 139-146) of the DNA-binding domain as an independent motif, generated the structure separately in an extended conformation, and optimized it independent of the remainder of the polypeptide. The amino terminus was then docked to the minor groove by following the model of Sluka *et al.* (Sluka, 1988; J. Sluka, A.C. Glasgow, M.I. Simon, and P.B. Dervan, personal communication) [as derived from chemical modification data (Sluka *et al.,* 1987)]. Constraints were provided between Arg^{140} and Arg^{142} and atoms in the minor groove (at base pairs 4, 5, and 6) and the structure was optimized as described above. The low-energy minor-groove (residues 139-146) and major-groove (residues 146-190) portions were then linked with an artificial distance constraint and the full 52-residue polypeptide was optimized.

Although there are approximations and restrictions inherent in such calcula-

tions, we believe that they account for proper steric and hydrogen bonding interactions and lead to a number of new structural details that help formulate experiments to test the structural elements.

RESULTS

The predicited structure of the Hin biding domain is sketched in Fig. 2, while Fig. 7 contains a stereo image including some specific contacts. The recognition helix (helix 3, residues 173-180) lies within the major groove and involves four major site specific sets of interactions, stabilized by two tyrosine-phosphate hydrogen bonds. Optimal tyrosine-phosphate hydrogen bonding seems to produce a somewhat wider major groove than the classic B-form DNA [this is consistent with the observation that methylation of A^{-10} at the minor groove atom N3 is reduced by Hin-*hix* complex formations (Glasgow *et al.*, 1980), presumably because of this modification of DNA conformation].

The recognition helix is held in place by the stabilization helix (residues 162-169) that has hydrophobic interactions with the recognition helix plus hydrogen bonds between side chains of Arg^{162}-Gln^{163}-Gln^{164} and $P^{13} - P^{11}$ of the phosphate backbone. Additional phosphate contacts are made by Lys^{146} (with P^{-9}) and Lys^{186} (with P^{-5}), further increasing the non-sequence-specific energy of interaction. Fig. 2 shows the orientation of Hin in the complex together with phosphate contacts made in the complex.

The specific interactions of Arg^{140} and Arg^{142} with the minor groove of $A^5A^6A^7$ are supported by the region 143-161, which includes a third helix lack-

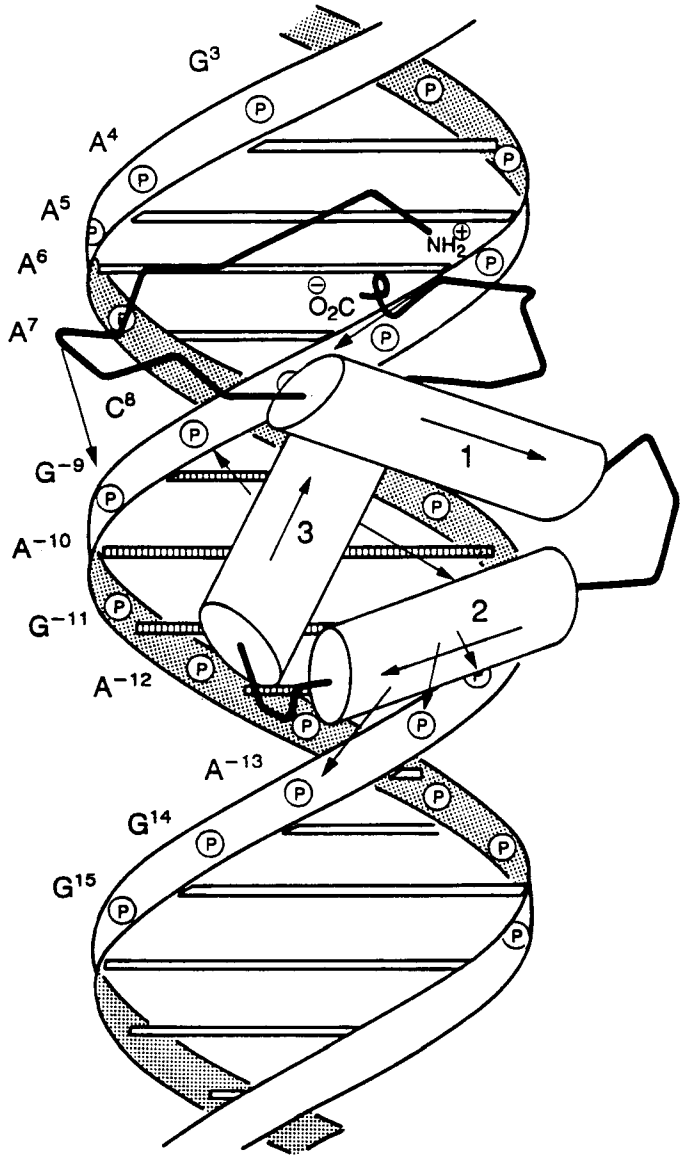


Figure 6: A model of the proposed structure of the DNA-binding domain of Hin recombinase with indications of the backbone structure and phosphate contacts (hydrogen bonding to phosphates is illustrated by arrows to the phosphates contacted). Hin helicies 1 (no sequence-specific interactions), 2 (stabilization helix), and 3 (recognition helix) are shown with arrows indicating the overall direction of the polypeptide backbone.

ing sequence-specific interactions with the DNA (Lys^{146} has a hydrogen bond to P^{-9}). A number of hydrophobic interactions between groups on the three helicies provides additional hydrophobic stabilization of the overall structure.

The protein-DNA contacts summarized in Fig. 7 are sufficient to explain the known sequence selectivit and the observed methylation interference and protection patterns characteristic of the Hin protein (Bruist *et al.,* 1987; Glasgow *et al.,* 1989) with only one exception noted below. Key points are as follows.

(*i*) The strong selectivity for $T^{12}T^{13}$ is generated by complementarity between the hydrophobic surface created by the side chains of residues in the turn region of the peptide, Ile-Gly-Val (residues 171-173), and the C5 methyl groups of the thymines.

(*ii*) The model structure shows no significant interaction between position 11 and the Hin protein (this agrees with the lack of sequence conservation at position 11 in *hix* sites and with the apparent lack of sequence selectivity associated with this site for Hin proteins).

(*iii*) The model has T^{10} strongly preferred due to a bridging set of hydrogen bonds between the complementary base A^{-10} and Ser^{174} of Hin. Such hydrogen bonding has been postulated to produce adenine specificity upon Cro binding, as suggested by Ohlendorf *et al.* (Ohlendorf *et al.,* 1982). Replacement of A^{-10} with either C or G would reduce optimal hydrogen bonding (neither has hydrogen bond donor capability in the major groove), while an $A^{-10} \rightarrow T$ transversion would eliminate the hydrogen bonding potential at this site.

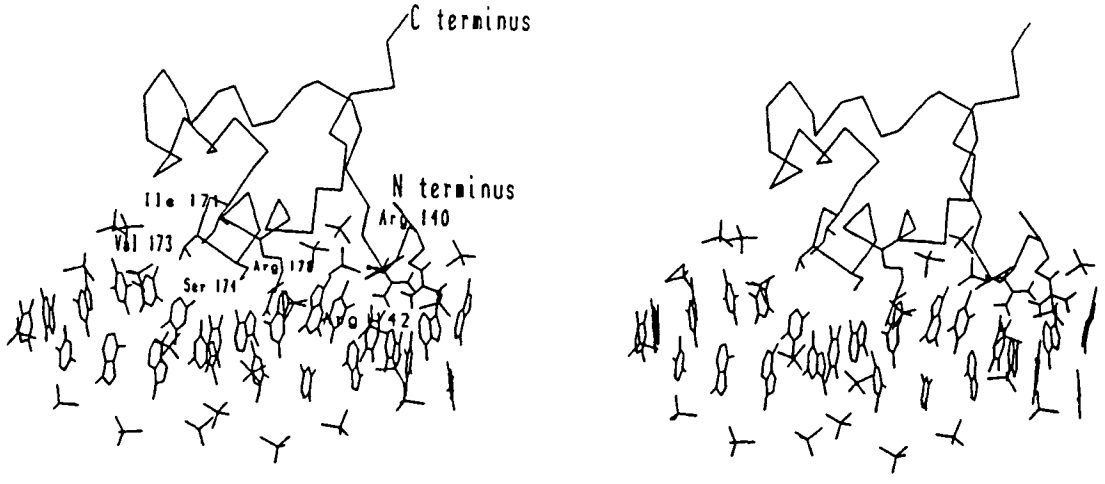


Figure 7: A stereo diagram of proposed sequence-specific contacts in the major and minor grooves. The C^{α} trace for the entire binding domain is shown with the side chains of those residues implicated in sequence-specific contacts. The deoxyribose sugars have been omitted for clarity.

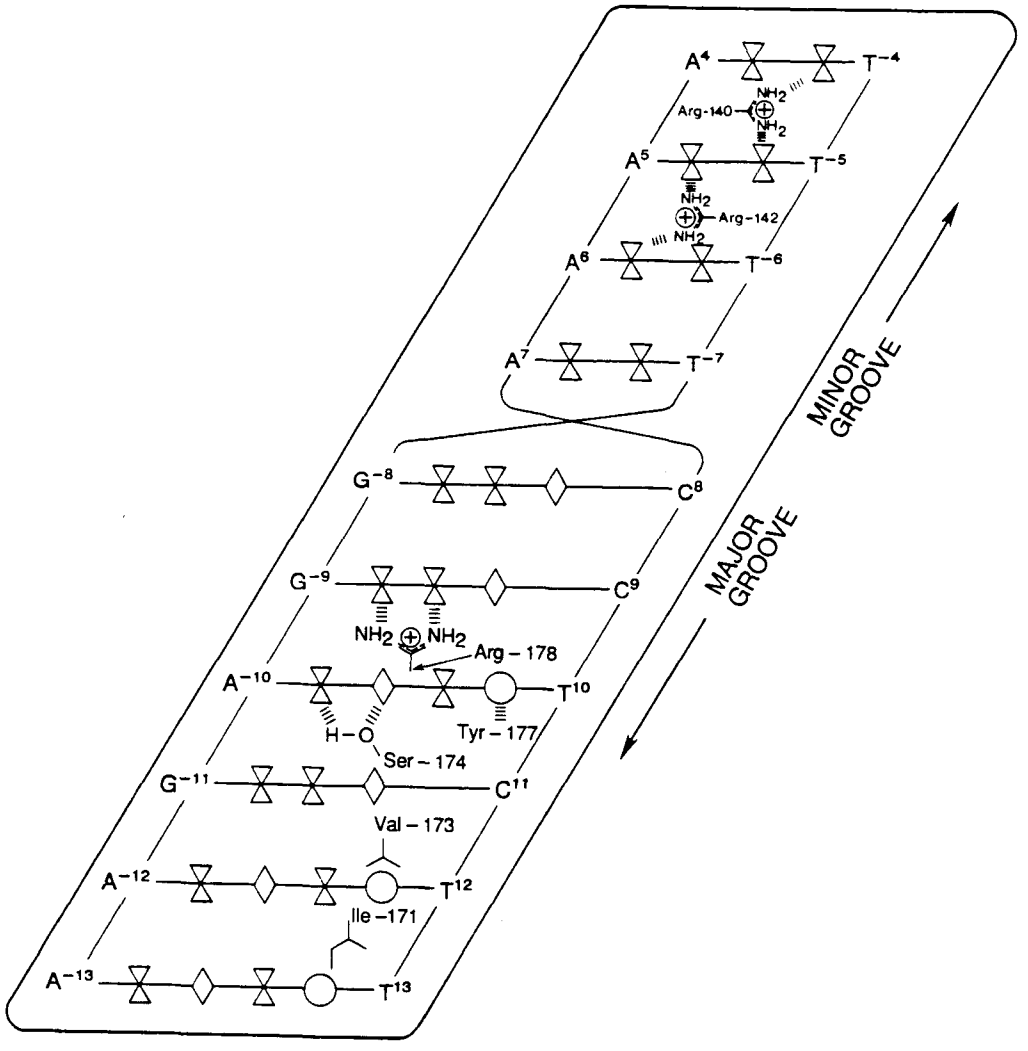


Figure 8: A diagram of the sequence specific contacts of the Hin-*hix* complex. Hydrogen bond donors are designated by diamonds, acceptors by hourglasses, and the 5-methyl group of thymine by a circle.

(*iv*) In the theoretical model, Arg^{178} is responsible for recognition of G^{-9}, donating hydrogen bondsto the N7 nitrogen and the C5 carbonyl groups. Consistent with this, methylation of the N7 nitrogen of G^{-9} is deleterious to Hin binding (Glasgow *et al.,* 1989). Replacement of G^{-9} could be tolerated since the guanidinium group of arginine coud also form a set of bridging hydrogen bonds across the N7 atoms of positions -9 and -8. Thus, alternative sequences providing full hydrogen bonding to Arg^{178} would also inlcude N^8C^9 as well as $(\frac{T}{C})^8T^9$ [corresponding to $(\frac{A}{G})^{-8}A^{-9}$].

(*v*) Position 7, a conserved adenine in the four known *hix* half-sites, does not appear to be contacted by the protein. This lack of selectivity is consistent with the observation that metylation of the N3 of A^7 does not reduce Hin binding, and Hin binding does not prevent methylation at this site (Glasgow *et al.,* 1989). We propose that the observed conservation at this site is coincidental and plays no role in DNA recognition.

(*vi*) Recognition at positions 4, 5, and 6 is provided by sequence-specific contacts between the amino terminus of the binding domain and atoms in the minor groove (Arg^{140} with the O3 atoms of $T^{-5}T^{-4}$ and Arg^{142} with the N3 atoms of A^6T^5).

(*vii*) The final positions, 3-1, are conserved among all known *hix* sites, but DNase I footprinting suggests that these positions are not contacted by the 52-residue binding domain (Glasgow *et al.,* 1989), and they are not contacted in our structure. It is possible that this region plays a role in site-specific recombination by interacting with the catalytic domain of recombinase.

SUGGESTED TESTS OF THE STRUCTURE

The theoretical model of the Hin binding domain suggests mutations (see Fig. 9 for a summary) and base-pair substitutions that should alter binding specificity.

The model predicts that residues 146, 162, 163, 164 and 186 donate hydrogen bonds to the phospahte backbone and that these positions can be satisfied with any residues with similar hydrogen bonding potential (such as arginine, asparagine, glutamine, or lysine). Tyr^{177} and Tyr^{179} are involved in hydrogen bonding to the phosphate backbone but may also play a role in modifying the width of the major groove. If this latter effect is importaint in recognition, their substitution with amino acids having equivalent hydrogen bond donor ability but lacking such a rigid side chain would lead to reduced binding.

Arginine is the only amino acid capable of bridging the minor-groove hydrogen bond acceptors and providing adeinine specificity at positions 4, 5, and 6. Any other substitution should alleviate selectivity at these positions and greatly diminish binding (due to the loss of hydrogen bonding). The calculations suggest that flexibility of the arginine side chain would also allow equivalent hydrogen bonding to A·T transversions. However, the C-2 amino group of guanine, located in the minor groove, would prevent hydrogen bonding to G·C containing sequences.

Substitutions of Ile^{171}, which is responsible for selective hydrophobic interactions with T^{13}, are probably limited to the sterically conservative replacements leucine and valine. Replacement of Ile^{171} with the hydrophilic amino acid glutamine

Hin Sequence									
Ile	Gly	Val	Ser	Thr	Leu	Tyr	Arg	Tyr	Phe
Functionally Acceptable Substitutes									
						Arg		Arg	
				Ala		Asn		Asn	Ile
Val				Cys	Ile	Gln		Gln	Leu
Leu	---	---	---	Ser	Val	Lys	Lys	Lys	Met
Substitutions that Alter Selectivity									
Ala		Ala							
Gln		Thr	Ala				Ala		

Figure 9: Predicted effects of various point mutations in the putative recognition helix. The wild-type Hin sequence is shown in the upper row. Conservative mutations, those that are predicted to affect neither the structure nor the binding characteristics of Hin, are listed in the middle rows. Mutations predicted to change sequence selectivity without structurally disrupting Hin are shown in the lower rows; the symbol Ala actually refers to any small residue that will not disrupt the overall structure of the binding domain. All other symbols represent the standard 3 letter code.

may lead to adenine selectivity at position 13 by providing a hydrogen bond donor and acceptor of the correct geometry.

Val^{173} (responsible for T^{12} selectivity) is a highly constrained position that can only be satisfied by valine. Replacement with a small amino acid such as alanine should remove selectivity at position 12 and reduce overall Hin affinity, and replacement of Val^{173} with threonin may generate A^{12} selectivity by providing a hydrogen bond acceptor and donor of the appropriate geometry.

Ser^{174} is a highly conserved position; no other amino acid is capable of forming such a tightly constrained set of hydrogen bonds. The functionally conservative replacement of a threonine is sterically forbidden due to the hightly constrained geometry of the hydrogen bonding, while replacement of this amino acid with one of smaller size (*e.g.,* alanine) would reduce binding by removing two hydrogen bonds.

The next residue, Thr^{175}, does not appear to make any contact with the DNA and presumably is constrained only by steric considerations. Substitution of small amino acids at this position (serine, cysteine, or alanine) should not reduce binding.

Leu^{176} serves a structural role as part of the hydrophobic core of the domain. Residues that participate in hydrophobic interactions are often highly constrained sterically (Reidhaar-Ohlson and Sauer, 1988), and thus few substitutions at this position would produce stable proteins.

Tyr^{177} and Tyr^{179} play a role in hydrogen bonding to phosphate, as discussed above. For Arg^{178}, the only replacement that would maintain selectivity at position 9 should be lysine, which is also capable of donating two hydrogen bonds.

Phe^{180} also plays a role in maintaining the hydrophobic core of the protien and, as such, is tightly constrained.

To facilitate additional experimental and theoretical investigation of this model, the coordinates of this proposed complex are being submitted to the Brookhaven Protein Database.

SUMMARY

Combinding molecular dynamics simulations with constraints based on current knowledge of protein structure leads to a a theoretical structure of the binding domain of Hin recombinase with the *hix* site of DNA. The model offers a mechanistic explanation of the presently known characteristics of Hin and predicts the effects of specific mutations of both protein and DNA. The predictions can be tested by currently feasible experiments that should lead to refinements in and improvements on the current theoretical model. Because current experimental and theoretical methods are all limited to providing only partial information about protein-DNA interactions, we believe that this approach of basing molecular simulations on experimental knowledge and using the results of these simulations to design new, more precise experimental tests will be of general utility. These results provide additional evidence for the generality of the helix-turn-helix motif in DNA recognition and stabilization of proteins on DNA.

Acknowlegments

We thank Professor Peter Dervan for helpful discussions. K. W. P. gratefully acknowledges support from the National Science Foundation in the form of a graduate fellowship. This project was initiated with support from Office of Naval Research/Defense Advanced Research Planning Agency and continued with support from Department of Energy-Energy Conversion and Utilization Technologies. The computations were carried out using BIOGRAF (from BioDesign) with modified routiens for torque mechanics and for C^{α} constraints. The computers (Alliant FX8/8 and DEX VAX 8650) and graphics systems (Evens and Sutherland PS330 and 390) were funded by Office of Naval Research/Defense Advanced Research Planning Agency. National Science Foundation-Materials Research Groups, Office of Naval Research-Special Research Opportunities, and Department of Energy-Energy Conversion and Utilization Technologies. This paper is Contribution 7771 from the Arthur Amos Noyes Laboratory of Chemical Physics.

References

BIOGRAF/POLYGRAF, (1992), Copyright, Molecular Simulations, Inc.

Bruist, M. F., Horvath, S. J., Hood, L. E., and Steitz, T. A., (1987), *Science*, **235**, 777 - 780

Chothia, C., and Lesk, M. A., (1986), *EMBO J.*, **5**, 823-826

Correa, P. E., (1990), *Proteins*, **7**, 366-377

Dodd, I. B., and Egan, J. B., (1990), *Nucl. Acids Res.*, **18**, 5019-5026

Dykes, D. C., Brandtrauf, P., Luster, S. M., Freidman, F. K., Pincus, M. R., (1993), *J. Biomolecular Structure and Function*, **10**, 905-918

Glasgow, A. C., Bruist., M. F., and Simon, M. I., (1989a), *J. Biol. Chem.*, **264**, 10072-10082

Glasgow, A. C., Huges, K. T., and Simon, M. I.,(1989b), *Mobile DNA*, Berf, D., and Howe, M, eds. (Am. Soc. Microbiol, Washington, DC)

Holm, L., and Sander, C., (1991), *J. Mol. Bio.*, **218**, 183-194

Holm, L., and Sander, C., (1992), *Proteins*, **14**, 213-223

Hughes, K. T., Youderian, P. and Simon, M. I., (1988), *Genes Dev.*, **2**, 937-948

Johnson, R. C., Bruist, M. F., Glaccum, M. B., and Simon, M. I., (1984) *Cold Spring Harbor Symp. Quant. Biol.*,**49**, 751

Johnson, R. C., Bruist, M. F., and Simon, M. C., (1986), *Cell*, **46**, 531-539

Johnson, R. C., Glasglow, A.M., and Simon, M. I., (1987), *Nature*, **329**, 462

Jones, T. A., Zou, J.-Y., Cowan, S. W., and Kjeldgaard, M., (1991), *Acta. Cryst.*, **47**, 110-119

Jordan, S. R., and Pabo, C. O., (1988), *Science*, **242**, 839-899

Karplus, M., and Petsko, G. A., *Nature*, (1990), **347**, 631-634

Koymans, L. M. H., Grootenhuis, P. D., and Haasnoot, C. A. G., (1993), *J. Royal Neth. Chem. Soc.*, **112**, 161-168

Mack, D. P., Sluka, J. P., Shin, J. A., Griffin, J. H., Simon, M. I., and Dervan, P. B., (1990), *Biochemistry*, **29**, 6561-6567

Mathiowetz, A.M., (1993) *Dynamic and Stochastic Protein Simulations: From Peptides to Viruses.*, Thesis, (California Institute of Technology, Pasadena CA)

Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y. and Mathews, B. W., (1982), *Nature*, **298**, 718-723

Ohlendorf, D. H., Anderson, W. F., Takeda, Y., and Mathews, V. W., (1983), *J. Biol. Mol. Struct. Dyn.*, **2**, 553-563

Otwinowski, Z., Shevits, R. W., Zhang, R. C., Lawson, C. L., Joachimaiak, A., Marmorstein, R. Q., Luisi, B. F., and Sigler, P. B., (1988), *Nature*, **355**, 321-326

Plasterk, R. H. A., and van de Putte, P., (1984), *Biochim. Biophys. Acta*, **78**, 111-119

Plaxco, K. W., Mathiowetz, A. M., and Goddard, W. A., III, (1989), *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 9841-9845

Purisima, E. O., and Scheraga, H. A., (1984), *Biopolymers*, **23**, 1207-1224

Reid, L. S., and Thornton, J. M., (1989), *Proteins*, **5**, 170-182

Reidhaar-Ohlson, J. T., and Saur, R. T., (1988), *Science*, **241**, 53-57

Rey, A., and Skolnick, T. B., (1992), *J. Comp. Chem.*, **13**, 443-456

Schmitz, U., Kumar, A. M., and James, T. L., (1992), *J. Am. Chem. Soc.*, **114**, 654-656

Simon, M. I., Zeig, J., Silverman, M., Mandel, G., and Doolittle, R. F., (1980), *Science*, **209**, 1370

Sluka, J. P., Horvath, S. J., Bruist, M. F., Simon, M. I., and Dervan, P. B., (1987), *Biochemistry*, **29**, 6551-6561

Sluka, J. P., (1988) *Design, Synthesis and Characterization of Sequence Specific DNA Cleaving Agents*, Thesis, (California Institute of Technology, Pasadena CA)

Sluka, J. P., Horvath, S. J., Glasgow, A. C., Simon, M. I., and Dervan, P. B., (1990), *Biochemistry*, **29**, 6551-6561

Srinivasan, S., March, C. J., Sudarsanam, S., (1993), *Protein Sci.*, **2**, 277-289

Zeig, J., Silverman, M., Hilmen, M., and Simon, M. I., (1977), *Science*, **196**, 170

Chapter 3

Experimental Investigations of the Hin-*hix* Complex

Abstract

The ultimate test of any structural or mechanistic model is the ability of the predictions that are generated by that model to withstand the scrutiny of experimental investigation. The atomic resolution model of the DNA binding domain of Hin recombinase reported previously was used to generate a diverse set of predictions regarding the biochemical and genetic characteristics of the complex. The experiments that have been conducted since the development of the model, including the high resolution genetic determination of the elements in Hin that are responsible for its sequence specificity, the qualitative characterization of the selectivity of the protein, and the identification of functional groups on the *hix* element that play roles in complex formation all indicate a close correlation between theory and experiment. Some inconsistencies have also been observed, which have led to further refinement of some aspects of the model structure. These results are indicative of the general utility of this type of modeling effort.

Introduction

The computational complexity of *ab initio* structural determination by molecular dynamics simulations limits the applicability of this technique for the determination of protein structures from primary sequence data. However, as we have illustrated in the previous chapter, experimentally constrained molecular dynamics approaches can be used to model molecular structure with the caveat that the accuracy of the final model may be limited by inaccuracies in the form or application of these constraints. For this reason, it is imperative that careful scrutiny be performed on models produced by such a technique, to determine both consistency with the experimental constraints used to bias the conformational search and to measure the predictive value of the model results.

Initial tests of the model building effort included an analysis of the agreement of the model with our knowledge of the rules of structure of proteins. The dihedral angles, hydrophobic packing, and surface characteristics of the model structure are consistent with the well characterized principles of protein structure, and were the first indications that the initial assumptions upon which the model building effort was made were reasonable.

Beyond the simple tests of consistency with the principles of protein structure, a wide variety of experimental investigations have been conducted on the genetics and biochemistry of the Hin-*hix* complex. Genetic investigations have included the generation of mutations of the putative binding domain of Hin that disrupt complex formation, the determination of the sequence selectivity of the protein

and the generation of Hin mutants that exhibit an increased affinity for DNA. Biochemical investigations have included the modifications of functional groups in the major and minor groove of the *hix* element that reduce binding and additional characterizations of the orientation of Hin as determined by the addition of chemical cleaving reagents. In addition, the structures of two putative close homologues of the DNA binding domain of Hin have been solved and agree closely with many aspects of the proposed Hin structure.

The predominant body of experimental work is consistent with our model of the structure of the Hin-*hix* complex. The few discrepancies between experiment and the model indicate certain aspects of the model that are in need of further refinement, as discussed below.

Structural Tests

Now that on the order of 400 protein structures have been solved by crystallographic and NMR based techniques, a body of empirical rules have been established regarding protein structures. These rules can be used to judge the feasibility, or at least the lack of feasibility, of a structural model. They include the role of hydrophobic packing in the interior of the protein, the makeup of the solvent accessible residues in the putative structure, and the nature of the dihedral angles observed in the backbone of the polypeptide. For model structures based on homology, various conserved interactions can also be harnessed to test the validity of the underlying assumptions of this type of model building effort.

A primary screen can be based on simple analysis of the environment of each residue in the model structure. Such questions as the occurrence of charged or other hydrophillic groups buried in the hydrophobic core of the protein, the existence of hydrophobic surfaces that do not correspond to functional aspects of the protein, and the observation of significant non-sequence specific DNA contacts are simple tests of model viability. Figure 1 illustrates the hydrophobic packing of the model structure. Note the dense packed hydrophobic nature of the interior. This is a sensitive test of our assumption that Hin is homologous to the helix-turn-helix family of proteins and that the phasing of the sequence lineup was correctly assigned.

Two hydrophobic surfaces are evident in the uncomplexed protein. One, which is evident in the turn region of the helix-turn-helix motif, appears to play an important role in site recognition and is highly conserved in the Hin family of recombinases. A second hydrophobic surface region is seen on the face of the protein opposite the DNA recognition surface. This region presumably corresponds to the surface of the binding domain that is packed against the enzymatic domain in the native protein.

The dihedral angles present in the model structure were analyzed using an analysis algorithm we have written that compares dihedral angles with a compilation of angle occurrences taken from the Brookhaven structural data base (Mathiowetz, 1993). No uncommon ϕ or ψ angles were observed.

The Hin model exhibited several interactions that, while not biased into the model, do reflect conservation between Hin and other members of the helix-turn-helix family of proteins. In particular, several hydrogen bond donating amino acids

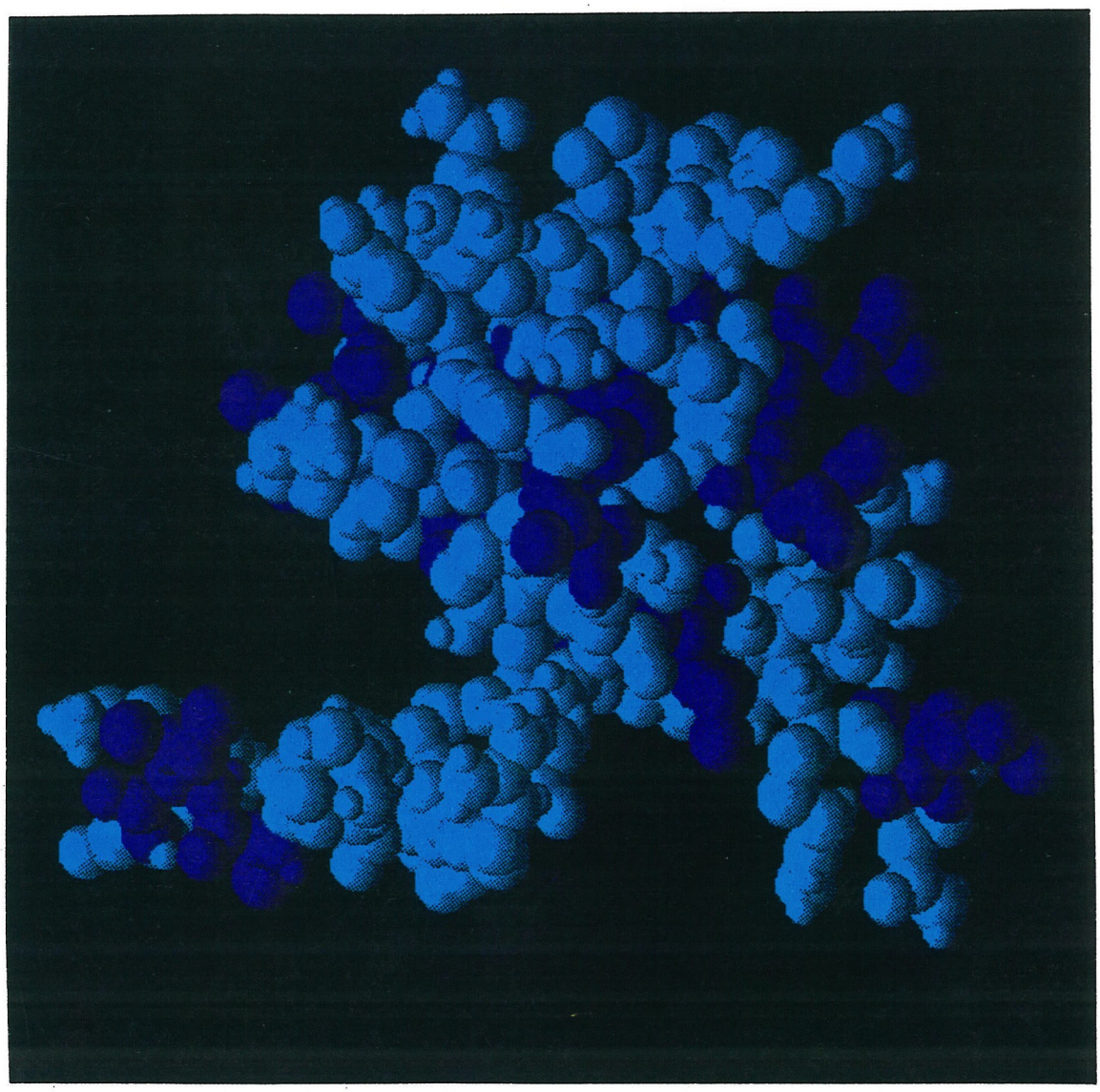


Figure 1: Hydrophobic packing in the Hin model structure. Note that hydrophobic groups in the structure (purple) are predominantly localized to packed interior positions, the turn region between the recognition and stabilization helicies and the face of the protein opposite the DNA binding surface.

in the structure emerged as possible contacts with the phosphate backbone of the *hix* site. As shown in chapter 2, at least three of the putative contacts observed are homologous to contacts known to occur in other members of the helix-turn-helix family of DNA binding proteins (Ohlendorf *et al.,* 1982; Jordan and Pabo, 1988), further evidence that the assignment of Hin to this protein family, as well as the overall model building effort, were substantially correct.

Identification of Residues on the DNA Binding Surface

The development of a simple genetic assay for Hin activity provided the possibility of conducting mutational analysis of Hin to characterize the location and nature of the binding surface of the protein. Of particular interest, two mutations in the putative DNA binding face of Hin were isolated, 174-ser to cys and 178-arg to gly, both of which are illustrated in figure 2. Both of these mutations are sterically conservative, so neither should prohibit binding unless they are involved in complex formation or reduce protein stability. As the latter was ruled out using a western blot analysis to analyze stability *in vivo*, the mutants apparently play a role in complex formation.

The determination of the relative binding affinities of these mutants would be very informative, as the model suggests that the 174 ser-cys mutation should only affect selectivity at the -9 (T·A) position of the binding site, leaving selectivity at other sites unchanged. Unfortunately, efforts to localize the interaction of these mutants on the *hix* element by assaying the reduction in sequence selectivity that occurs in the mutant protein failed because the total affinity of these mutants is

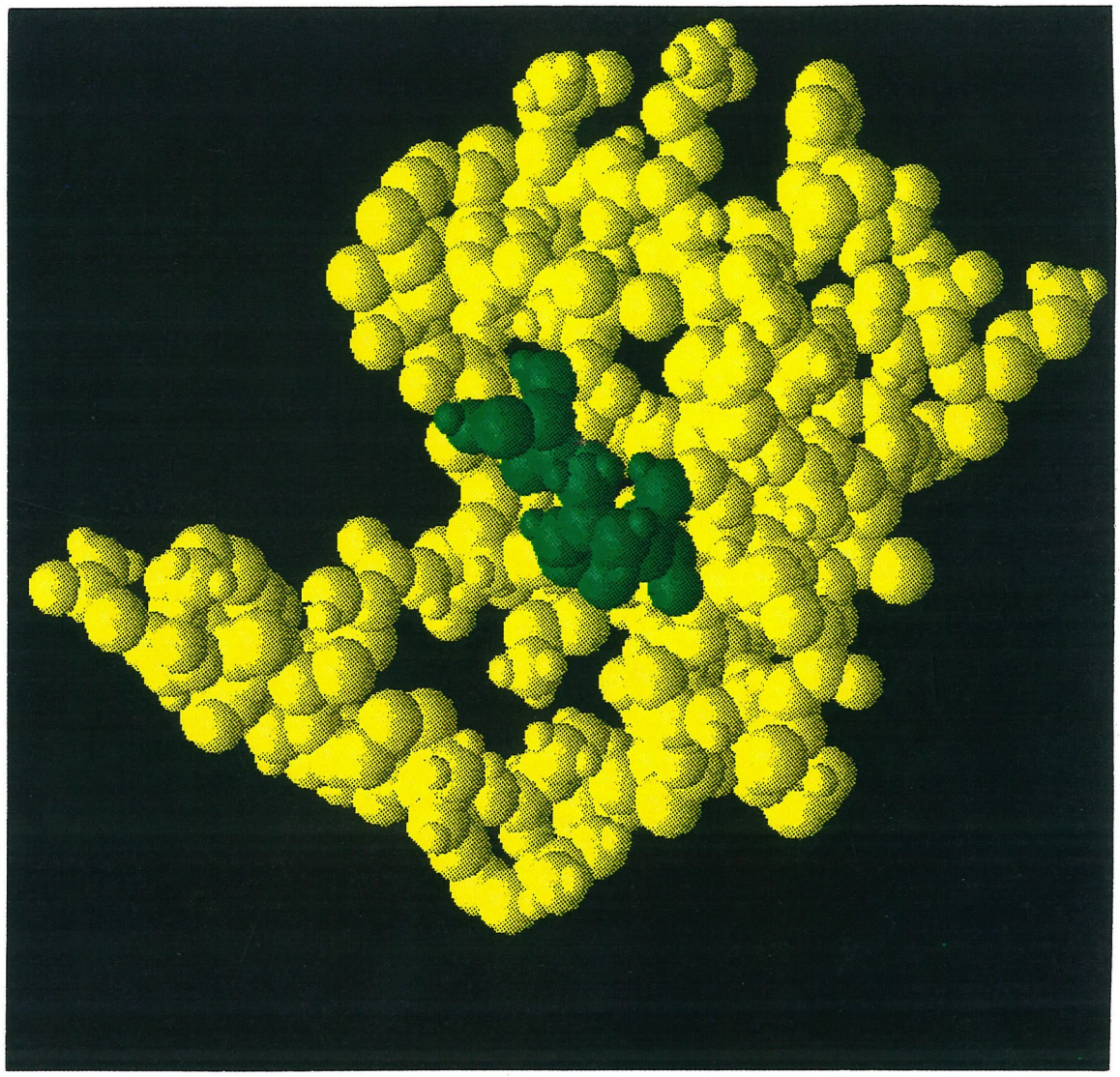


Figure 2: Highlighted are the positions of the two structurally conservative mutations of the putative DNA recognition helix of Hin demonstrated to inhibit DNA binding. They are 174 ser-cys and 178 arg-gly.

less than can be assayed by the phage challenge assay used to determine sequence specificity (see below). For both mutants, the binding signal observed by the phage challenge assay was reduced by 8 orders of magnitude over the wild-type Hin protein. It is interesting to note that, as shown below, mutations at positions 9 and 10 of the *hix* site, thought to interact with 178-arg and 174-ser, also demonstrate an 8 order of magnitude reduction in binding signal (Hughes *et al.*, 1992).

Contributions to the Free Energy of Binding by Thymine Methyl Groups

Our modeling results indicate that the terminal thymine residues in the *hix* sequence play a significant role in determining the sequence selectivity of the Hin protein. The nature of this contact is illustrated in figure 3. Desolvation of this relatively hydrophobic group by complex formation should contribute several kcal/mol to the free energy of binding to A·T-containing sequences. Furthermore, the model indicates that no hydrogen bonds are being formed at these positions in the binding site. If the interactions are the product of hydrophobic interactions rather than specific hydrogen bonding, then replacement of the thymines in these positions with the otherwise equivalent base pair deoxyuracil·adenine, which lacks a 5-methyl group, should reduce the binding affinity of Hin, while replacement with a 5-methyl cytosine·guanine should not.

We have investigated this interaction using gel shift assays (described in the materials and methods section below) with modified *hix* elements containing uracil and 5-methyl cytosine at positions 12 and 13, as shown in figure 4. The thymine methyl group at position 12 was demonstrated to have a significant effect on the

Figure 3: Highlighted are the hydrophobic turn of Hin and the thymine bases at positions 12 and 13 of the *hix* element.

Table 1: The effects on the free energy of complex formation for various base pair substitutions at positions 12 and 13 of the *hix* element. Relative free energies are reported with respect to the wild *Hix L* sequence. Reported values are the average of several determinations, standard deviations were on the order of %5-10 for K_d values.

Species	K_d (% *hix*)	$\Delta\Delta G_{bind}$
$U_{12}U_{13}$	7 %	1.6 kcal/mol
U_{12}	11 %	1.3 kcal/mol
U_{13}	95 %	0.0 kcal/mol
$5mC_{12}5mC_{13}$	73 %	0.2 kcal/mol
$5mC_{12}$	79 %	0.1 kcal/mol
$5mC_{13}$	91 %	0.1 kcal/mol

free energy of complex formation, accounting for a 1.3 kcal/mol contribution to the sequence specific binding at that position. The contribution of the methyl group at position 13 was significantly less, at 0.3 kcal/mol for the dimeric protein complex. Substitution of 5-methyl cytosine·guanosine base pairs at these positions led to a free energy of complex formation that was almost identical to that of the wild-type sequence, indicating that the 5-methyl group is both necessary and sufficient for sequence specificity at these positions in the *hix* element. These results are in

keeping with the predictions made by the model building effort that no sequence specific hydrogen bonding or steric interactions are occurring at these base pairs. The effects of the substitutions investigated at positions 12 and 13 are listed in table 1.

Materials and Methods

Media: Minimal media was the E medium of Vogel and Bonner, supplemented with 0.2% glucose (Vogel and Bonner, 1956). LB medium (Difco Bacto tryptone 10 g/l, Difco yeast extract 5 g/l, and NaCl 5 g/l) was used as rich medium. Antibiotics were included in media as needed to a final concentration of: kanamycin sulfate (kan), 50 μg/ml; spectinomycin (spec), 100 μg/ml; streptomycin (strep) 50 μg/ml.

Cassette Mutagenasis: Single stranded oligonucleotides corresponding to amino acids 169 to 182 were chemically synthesized incorporating a small percentage (1.0%) of each of the other three bases in the synthesis mix for each nucleotide in the codons for these amino acids. The complementary strand was synthesized without incorporating non-wild type nucleotides. Both oligonucleotides were EtOH precipitated/washed, lyophilized, taken up in TEA and quantified by UV spectroscopy. Equal molar amounts of each were taken up in 0.1 M NaCl in TEA, taken to boiling and cooled to room temperature over 1 hour. One μg of the plasmid pKH-66 (Hughes *et al.*, 1988) was cut at the corresponding restriction sites (sac II and sca I), phenol extracted, EtOH precipitated/washed/lyophilized. The hybridized oligonucleotides were added in 1000 fold M excess to 1 μg cut plasmid and ligated in a total volume of 16.5 μl ligase buffer with 600 units of T4 DNA

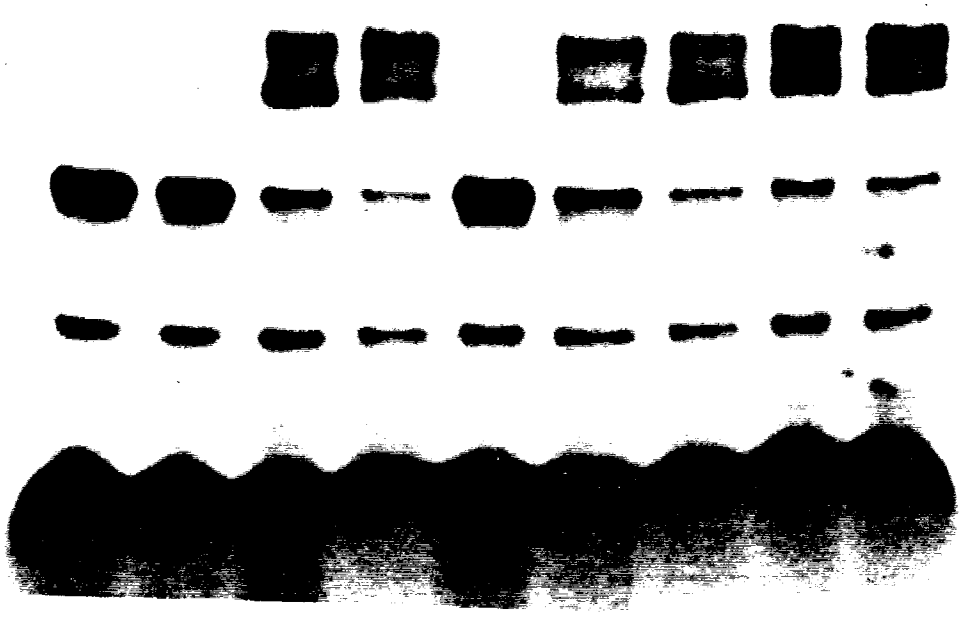


Figure 4: A gel shift assay of the effect of various base pair substitutions at positions number 12 and 13 on Hin-*hix* complex formation. Rows 1-9 correspond to: $U_{12}U_{13}$, U_{12}, U_{13}, $C_{12}C_{13}$, $5mC_{12}$, $5mC_{13}$, $5mC_{12}5mC_{13}$, and *Hix L*. The upper band is the Hin-*hix* complex, the second band is uncomplexed duplex DNA. The lower bands represent hairpin DNA and free radiolabel.

Ligase overnight at room temperature. The resulting ligation mixture was EtOH precipitated/washed/lyophylized and taken up in 100 μl TE/100mM NaCl, placed in a boiling water bath for 5 minutes,cooled to room temperature, EtOH precipitated, washed and taken up in 10 μl TE. 5 μl of this material was transformed into the RZ211-λ fla 406 strain of *E. Coli* (Bruist and Simon, 1984) made competent by the DMSO method (Hanahan, 1983). These cells were plated (100 μl/plate) on strep/spec MaConkey indicator plates (Maniatis *et al.,* 1982) and grown overnight at 37°. Red colonies were scored as Hin^+ mutants, picked, streak purified, inoculated into 2 ml minimal media with strp/spec selection, and grown overnight at 37°. These cultures were then minipreped by the boiling alkali method (Maniatis *et al.,* 1982), and sequenced (Sanger *et al.,* 1977). Two plasmids isolated in this manner, corresponding to the mutations 174 ser-cys and 178 arg-gly were designated pKWP111 and pKWP115 and isolated in larger quantities for further investigations.

Western Blot Analysis: The ProtoBlot Immunoscreening System protocol (Promega) was used to assay mutant and wild-type Hin levels expressed from plasmids pKWP 111, pKWP115 and pKH66 under different induction conditions. Strains of MS1883 with each of the three plasmids were grown overnight at 37° in LB medium containing Strep and Spec to maintain plasmid selection. The culture was diluted 400-fold into LB plus Strep and Spec and the following concentrations of IPTG (final molar concentrations after dilution): 2 x 10^{-5}, 1 x 10^{-4}, and 8 x 10^{-4}. The cultures were grown for 2.5 additional hours at 37° to allow for induction of Hin (final density approximately 100 Klett units). Ten ml of cells from each of the 9 samples were pelleted by centrifugation and resuspended in 0.25 ml of LB. Ten μl 2.5X SDS-PAGE loading buffer (Maniatis *et al.,* 1982) was added to

a 0.015 ml sample of each culture and the resulting mixtures denatured at 100°, electrophoresed in 15% SDS-polyacrylamide (with 4% stacking gel), transferred to nitrocellulose, and hybridized with Hin antiserum prepared previously (Bruist and Simon, 1984). Purified Hin recombinase (Johnson *et al.,* 1986), was also run as a positive control.

Phage Challenge Assay: Overnight cultures of MS1868 carrying the Hin-mutant producing plasmids pKWP111 and pKWP115 were diluted 100-fold into LB plus Strep and Spec to maintain plasmid selection and was grown to a density of 100 Klett units (approximately 6 x 10^8 cells/ml). The cells were diluted fourfold into the same medium plus varying amounts of the inducer IPTG and grown for 1 hr to permit the induction of Hin expression. P22 *hix* or other double symmetric mutant challenge phage was added to a multiplicity of infection of 20, and the infected cells were incubated for 1 hr at room temperature to allow the expression of the Kan^r phenotype. After 1 hr, dilutions were plated onto LB-Kan plates containing the same concentration of IPTG used for induction and incubated overnight at 37°.

Gel Shift Assay: A series of 60 nucleotide self complimentary oligonucleotides were synthesized with the following sequences:

5'-GTCGACCCGGGTTAACCxxATCAAAAACCA-

TGGTTTTTGATxxGGTTAACCCGGGTCGAC-3'

Substitutions were made at positions 12 and 13, indicated by x, as listed in table 1 using the appropriate phosphoramidites from Glenn Ridge Laboratories. The oligonucleotides were gel purified by urea-PAGE, eluted in 0.15 M NaCl in TEA at

37° overnight, separated by filtration, EtOH/NaOAC precipitated, washed twice in 70% EtOH, lyophilized, and taken up in 0.5 ml TEA. The resulting DNA concentration was determined by UV spectroscopy, and an aliquot diluted to 10 fmol/μl. This aliquot was end labeled with P-32 γ ATP and T4 poly nucleotide kinase (Maniatis *et al.,* 1982). Gel retardation assays were performed with these oligonucleotides as described (Glasgow *et al.,* 1989a), using Hin generously provided by A. Glasgow, purified as previously reported (Johnson *et al.,* 1986). The gels were dried, exposed to film and the bands excised and counted by Cerenkoff counting.

Consistency with the Literature

Part of the motivation for studying Hin recombinase is the diverse set of experimental techniques that have been and still are being conducted on the structure of the Hin-*hix* complex. While crystallographic and NMR based investigations into the structure of the complex have not yet come to fruition, a variety of genetic and biochemical investigations have been conducted that largely corroborate the structural model we have presented.

Sequence Specificity

One of the most basic characteristics of a sequence specific DNA binding protein is, of course, that it binds a specific sequence of DNA. From our modeling work, we have derived a set of rules for the relative specificity of the Hin-*hix* complex, as indicated in chapter 2. This model is in reasonable agreement with the crude assay of relative specificities observed for the complex by the phage challenge assay,

as shown in figure 5. The phage challenge assay is qualitative and somewhat limited in applicability by the occurrence of *Salmonella* proteins other than Hin that bind with high affinity to a few of the modified *hix* sites. Still, the general outline of the sequence selectivity of Hin can be observed in that positions 12, 10, 9, 5, and 6 play a significant role in selectivity. The primary discrepancy between this data and the model is that no specificity is observed at position 4.

The phage challenge assay was designed to characterize the binding of sequence specific DNA binding proteins to their target sites *in vivo* (Benson *et al.*, 1986). This assay takes advantage of the regulatory properties of the immunity I region of the *Salmonella*-specific bacteriophage P22. P22 carries a lysogenic repressor, C2 (equivalent to the CI repressor of λ), which represses other phage genes during lysogeny. Unlike bacteriophage λ, P22 has a second immunity region, the immunity I region, which is responsible for the regulation of expression of the *ant* gene product. Anti-repressor (Susskind and Youderian, 1983) inhibits the C2 lysogenic repressor of P22 thereby inducing lytic phage growth. During lysogeny, the Mnt repressor binds at an operator site for the *ant* gene. In the challenge phage, the *mnt* is replaced by a kanamycin resistance gene and the normal operator sequence for P_{ant} is replaced by a DNA site whose *in vivo* binding properties are to be assayed, as shown in figure 6. When the DNA binding protein binds to its specific DNA sequence placed at the P_{ant} operator site, it represses transcription of the *ant* gene from P_{ant} permitting lysogeny. Lysogens are selected for with kanamycin and scored. The higher affinity a protein has to the artificial P_{ant} operator site, the greater the frequency of kanamycin resistance due to lysogenization of the phage (Hughes *et al.*, 1992).

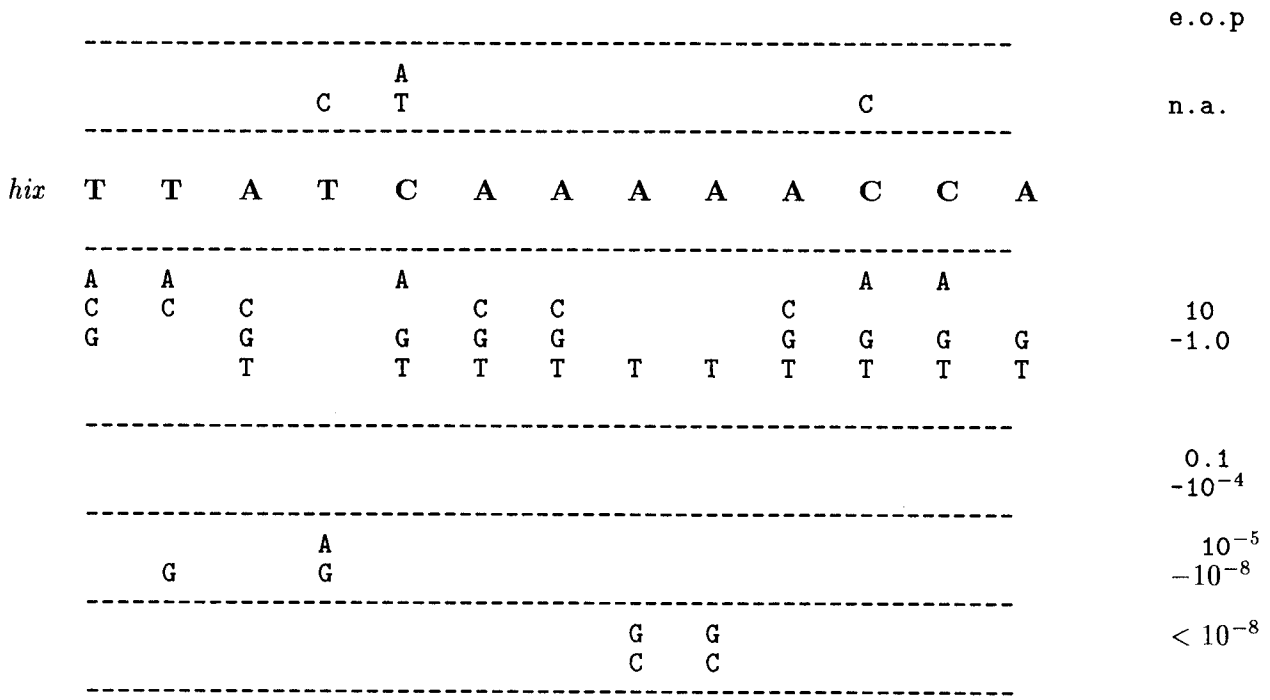


Figure 5: The experimentally observed sequence specificity of Hin Recombinase measured as efficiency of plating (e.o.p.). The wild-type sequence is shown and individual point mutations listed below, in order of their effect on Hin binding (Hughes *et al.*, 1992). Mutations listed above were not assessible due to the masking effects of naturally occurring *E. coli* proteins that fortuitously bind to those sequences (our results).

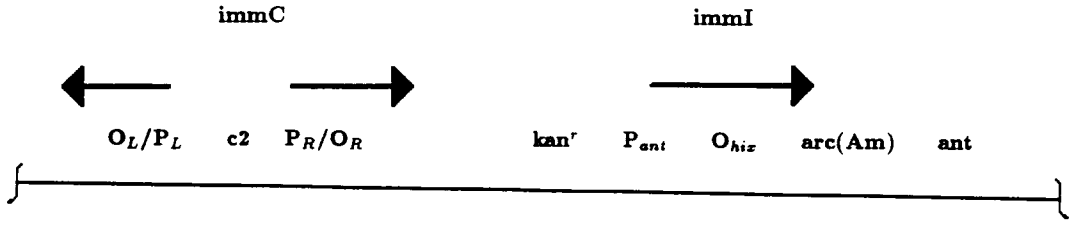


Figure 6: The P22 challenge phage assay with a *hix* challenge phage. A *hix* site is placed at the normal operator site for the *ant* gene promoter in phage P22. When this *hix* challenge phage is used to infect *Salmonella* cells producing Hin, Hin binds at the *hix* site in the phage and represses transcription from the *ant* gene promoter. Repression of the *ant* promoter is necessary for the establishment of lysogeny and the occurrence of kanamycin resistance (Hughes *et al.*, 1988).

Increased Affinity Mutants

Effort has gone into selecting mutations in Hin that demonstrate altered sequence specificity. To date, only one mutant has been isolated that exhibits enhanced binding characteristics (Kelley Hughes, pers. com.). The mutation in question, 170 gly-ser, exhibits an increased binding affinity for all sequences of DNA,

but does not exhibit altered sequence specificity. It was isolated by a selection designed to select for altered specificity because the increased affinity for DNA that it exhibits is sufficiently large that it can bind the altered sequence, although more poorly than it binds to the wild-type *hix* element. A mechanism by which this interaction might occur is readily apparent from our model structure: the new serine hydroxyl group in this mutant is in an excellent position to form a non-sequence dependent hydrogen bond to the phosphate backbone of base -13, thus increasing the free energy of interaction in a non-sequence dependent manner. The geometry of this proposed interaction is shown in figure 7.

Chemical Modifications of Hin and *hix*

The addition of chemical cleaving reagents to the carboxy terminus of the DNA binding domain of Hin has confirmed the general orientation of the recognition helix in the Hin structural model (Mack *et al.*, 1990). The addition of an Fe(II)-EDTA moiety to the amino acid 183 (mutated from the wild-type serine to lysine) of the domain caused cleavage consistent with our localization of this part of the domain slightly above the major groove in a position towards the center of the *hix* site, near base pair 6.

The effect of the deletion of specific functional groups from the minor groove contact portion of the *hix* element (positions 5 and 6) has recently been investigated. Substitution of I·C (equivalent to A·T base pairs in the minor groove) has no effect on complex formation. Substitution at these sites by the modified base pair 5-deazadenine·T indicates that minor groove hydrogen bond contacts occur only to hydrogen bond acceptors on the strand designated as the minus strand (see chapter

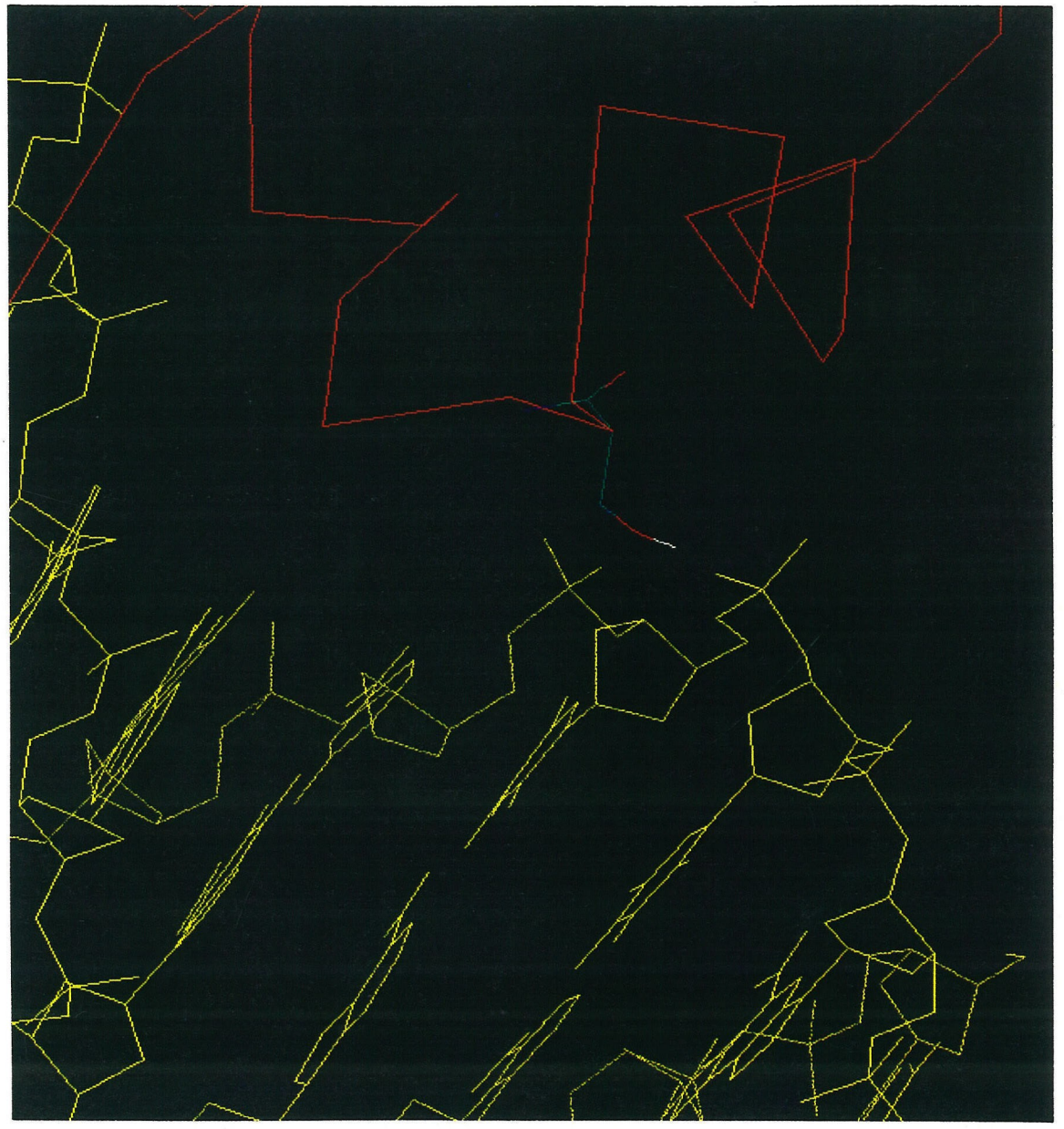


Figure 7: The proposed interaction geometry for the 170 gly-ser mutation in Hin that increases the non-selective binding affinity of the protein. Shown are the *hix* element (yellow), the Hin backbone trace (red), and the mutant side chain.

2) (Hughes *et al.*, 1992).

Comparison to the Homeodomain Structure

The homeodomain is a conserved structural element that exists in a family of eukaryotic regulatory proteins and has recently been demonstrated to represent a DNA binding element (Gehring *et al.*, 1990) and suggested to be homologous to the binding domain of Hin recombinase (reviewed in Affolter *et al.*, 1990). Two examples of this domain in complex with their cognate sites have been structurally determined: the homeodomain of *Antennapedia* (*Antp*) and that of *engrailed* (*en*) (Otting *et al.*, 1990; Kissinger *et al.*, 1990). These proteins share significant homology to the helix-turn-helix family of proteins in general (Qian *et al.*, 1989) and to the Hin family of recombinases in particular (Steitz, 1990). The putative homology to the Hin family extends to the minor groove binding elments of the protein. Figure 8 illistrates a comparison of the experimentally defined structural assignments of the *Antp* and *en* homeodomains and the structural assignments we have postulated for Hin. The best fit sequence alignment introduced one deletion in Hin and one in *en*, both of which occur in random coil regions of the model structure. There is a high degree of correspondence between internal residues in the Hin model and those observed to be internal in the two solved structures. Several putative phosphate contacts in Hin are observed to occur in the homeodomains and are also indicated in figure 8. In general, an impressive degree of correspondence has been observed between the solved homeodomain structures of *Antp* and *en* and our proposed Hin binding domain structure.

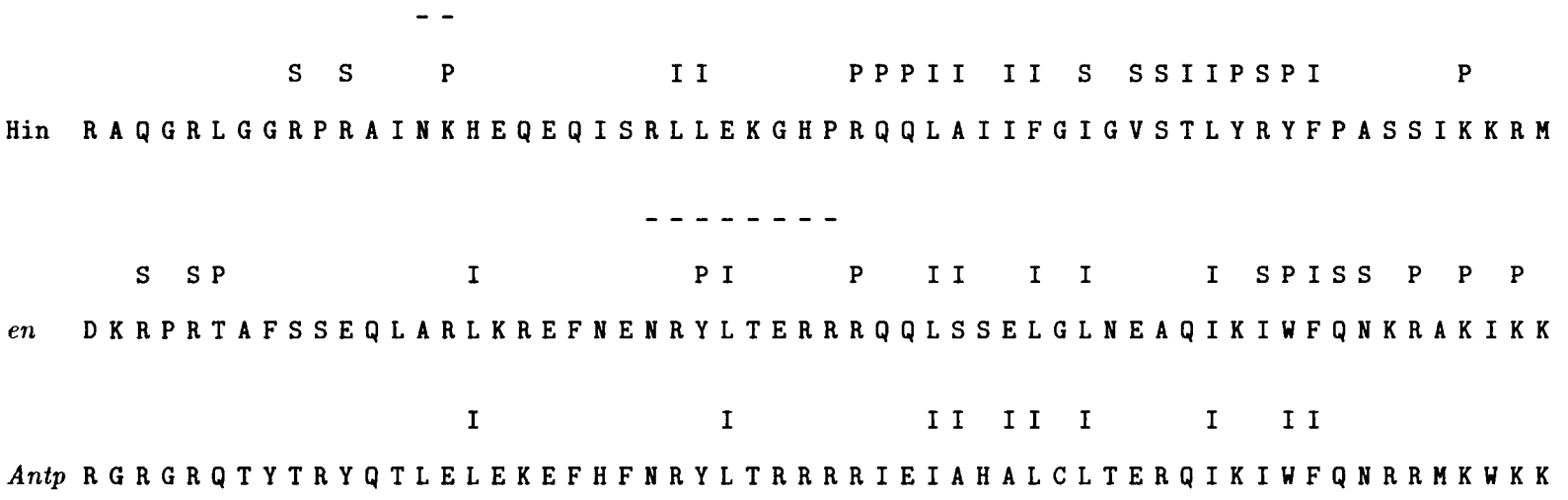


Figure 8: A comparison of the Hin DNA binding domain the *en*, and the *Antp* homeodomains. Hin helicies 1-3 (from chapter 2); *en* homeodomain: α1, α2, and α3 helical segments (from Kissinger *et al.*, 1990); *Antp* homeodomain: I to IV helical regions (from Otting *et al.*, 1990). I demarks residues postulated to be buried in Hin recombinase and known to be core elements in the homeodomains. P represents residues postulated to make non-sequence specific hydrogen bonding contacts with the DNA in Hin recombinase and those residues demonstrated to make such contacts in the homeodomains. S denotes postulated and observed sequence specific DNA contacts. The sequence specific and non-specific contacts of *Antp* have not been well characterized. For optimal alignment two insertions were postulated in *en* (8 residues) and one in Hin (2 residues), which are noted by dashes (from Affolter *et al.*, 1991).

Structural Refinements

The primary discrepancy between theory and experiment has been the experimental determination that sequence specific contacts in the minor groove are limited to a single set of bidentate hydrogen bonds to hydrogen bond acceptors on the (-) strand of *hix* element at positions -5 and -6. This interaction is seen to occur in the homeodomain structures discussed above. We suspect that the discrepancy between theory and experiment occurred because the simulations were conducted on the free binding domain, which, in terminating one residue beyond the minor groove binding element, leads to a highly mobile terminal arginine residue that can make promiscuous hydrogen bonding contacts. We predict that had the chemical modification work and NMR and crystallography on the homeodomain been conducted on the isolated binding domain, then similar erroneous hydrogen bonding contacts would have been observed. Simulations conducted that lack this residue, or that are constrained to prevent it from forming contacts inside the minor groove, should further refine the structure of this element of the DNA binding domain of Hin.

Conclusions

While no one has succeeded in accurately predicting the three-dimensional structure of a protein from its amino-acid sequence from first principles, it has proven possible to use experimentally constrained approaches to generate atomic resolution models of protein structure. We have reported such a model structure

for the DNA binding domain of Hin recombinase in complex with the *hix* element of DNA. The model explains the observed sequence specificity of Hin recombinase and leads to a number of predictions concerning the structure, genetics and biochemistry of the Hin-*hix* complex. With few exceptions, these predictions have withstood the scrutiny of experimental investigation and have demonstrated the general validity of the model structure we have produced. The minor discrepancies noted between experiment and theory can be accommodated and rather than discrediting the model have lead to refinements of it. We feel that our results are indicative of the general utility of this type of modeling effort, and hope that it continues to gain in use.

Acknowledgments

The author wishes to thank Dr. K. Hughes for help on the experimental portions of this work and for communicating prepublication results regarding higher affinity mutants. Thanks are also in order to Dr. A. Glasgow for her help with the gel shift assays and for generously providing the purified Hin recombinase used in these assays, and Dr. M. Simon under whose guidance this work was conducted.

References

Affolter, M., Percival-Smith, A., Muller, M., Billeter, M., Qian, Y. Q., Otting, G., Wuthrich, K, and Gehring, W. J., (1990), *Cell*, **64**, 879-880

Benson, N., Sugiono, P., Bass, S., Mendelman, L. V., and Youderian, P., (1986), *Genetics*, **118**, 21-29

Bruist, M. F., and Simon, M. I., (1984), *J. Bacteriol.*, **159**, 71-79

Gehring, W. J., Muller, M., Affolter, M., Perciveal-Smith, A., Billeter, M., Qian, Y. Q., Otting, G., and Wuthrich, K., (1990), *Trends Genet.*, **6**, 323-329

Hanahan, D., (1983), *J. Mol. Biol.*, **166**, 557-580

Hughes, K. T., Youderian, P., and Simon, M. I., (1988), *Genes and Development*, **2**, 937-948

Hughes, K. T., Gaines, P. C. W., Karlinsey, J. E., Vinayak, R., and Simon, M. I., (1992), *EMBO J.*, **11**, 2695-2705

Glasgow, A. C., Bruist, M. F., and Simon, M. I., (1989a), *J. Biol. Chem.*, **264**, 10072-10082

Glasgow, A. C., Huges, K. T., and Simon, M. I.,(1989b), *Mobile DNA*, Berf, D., and Howe, M, eds. (Am. Soc. Microbiol, Washington, DC)

Johnson, R. C., Bruist, M. F., and Simon, M. I., (1986), *Cell*, **46**, 531-539

Jordan, S. R., and Pabo, C. O., (1988), *Science*, **242**, 839-899

Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornber, T. B., and Pabo, C. O., (1990), *Cell*, **63**, 579-590

Mack, D. P., Sluka, J. P., Shin, J. A., Griffin, J. H., Simon, M. I., and Dervan, P. B., (1990), *Biochemistry*, **29**, 6561-6567

Maniatis, T., Fritch, E. F., and Sambrook, J., (1982), *Molecular Cloning: a Laboratory Manual*, Cold Springs Harbor Laboratories, (Cold Springs Harbor, NY)

Mathiowetz, A. M., (1993), *Dynamic and Stochasitc Protein Simulations: From Peptides to Viruses.*, Thesis, (California Institute of Technology, Pasadena, CA)

Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y. and Mathews, B. W., (1982), *Nature*, **298**, 718-723

Otting, G., Qian, Y.Q., Billeter, M., Muller, M., Gehring, W. J., and Wuthrich, K., (1990), *EMBO J.*, **9**, 3085-3092

Qian, Y. Q., Billeter, M., Otting, G., Muller, M., Gehring, W. J., and Wuthric, K., (1989), *Cell*, **59**, 573-580

Sanger, F., Nicklen, S., and Coulson, A., (1977), *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5463-5476

Steitz, T. A., (1990), *Quart. Rev. Biophys.*, **23**, 205-280

Susskind, M. M., and Botstine, D., (1978), *Microbiol. Rev.*, **42**, 385-413

Susskind, M. M., and Youderian, P., (1983) in *Lambda II*, Hendrix, I. L., and Roberts, J. W., eds. (Cold Springs Harbor Laboratory Press, New York)

Vogal, H. J., and Bonner, D. M., (1956), *J. Biol. Chem.*, **218**, 97-106

Chapter 4

Perturbation Thermodynamics

Abstract

The second law of thermodynamics establishes that all systems are driven to minimumize a characteristic property called free energy. All other thermodynamic properties can be obtained from a knowledge of free energy and its derivatives. Because of the fundamental nature of this property particular focus on it is necessary when comparing theory and experiment. Unfortunately, free energy is not readily obtained in conventional molecular dynamics simulations. The problem originates in the intrinsic difficulty of calculating the entropy of a system, because entropy is an average quantity that reflects the contributions to the total free energy of many individual conformations in the ensemble of conformations available to the system. Because molecular dynamics simulations can be used to generate realistic ensembles of conformations, they can be used to calculate average thermodynamic properties like free energies. While the ensemble of states over which this average must rigorously be taken is infinite, a variety of approaches have developed to generate convergent averages over limited sets of conformations. One such technique, termed pertubation thermodynamics, has been used successfully to calculate the relative free energies of a variety of biologically relevant interactions. In this chapter we will report on the methods and limitations involved in the sucessful application of this technique to several systems that we have studied.

Introduction

The difficulty in calculating free energies originates in the intrinsic difficulty of calculating the entropy of the system. Conventional molecular dynamics simulations sample only a restricted subset of conformation space, the low enerergy region. Because these low energy conformations are the dominant contributors to a systems enthalpy, a relatively good estimate of this property can usually be made. However, unlike enthalpy, estimates of entropy–a measure of the relative disorder of the system–require an often prohibitively extensive knowledge of the conformations available to the system. Fortunately, new methodologies have been introduced to provide a means for the determination of at least relative free energies for molecular systems via computer simulations. Even so, very intensive computational procedures are required in all but the simplest of cases. With the advent of massively parallel computers and improved simulation techniques, these computational barriers are being overcome and fee energy simulations for a variety of important chemical and biological systems have become feasible.

The basic equation of molecular mechanics:

$$E = \sum_{i,j}^{N} \frac{B_{ij}}{r_{ij}^{12}} - \frac{A_{ij}}{r_{ij}^{6}} + \frac{Cq_iq_j}{r_{ij}^{2}} + \sum_{i,j}^{M} K_{bond}(r_{ij} - r_{eq})^2 \ldots$$

calculates the enthalpy of a given conformation relative to some ground state for each angle, bond and nonbond interaction in the system defined by the forcefield. Enthalpy (Δ H) is a useful characteristic, as it can offer a measure of the relative steric, electrostatic, and bonding merits of a set of conformations, but it is not

the characteristic of greatest interest as it is only one of the components of the characteristic that drives a reactions, free energy (ΔF.E.).

$$\Delta\text{F.E.} = \Delta\text{H} - \text{T}\Delta\text{S}$$

The second term in this equation, TΔS, is the entropic contribution to the driving force of reactions. Entropy is a measure of the average degree of disorder for a system and as an "average property" cannot be calculated from a single static conformation. Molecular dynamics, which can be used to generate an ensembles of molecular configurations for a molecular system, can be utilized to calculate this average property and thus can be used to determine free energies as well as enthalpies.

Zwanzig (Zwanzig, 1956) has demonstrated that the relative free energy of two systems can be calculated as the log of an exponential of the enthalpy difference between the two systems averaged over the ensemble of one of them:

$$G(\lambda) - G(\lambda_i) = -RTln < e^{-(V_\lambda - V_{\lambda_i})/RT} >_{\lambda_i} .$$

This equation is rigorously true for any two systems λ and λ_i. (For a proof of this relationship, see Appendix two.)

Using this relationship we should be able to calculate free energies of solvation for such systems as:

$$\text{H}_2\text{O}_{(pure)} + \text{Cl}^-_{(g)} \longrightarrow \text{Cl}^-_{(aq.)}$$

by calculating an ensemble for pure water, and determining the change in energy of each conformation on that ensemble when a chloride ion is placed in it. Unfortunately, while the equation above is true for any two states, it is only rigorously true if the average is take over all of conformation space, which, by definition, is infinite. For the water-chloride case, some water conformations will be in the correct geometry to solvate chloride ions. These conformations are low in energy and thus heavily weighted in the exponential average. Unfortunately, for such a drastic change as placing a chloride ion in pure water, such low energy conformations are rare so the average is slow to converge. Typically, convergence is considered prohibitively slow if the change in free energy is more than 2kT, or about 1.2 kcal/mol at room temperature.

This slow convergence problem has been overcome somewhat by a technique termed *umbrella sampling* (Pangali *et al.,* 1979; Berkowitz *et al.,*, 1984), in which the fact that free energy is a state function (*i.e.*, the determination of the relative free energy between two states will be the same irrespective of the path taken between the states) is used to generate a series of relative free energies along a reaction pathway. For the solvation of chloride example, the chloride ion would be slowly immersed into liquid water. That is, the average is first generated for an ensemble for pure water and the difference in enthalpy determined for those conformations with a chloride ion suspended in the gas phase above it. As a chloride ion suspended far above the liquid interface of water will have little effect on the water, the relative free energy determination for the two systems pure water and water plus a suspended chloride will converge quickly. The next step is to translate the chloride ion a short distance towards the water, and then calculate an ensemble

for this system and compare it with the energy of the next system, *i.e.*, the chloride ion slightly closer to the gas-water interface. The resultant relative free energies (all the way from a distant, gas phase chloride to a solvated chloride) can be summed to determine the free energy of solvation of the chloride ion. This technique has been utilized for a few simple cases (Berkowitz *et al.*, 1984; Chandrasekhar *et al.*, 1985), but unfortunately suffers from the fact that interactions at the interface are usually greater than 2kT unless prohibitively small translations are used. This problem has greatly limited the general utility of this approach.

A second approach to minimizing convergence time is possible because simulations need not be limited to physical transformations. If we limit ourselves to the determination of relative free energies, we can use a thermodynamic cycle to calculate a relative free energy of interest from the relative free energy of two less drastic transformations:

$$\begin{array}{ccc} L(g) & \rightarrow & M(g)\ \Delta\text{F.E.}_3 \\ \downarrow \Delta\text{F.E.}_2 & & \downarrow \Delta\text{F.E.}_1 \\ L(aq) & \rightarrow & M(aq)\ \Delta\text{F.E.}_4 \end{array}$$

The $\Delta\Delta$F.E. of interest, ΔF.E.$_2$ - ΔF.E.$_1$, can be calculated from ΔF.E.$_4$ - ΔF.E.$_3$ much more rapidly than the direct simulation of L(g) $\rightarrow$ L(aq) and M(g) $\rightarrow$ M(aq). This technique was first used by McCammon and coworkers to determine the relative free energy of solvation of chloride and bromide ions (Lybrand *et al.*, 1985a). For this system, convergence was rapid (30 picoseconds simulation time) and accurate (3.35 $\pm$ 0.15 kcal/mol versus the experimentally determined 3.3 kcal/mol, despite the fact that the change in free energy of the system is somewhat larger than the

above mentioned limit of 2kT.

A further improvement in convergence can be obtained if the conversion of L to M is taken in a stepwise fashion with equilibration at each step around a "hybrid" molecule defined by a variable λ, termed a coupling parameter, such that the potential of each intermediate hybrid (V_{λ_i}) is given a combination of the potentials of the starting and ending states:

$$V_{\lambda_i} = \lambda_i V_L + (1 - \lambda_i) V_M$$

An ensemble is generated for each intermediate, a relative free energy difference from the last hybrid determined and the free energy difference of all of the steps summed to calculate a total free energy change for the system. The coupling parameter defines the mixing of all of the parameters that differ between the two systems. This can be conceptualized by considering two noble gas atoms interacting with each other *in vaccuo*. If state A contains two neon atoms, and state B contains one neon atom and one argon atom, λ would linearly change the van der Waals parameters of the potential from describing the interaction between two neon atoms to the interaction of one neon atom with one argon atom over the course of the free energy simulation. Besides perturbations of van der Waals characteristics, other examples can include a mixing of two bond lengths, two torsion angles, a change in relative placement, a modification or one functional group into another, or even the creation of annihilation of a functional group, all of which are illustrated in figure 1. The coupling parameter is usually varied linearly with respect to simulation time, but intermediate hybrids can be generated more or less frequently during a given portion of the simulation if the chemical potential of the sytem is changing

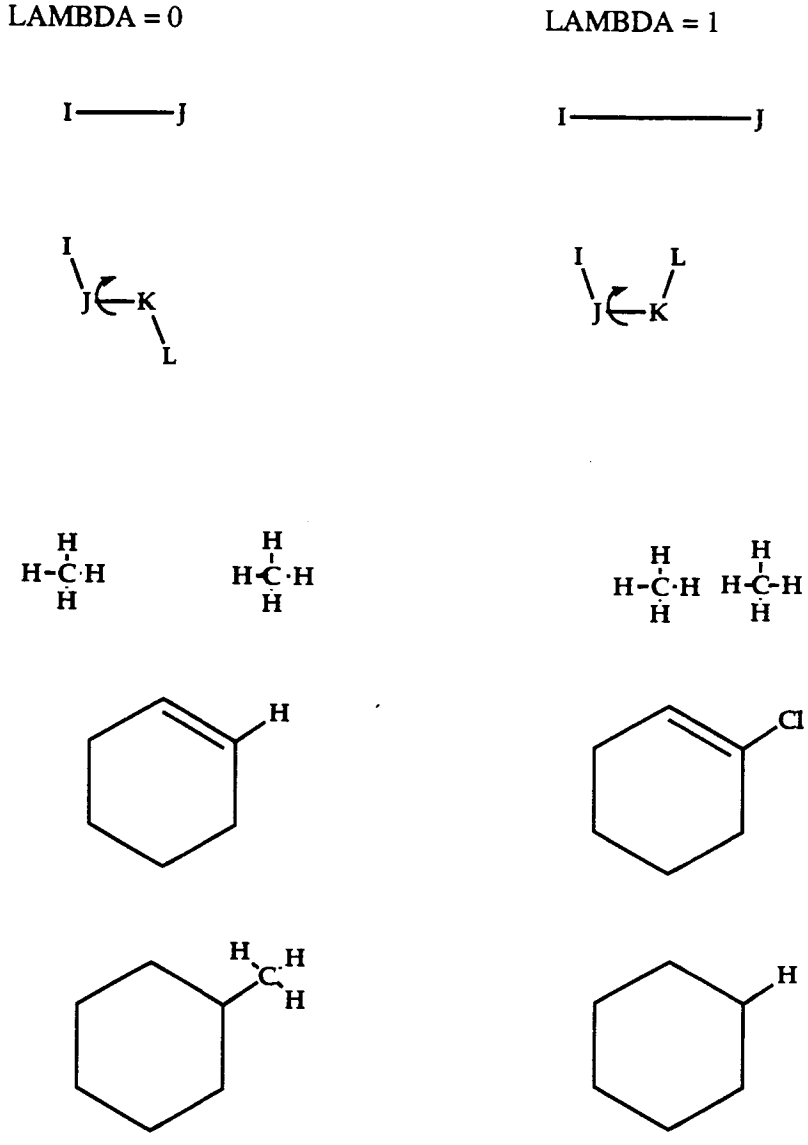


Figure 1: Coupling parameters (λ) can be used to define a perturbation from one species to second that differ in an internal coordinate, a relative position, the characteristics of a functional group, or even in the existence of a functional group.

rapidly (Mezei and Beveridge, 1986). This "perturbation" technique lies at the heart of the general utility of perturbation thermodynamics because it allows the quick convergence of the pertinent average for systems of reasonable complexity. (reviewed in Beveridge and DiCapua, 1989).

Methodology

Generating the Ensemble

The ensemble over which perturbation thermodynamic averages are taken can be generated using either Monte Carlo or molecular dynamics based techniques (reviewed in Beveridge and DiCapua, 1989). Monte Carlo techniques, as the name implies, build an ensemble of states by randomly generating conformations and, in the Metropolis implementation (Metropolis *et al.*, 1953), accepting only those conformations that fit the criteria for inclusion in the ensemble. The criteria is as follows: For a given conformation i, a new conformation, $i+1$, This new conformation will be accepted as part of the ensemble (and as a starting point for the next conformation) if it is lower in energy that i. If conformation $i+1$ is higher in energy than conformation i, then conformation $i+1$ will be included in the ensemble some fraction of the time, with the probability, ρ, of acceptance dependent on the energy of the conformations with the relationship:

$$\rho = e^{\frac{E_i - E_{i+1}}{k_b T}}$$

where T is the bath temperature of the simulation. This procedure can be used to generate a canonical ensemble of states over which thermodynamic averages may be taken.

Molecular dynamics based techniques for deriving ensembles are somewhat more efficient than Monte Carlo techniques for systems that do not exhibit a variety of stable conformations separated by substantial (*i.e.*, greater than a few times kT) barriers (Jacucci and Rahman, 1984). This is due to the fact that, while Monte Carlo methods have a greater ability to sample conformations separated by high energy barriers, this occurs at the cost of discarding a large fraction of all of the conformations generated. Molecular dynamics, on the other hand, is deterministic (*e.g.*, physical properties of the system such as coordinates and forces, determine its time evolution), and therefore has the advantage that every conformation generated is part of the ensemble. Because the systems reported on here do not exhibit multiple stable conformations separated by high energy barriers (see below) the more efficient molecular dynamics method was used.

The effects of various methods of temperature scaling on perturbation thermodynamic ensemble averages are reported in table 1, in which are reported the results of the simulated transformation of chloride to bromide in water. As can be seen, even the naive approach of frequent temperature scaling generates accurate and quickly convergent averages.

The Forcefield Description

The original implimentations of most of the common molecular dynamics

Table 1: The results of perturbation thermodynamic analysis of the transformation of Cl^-(aq) to Br^-(aq) under a variety of simulation techniques. All simulations were conducted for 40 picoseconds with a 1 femtosecond timestep. The simulations included 214 SPC (Berendsen *et al.*, 1981) water molecules in PBC, Cl and Br as described (Lybrand *et al.,* 1985a), and started with a previously equilibrated conformation. The experimentally observed value for this relative free energy is 3.15 kcal/mol (see Lybrand *et al.*, 1985a). The types correspond to canonical (Nosé, 1984), and temperature scaled, which was scaled with the indicated number of femtoseconds between scalings. The error reported is estimated as per the literature (Fleischman and Brooks, 1987). See chapter 1 and appendices III and IV for details on the implimentation of these molecular dynamics routines.

Type	Scale Freq.	R.F.E.
Canon.	na	3.43 ± 0.30
Scaled	1000	3.24 ± 0.22
Scaled	200	3.20 ± 0.25
Scaled	1	3.25 ± 0.19

forcefields were predicated on the united atom approximation (Weiner *et al.,* 1984) which did not explicitly represent hydrogens on carbon atoms, rather these hydrogens were represented as modifications of the carbon atoms charge and van der

Waals parameters. While this approach has been well established for molecular dynamics conformational searches, it has been shown to lead to inaccuracies in the more demanding simulations required for thermodynamic analysis (Sun *et al.*, 1992). A simple solution to this problem is to avoid the united atom approximation and conduct simulations using the full atom representation, as is done in the work reported here.

Because the DREIDING forcefield was optimized for very general use, it does not report charges, as the charge distribution on a given atom type can vary dramatically depending on the molecular structure in which it resides. For the simulations reported here that use the DREIDING forcefield, atomic charges were calculated using pseudospectral generalized valence bond (ps-GVB) (Ringnalda *et al.*, 1993) calculations on the valence minimized (*i.e.*, all bonds, angles, torsions and inversions at equilibrium) geometry. Ps-GVB is a novel numerical technique that uses both a basis set and a Cartesian grid to reduced Hartree-Fock calculations from order n^4 to n^3 which results in a significant speedup for calculations on the large (by *ab initio* standards) molecules simulated here, without a noticable reduction in accuracy.

Equilibrium and Convergence

Free energy is an equilibrium property of systems, and as such can only be derived from equilibrium ensemble of states. The determination of when equilibrium has been reached for a simulation can be made by observation of such experimentally verifiable equilibrium properties such as heat capacities and diffusion constants. For

the systems reported on here, heat capacities were used as a guideline for judging achievement of equilibrium.

The free energy of a system is also an average value averaged over all of conformation space, which by definition is infinite in extent. Since any simulation possible will obviously only be finite, there is no guarantee that a sufficiently large sample of the available conformation space has been sampled to generate an accurately converged average. One test of this is to run simulations in both directions along the reaction, if the two determinations of relative free energy agree, a high degree of confidence can be assigned to the derived average. The divergence from the truly convergent answer can be estimated by splitting a given ensemble into ten equal parts and calculating a value for each fraction. The deviation in this average is widely considered an accurate measurement of convergence (Fleschman and Brooks 1987). In general, convergence will improve as the number of intermediate hybrids increases, until it reaches some maximum, after which some increased divergence is seen. For the extreme case, called the "slow growth technique" because the species under investigation grows slowly towards the end species, changing with each dynamics step, significant errors in convergence can occur (Mitchell and McCammon, 1991). In chapter 5 (figure 5) we discuss a complete analysis of how convergence depends on the number of intermediate hybrids for one of the simulations reported here.

Some issues that slow convergence include the existence of rotational or other isomers that interconvert slowly on the timescale of the simulation (Straatsma and McCammon, 1989). For the two systems investigated here, the conversion of a

hydrogen to a rotationally fixed amino group and to a rotationally highly symmetric methyl group, no such slowly interconverting isomers exist so convergence times are relatively rapid. Typically, the observed average was reasonablly well converged in 25 picoseconds, although longer simulations (up to 200 picoseconds for some systems) were conducted to check for deviations, as reported in chapter 5.

Estimation of Error

There are several possible sources of error in results obtained using the perturbation thermodynamic technique, including errors in the molecular description and errors in convergence.

Table 2: A collection of molecular dynamic perturbation thermodyamics results for a variety of systems, indicating the type of accuracies achievable with the force-field/molecular dynamics description of molecular systems (from Lybrand *et al.*, 1985a; Lybrand *et al.*, 1985b; Jorgensen *et al.*, 1988; Rao *et al.*, 1987).

Transformation	Simulation R.F.E.(kcal/mol)	Experimental R.F.E.(kcal/mol)
$Cl^- \rightarrow Br^-$	3.2 ± 0.2	3.15
SC-24:$Cl^- \rightarrow$ SC-24:Br^-	4.15 ± 0.35	4.0
CH_4(aq) $\rightarrow$ Cl^-(aq)	-79.1 ± 2.0	-77.0
Subtilisin 155Asn→Ala Δ G_{cat}	3.4 ±1.1	3.67

Errors in the forcefield description of molecular energetics, which is covered in more detail in the first chapter of this thesis, is not particularly easy to systematically assess, but can usually be estimated, at least qualitatively, by comparison with related cases–or in the many perturbation thermodynamics "test" cases done to date, with the experimental determination of exactly the same transformation. Several diverse test cases are listed in table 2, which represent the broad range of perturbation simulations that have been conducted and are illustrative of the degree of accuracy that can be achieved by the forcefield description. For the two systems investigated in the following chapters, the pertinent experimental work is discussed.

Errors in simple convergence are somewhat more easily estimated in a systematic fashion than errors in the forcefield. When simulations are conducted in a single direction, the traditionally accepted technique has been to determine the standard deviation of a set of ten internal perturbation thermodynamic averages (Fleischman and Brooks, 1987). This is, of course, impossible if less than ten data points are taken for each hybrid, for techniques such as the slow growth technique. In such situations, error is usually calculated by estimating hysterisis (Dang *et al.*, 1989; Mitchell and McCammon, 1991).

A more troublesome problem with convergence errors stems from the fact that some molecular systems can populate multiple conformational families, each of which should contribute to the free energy, but which are separated by significant kinetic barriers. Unless simulations long enough to overcome these kinetic barriers has been conducted, the convergence observed will be indicative only of the fact that one subpopulation of the conformations available to the system has been sufficiently

well sampled. While a solution to this problem has been formulated (Straatsma and McCammon, 1989), it is of little impact on the work reported here because the systems being simulated in this work do not exhibit the type conformational isomerism that causes this problem.

For the work reported here convergence was estimated by determining the standard deviation for multiple perturbation simulations. Each perturbation result reported was the average of two or more simulations, with equal numbers of simulations conducted in each of the two reaction directions. The observed hysterisis is a rather rigorous estimate of non-convergence errors (Mitchell and McCammon, 1991).

It has be appreciated for some time (Bash *et al.*, 1987; Fleischman and Brooks, 1987) that for free energy calculations which are dominated by electrostatic or hydrogen bond changes (*e.g.*, methanol to ethane), simulations quickly converge on accurate free energies. This is because dipolar reorientation of the solvent, which is the dominant contributor to the free energy of these interactions, is relatively rapidly achieved. Unfortunately, nonpolar perturbations, such as the uracil to thymine case reported in chapter 5, are determined mainly by exchange repulsion and dispersion attraction van der Waals terms. To accurately sample the structural changes that accompany such free energy changes requires considerably more simulation time, as large scale solvent reorientation to fill or create voids must be achieved (Sun *et al.*, 1992). This discrepancy in convergence times is illustrated in the following chapters, where it is reported that the transformation of a hydrogen to an amino group (chapter 6) is much more rapidly convergent than the transformation of a

hydrogen to a methyl group (chapter 5).

Biological Perturbation Thermodynamics

Advances in methodology and computational power have allowed the application of perturbation thermodynamic simulations techniques to a variety of research questions on the energetics of biological macromolecule interactions. Particular focus has been made on studies of protein-substrate interactions, while substantially less effort has gone into simulations of the nucleic acids.

A paradigm example of biological perturbation thermodynamic analysis are the set of simulation experiments conducted on the protease trypsin. Several free energy simulations have explored inhibitor binding and the catalytic mechanism in the native and mutant forms of the enzyme. The relative binding affinities of two inhibitors, benzamidine and a *p*-flouro derivative of benzamidine, have been successfully determined by a 22 picosecond perturbation thermodynamics simulation conducted on the native enzyme (from the crystal structure) solvated by 5000 water molecules. The calculated $\Delta\Delta A$, 0.91 ± 0.53 kcal/mol (Wong and McCammon, 1986), is in good agreement with the experimentally observed value of 0.5 ± 0.31 kcal/mol (Mares-Guia *et al.*, 1977). Studies undertaken with the protein mutant Gly216→Ala216 have also been conducted, for which the simulation value, 1.33 ± 0.51 kcal/mol, is also in good agreement with experimental value of ~ 2 kcal/mol (Wong and McCammon, 1986).

Another protease, thermolysin, has also been the subject of intensive theoretical investigation. Simulations have been conducted to illucidate the relative free

energy of binding of a class of phosphonamidate peptide analog inhibitors, which are thought to mimic transition-state configurations of enzyme-substrate complexes. These simulations included a calculation of the effects of relative solvation energies and concluded that the relative binding energy of the two inhibitors studied was 4.21±0.54 kcal/mol, in excellent agreement with the experimental result of 4.1 kcal/mol (Bash *et al.,* 1987). The actual ΔG of binding was calculated to be 7.5 kcal/mol, but a -3.58 kcal/mol ΔG of solvation also contributed to the total relative free energy.

Catalysis has also been investigated with some degree of sucess using perturbation thermodynamics. Utilizing transition state analogs, perturbation thermodynamics analysis were conducted to calculate the $\Delta\Delta G_{cat}$ and $\Delta\Delta G_{bind}$ for the enzyme subtilisin and two of its substrates (Rao *et al.,* 1987). While only reasonable agreement was seen between the theoretical and experimentally derived binding energies (0.11 $\pm$ 0.80 versus 0.41 kcal/mol), excellent agreement was achieved for $\Delta\Delta G_{cat}$ (3.40 $\pm$ 1.13 versus 3.67 kcal/mol). Similar studies undertaken with the enzyme α-lytic protease (Caldwell *et al.,* 1991) and several different peptide substrates have also been reported. In these studies, the correlation between theory and experiment was also good, with a calculated $\Delta\Delta G_{cat}$ of 0.88 versus the experimentally derived 1.32 kcal/mol.

Presently, very few free energy simulations have been reported for nucleic acid systems such as DNA and RNA. A qualitatively accurate simulation has been reported for the relative free energy of formation of cytosine and 5-methylcytosine containing triple helical DNA (13.5 kcal/mol versus 7.5 kcal/mol experimental;

Hausheer *et al.*, 1992). More refined work has been conducted on the absolute free energy of solvation of several of the nucleotides (Ferguson *et al.*, 1992), for which there was reasonable agreement between the theoretical results (-14.9 and -9.43 kcal/mol for 9-methyladenine and 1-methylthymine respectively) and experiment (-13.6 $\pm$ 1.1 and $\sim$ -10 kcal/mol).

Conclusions

Computer simulation techniques for the determination of free energies have come of age and are of increasing utility in the study of many systems of biological relevance. Despite the approximations inherent in the forcefield description of molecules and molecular motion, and despite the computational complexity of simulating long periods of molecular motion, demonstrably realistic ensembles can be generated for a variety of molecular systems and from them, an accurate estimation of free energy can be made. These estimates are often difficult to achieve experimentally, thus simulation offers insights into the mechanisms behind the formation of these systems that would otherwise be impossible to achieve.

References

Bash, P. A., Chandra Singh, U., Brown, F. K., Langridge, R., and Kollman, P. A., (1987), *Science*, **235**, 574-576

Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., and Hermans, J., (1981), in *Intermolecular Forces*, B. Pullman ed. (Reidel, Cordrecht, Holland) 331

Berkowitz, M., Karim, O. A., McCammon, J. A., and Rossky, P.J., (1984), *Chem. Phys, Lett.* , **105** 154

Beveridge, D. L. and DiCapua, F. M., (1989), *Annu. Rev. Biophys. and Biophys. Chem.*, **18**, 431-493

Caldwell, J. W., Agard, D. A., and Kollman, P. A., (1991), *Proteins*, **10**, 140-148

Chandrasekhar, J., Smith, S. F., and Jorgensen, W. L., (1985), *J. Am. Chem. Soc.*, **107**, 577 - 580

Dang, L. X., Mertz, K. M., Kollman, P. A., (1989), *J. Am. Chem. Soc.*, **111**, 8505-8509

Ferguson, D. M., Pearlman, D. A., Swope, W. C., and Kollman, P. A., (1992), *J. Comp. Chem.*, **13**, 362-370

Fleischman, S. H., and Brooks, C. L., III., (1987), *J. Chem. Phys.*, **87**, 3029-3034

Hausheer, F. H., Singh, U. C., and Saxe, J. D., (1992), *J. Am. Chem. Soc.*, **114**, 5356-5362

Jacucci, G., and Rahman, A., (1981), *Il Nuovo Cimento*, 4, 341-355

Jorgensen, W. L., Blake, J. F., and Buckner, J. K., (1988), *Chem. Phys.*, **129**, 193-200

Lybrand, T. P., Ghosh, I., and McCammon, J. A., (1985a), *J. Am. Chem. Soc.*, **107**, 985-986

Lybrand, T. P., McCammon, J. A., and Wipff, G., (1985b), *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 833-836

Mares-Guia, M., Nelson, D. L., Rogana, E., (1977), *J. Am. Chem. Soc.*, **99**, 2336-2339

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M.N., and Teller, A. H., (1953), *J. Chem. Phys*, **21**, 1087-1092

Mezei, M., and Beveridge, D. L., (1986), in *Computer Simulations of Chemical nad Biomolecular Systems*, Beveridge and Jorgensen, eds, (New York Acadamy of Sciences Press, New York), 1-23

Mitchell, M. J., and McCammon, J. A., (1991), *J. Comp. Chem.*, **12**, 271-275

Nosé, S., (1984), *J. Chem. Phys.*, **81**, 511

Pangali, C., Rao, M., and Berne, B.J., (1979), *J. Chem. Phys.*, **71**, 2975 -2976

Rao, S. N., Singh, U. C., Bash, P. A., and Kollman, P. A., (1987), *Nature*, **328**, 551-554

Muller, R. P., Marte, B., Won, Y., Donnelly, R. E. Jr., Pollard, W. T., Miller, G. H., Goddard, W. A. III, and Freisner, R. A., (1993),*PS-GVB*, v0.08, Schroedinger, Inc., (Pasadena, CA)

Straatsma, T. P., and McCammon, J. A., (1989), *J. Chem. Phys.*, **90**, 3300-3305

Sun, Y., Spellmeyer, D., Pearlman, D. A., and Kollman, P. A., (1992), *J. Am. Chem. Soc.*, **114**, 6798-6801

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. Jr., and Weiner, P., (1984), *J. Am. Chem. Soc.*,**106**, 765-783

Wong, C. F., McCammon, J. A., (1986), *J. Am. Chem. Soc.*, **108**, 3830-3835

Zwanzig, R. W., (1956), *J. Chem. Phys.*, **22**, 1420

Chapter 5

Analysis of the Free Energy of Solvation of the Thymine Methyl Group: The Role of Solvation in Sequence Specificity

Abstract

Experimental results indicate that interactions with the 5-methyl group of thymine often account for around 1 kcal/mol of the total selectivity at A·T base pairs in protein-DNA complexes. The limited ability of methyl groups to form non-covalent interactions of this magnitude has led to the hypothesis that the energy of solvation of this hydrophobic element is responsible for the observed contribution to selectivity, but this has proven difficult to test experimentally. Here we report a perturbation thermodynamic molecular dynamics (PTMD) analysis of the relative free energy of solvation of thymine and uracil, both as the free bases and in the context of double stranded DNA. The use of PTMD allows us to quantitatively estimate the role that the solvation of this group plays in the generation of sequence specificity in protein-polynucleotide interactions. Our results indicate that the effect of shielding this group from solvent should account for 0.90 kcal/mol of the observed contribution to specificity in protein-DNA complexes. The relative free energy of solvation is 2.3 kcal/mol for the more solvent exposed interactions that occur in the free bases. These observations have importaint implications in the mechanism of sequence specific DNA recognition and the evolution of the deoxynucleotide synthesis pathways.

Portions of the text of this chapter will compose an article coauthored with William A. Goddard III. It is to be submitted to the *Journal of Biological Chemistry.*

Introduction

The role of the thymine methyl group as a major element in determining the sequence specificity of many DNA binding proteins that recognize major groove elements has been experimentally recognized for more than a decade (Goeddel *et al.*, 1977). This functional group has been demonstrated to contribute to the selectivity of a wide variety of protein-DNA complexes, including the Lac repressor, the Cro repressor, the Catabolite Gene Activator protein, and RNA polymerase, in which it makes contributions of between 0.6 and 1.6 kcal/mol to the total sequence selectivity at individual base pairs in the recognition sequence of these proteins (Goeddel *et al.*, 1977; Takeda *et al.*, 1989; Gunasekera *et al.*, 1992; and Dubendorff *et al.*, 1987). As shown in figure 1, this hydrophobic element is highly solvent accessible via the major groove of DNA and easily interacts with protein structures that bind this feature of polynucleotides.

The thymine methyl group projects, as shown in figure 1, into the major groove of DNA where it is accessible both to proteins in protein-DNA complexes and to the solvent in free DNA. Traditionally, this selectivity is said to be imparted to thymine containing protein-DNA complexes by van der Waals interactions between the protein and this methyl group. However, the ability of methyl groups to form strong intermolecular interactions is extremely limited, with typical calculated (Bohm *et al.*, 1984; Novoa and Whangbo, 1991) and experimental (Schamp *et al.*,

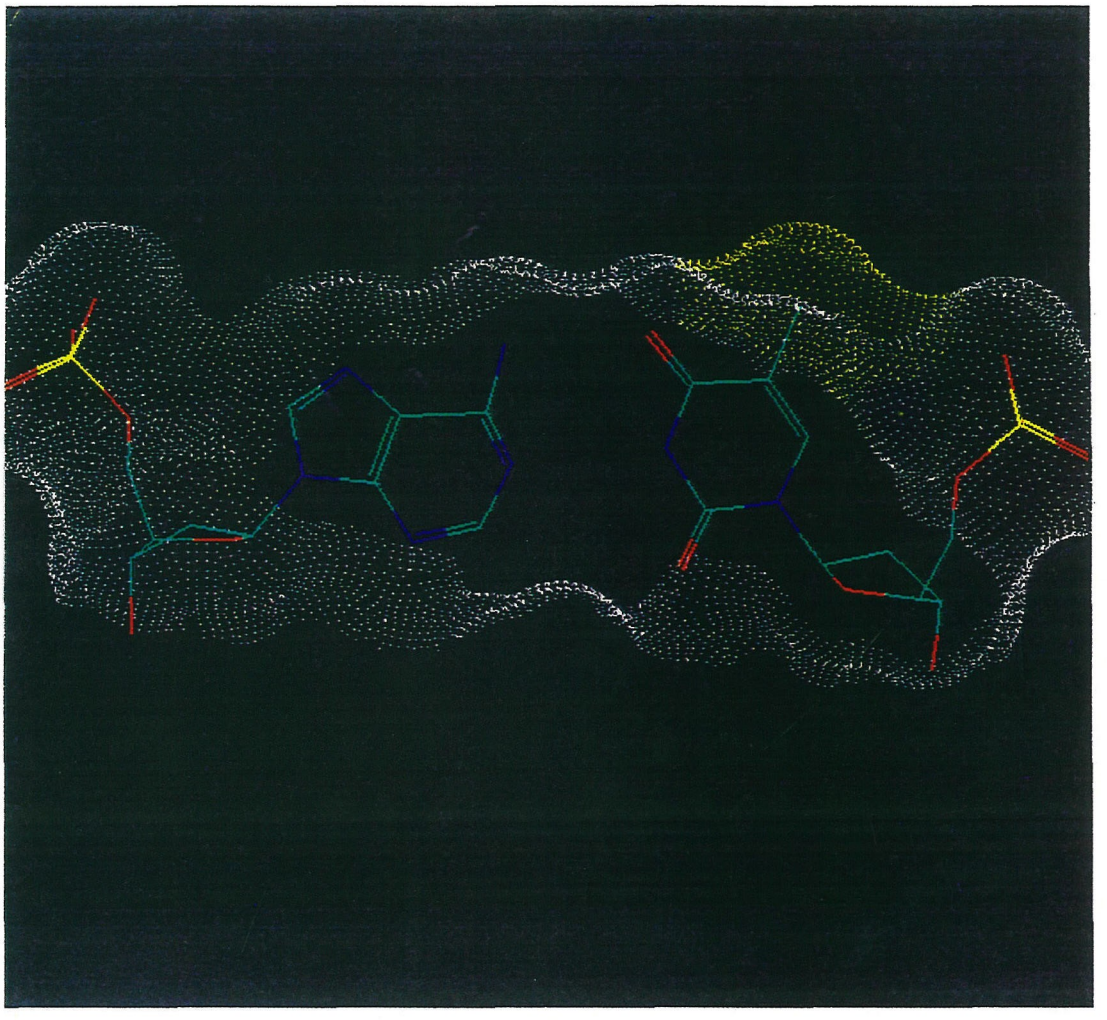


Figure 1: The solvent accessable surface of thymine·adenine (yellow) and uracil·adanine (white). Note that the thymine methyl group protrudes into the major groove generating a large hydrophobic surface area (yellow). The solvent accessable surface was calculated using 1.4 Å probe with atomic radii from AMBER (Weiner *et al.*,, 1984)

1958; Hanley and Klein, 1972; Matthews and Smith, 1976) gas phase interaction energies of about 0.2-0.5 kcal/mol, substantially less than the energy of selectivity observed in many protein-polynucleotide complexes. Because of this discrepancy it has been postulated that the bulk of the selectivity imparted by the thymine methyl group stems from the fact that complex formation results in the desolvation of this hydrophobic element. Such elements destabilize the solvated, uncomplexed form of the polynucleotide by increasing the degree of order in the solvent molecules that are in contact with the group, increasing the entropic cost of solvation. This explanation for the contribution to sequence selectivity of the thymine methyl group has been alluded to in the literature (Aiken and Gumport, 1991) and while it is commonly used to describe features of biological molecular recognition (*e.g.,* Yagi, 1973) an inability to experimentally determine relative solvation energies for highly soluable species such as mono- and poly- nucleotides has precluded experimental verification of the hypothesis and it has not entered the lexicon of molecular biology.

The improved accuracy and speed of computer simulation techniques have made it practical to accurately assess this important, but experimentally inaccessible, value. We have used a pertubation thermodynamics based molecular dynamics (PTMD) simulation to determine the relative free energy of solvation of thymine and uracil, both as the free bases and in the context of double stranded DNA, in order to estimate the role solvation energies play in the selectivity of sequence specific DNA binding proteins. Our results have important implications on general issues of sequence selectivity of DNA binding, the metabolism of the deoxynucleotides, and the role of solvation in molecular recognition.

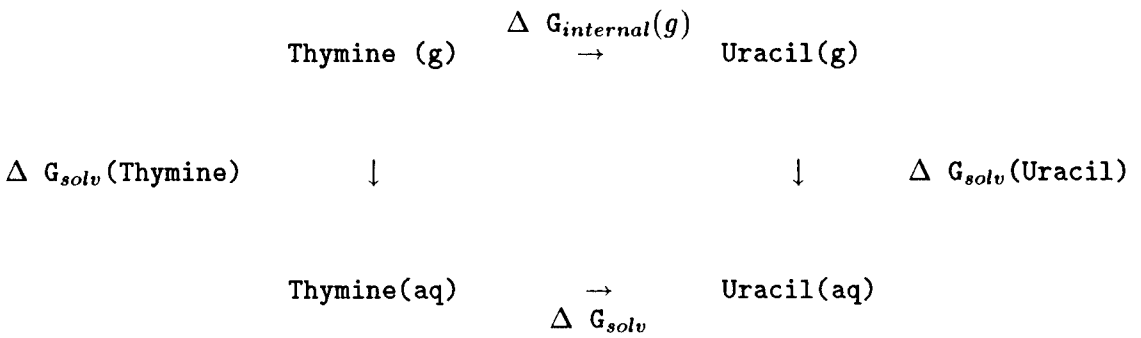


Figure 2: One of the thermodynamic cycles investigated. The simulation assessable values are $\Delta G_{internal}(g)$, and ΔG_{solv}, from which the relative free energy of solvation, $[\Delta\Delta\ G_{solv} = \Delta\ G_{solv} - \Delta\ G_{int}(g)]$ can be derived.

Methods

Perturbation thermodynamic molecular dynamics, which was used to calculate the relative free energy of the pertinent molecular transformations recorded here, has been described in detail in chapter 4. Issues that specifically impact the validity of this work are described in detail below.

A Forcefield Description of Thymine and Uracil

The accuracy of perturbation thermodynamic analysis is, of course, highly dependent on the accuracy of the parameter set used to describe the species involved. For the simulations reported here, accurate forcefield descriptions were needed for both the bases under investigation, and those elements that form the rest of the environment around solvated pyrimidine bases in the context of double stranded DNA.

The van der Waals characteristics and valence of thymine and uracil were taken from the AMBER forcefield (Weiner *et al.*, 1984) (see chapter one for details). This forcefield was specifically optimized for nucleic acids and has been used sucessfully for numerous simulations of nucleic acid hydration (*e.g.,* Miaskiewicz *et al.*, 1993) The amber forcefield was, however, optimized for the united atom representation of molecules (in which C-H groups are characterized as a single atom centered on the carbon nucleus with an increased van der Waals size and the summed charge of the two atoms). This has been shown to be a relatively inaccurate description for solvation energy estimation (Sun *et al.*, 1992), so the full atom representation of thymine

and uracil were used. This entailed the use of quantum mechanical calculations to determine charge distributions for the two molecules. For this, pseudospectral-Generalized Valence Bond calculations (ps-gvb) were conducted (Ringnalda *et al.*, 1993) on the free bases.

Two sets of charges were calculated: one for thymine and one for uracil. Except at positions five and, of course, the C-5 thymine methyl/uracil hydrogen, the charges differed by no more than 10%. These charges were averaged to derive the final forcefield used in these calculations because early simulation work indicated that no significant error in the description of these bases was introduced by this simplification and better convergence was obtained. Further, for investigations of the relative free energy of solvation of the methyl group in the major groove, eliminating perturbations in the minor groove segment of the molecule is a better model of protein induced desolvation than the full perturbation. The three methyl hydrogens (which exhibit differences in charge of 2%, presumably due to rotational electronic anisotropy effects) were each assigned the average charge of 0.1601. The C-5 carbon atom was assigned the calculated charge of -0.1180 in thymine, and the scaled (by 0.1010 q_e) charge of -0.3540 in uracil to create an overall neutrally charged molecule. The calculated charges for uracil and thymine bases are shown in figure 3 with the base charges used in this work. Geometries taken from equilibrium valence descriptions in AMBER (Weiner *et al.*, 1984).

The simulations reported here are simulations used to determine the relative free energies of solvation of several speciecs, and thus highly dependent on the use of an accurate water potential. The water forcefield used in this work, Transferable

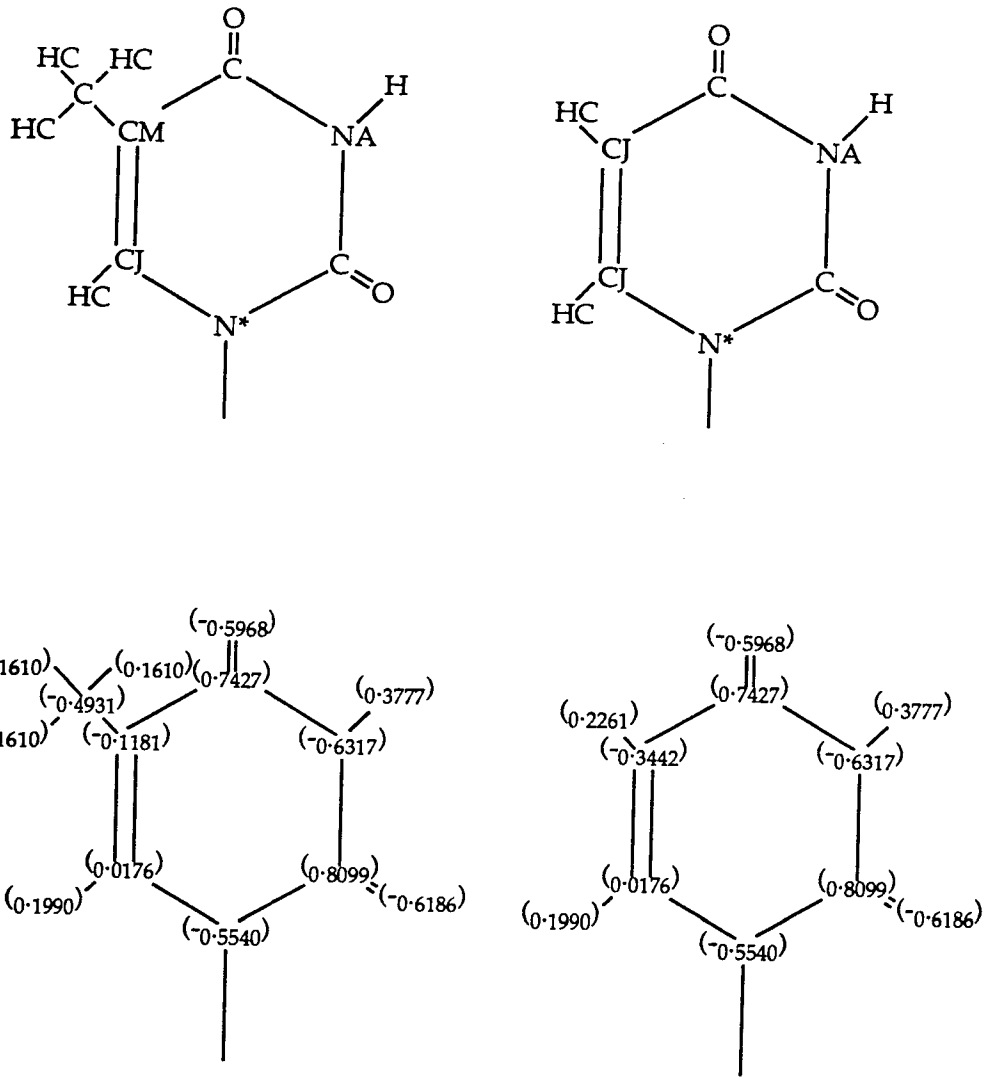


Figure 3: AMBER (Weiner *et al.*, 1983) atom types and ps-GVB derived charges for thymine and uracil.

Intermolecular Potential with three Point charges (TIP3P) (Jorgensen *et al.,* 1983), describes water as a rigid triatomic species with charges located on each nucli and a single van der Waals term centered on the oxygen. This water potential was chosen because it exhibits a high degree of accuracy in describing the degree of order experimentally observed in water, and should be well suited to describe the entropic contributions that dominate the solvation energies of hydrophobic groups. This forcefield has been used successfully to conduct accurate perturbation thermodynamic simulations on both proteins and DNA in solutions (*e.g.*, Caldwell *et al.*, 1991; Bash *et al.*, 1987; Rao *et al.*, 1987). In addition to its reasonably high degree of accuracy, the forcefield is computationally relatively simple, allowing for the rapid calculation of conformational energies and thus for the simulation of large ensembles over which to average. TIP3P and other water potentials are described more fully in Appendix II.

That portion of the DNA molecule not being perturbed was described using the AMBER forcefield (Weiner *et al.*, 1984). This forcefield was specifically optimized for DNA and proteins and has been used successfully for a variety of simulations of DNA hydration (*e.g.,* Miaskiewicz *et al.*, 1993). Two modifications of the standard AMBER description were made. First, the explicit hydrogen bonding interaction was omitted, because the TIP3P forcefield has been optimized to characterize hydrogen bonding without such an addition by the inclusion of explicit hydrogen charges. The second modification of the forcefield was to the phosphate sp^2 oxygen charges, which were reduced from -0.85 to -0.35 electrons to simulate counterion effects. This was necessary, because counterion movements cause large fluctations in the electrostatic environment around DNA which is very slow to con-

verge, and has been shown to be a reasonable approximation (Mills *et al.*, 1992).

The determination of solvent accessible surfaces and volumes was made with the BIOGRAF simulation package (BIOGRAF, 1992) using a 1.4 Å radius probe. The coordinates of B form DNA were taken as indicated in BIOGRAF.

Simulation Methods

The simulations reported here were conducted at constant temperature and constant volume with a molecular dynamics routine we have written for the KSR 1/64 parallel supercomputer. More complete descriptions of this routine appear in Appendicies III and IV. Because the simulations were conducted at constant temperature and volume the free energies calculated rigorously correspond to Helmholtz free energies, but at the level of accuracy expected from this type of simulation no significant difference is expected from the more traditional Gibbs free energy (constant temperature and pressure).

The perturbations conducted here were scaled linearly in λ over the course of the simulations. The perturbations modified bond lengths, charges, and the van der Waals characteristics (note van der Waals terms were linear in r_{eq}^6 not in r_{eq}). The perturbations were conducted by defining a set of atoms, termed "Dummy atoms" (D), at the position of the uracil C-5 hydrogen. These dummy atoms, which lacked van der Waals terms and charges, were systematically given both characteristics and moved towards the coordinates of the thymine methyl hydrogens as the simulation progressed. Some test simulations were conducted without scaling bonds which did

not significantly differ from those with scaled bonds. The perturbation conducted in illustrated in figure 4.

Simulations were conducted independently for the free bases and helical DNA. The simulations of free bases were conducted by placing the base in the center of a 18.94810 Å periodic boundary conditions cube and solvated with 214 randomly placed TIP3P water molecules, as shown in figure 5. This volume consists of the volume of 214 water molecules at 1.0 gram/ml plus the solvent excluded volume of a thymine base. This starting configuration was exhaustively equilibrated (greater than 100 picoseconds) before data collection was begun. The simulations of DNA were conducted using ten base pairs of dA·dT fixed in the B form (BIOGRAF, 1992) and placed in a periodic boundry conditions box of dimensions 40 Å x 40 Å x 33.8 Å. These dimensions mimic the natural periodicity of DNA and create the effect of an infinite helix of DNA, eliminating end effects. This construct was solvated by removing all water molecules within 1.4 Å of any solute molecule from a 1.0 g/ml bath of TIP3P water molecules pre-equilibrated with 100 picoseconds of dynamics and re-equilibrated with the solute molecule for at least 50 picoseconds. These simulations required 1572 water molecules. An illustration of a solvated turn of DNA appears in figure 6. All simulations were conducted using a 1 femtosecond timestep and the shake algorithm (Ryckaert and Berendsen, 1977) to maintain rigid water bonds and angles. A cubic spline cutoff, described more fully in appendix IV, was used to scale nonbond interactions over the range zero to 12.0 Å. Simulation work has indicated that this scaling leads to the best fit of phonon dispersion curves (Goddard, 1993) and leads to good agreement with the observed heat capacity in simulations of bulk TIP3P water (see appendix II).

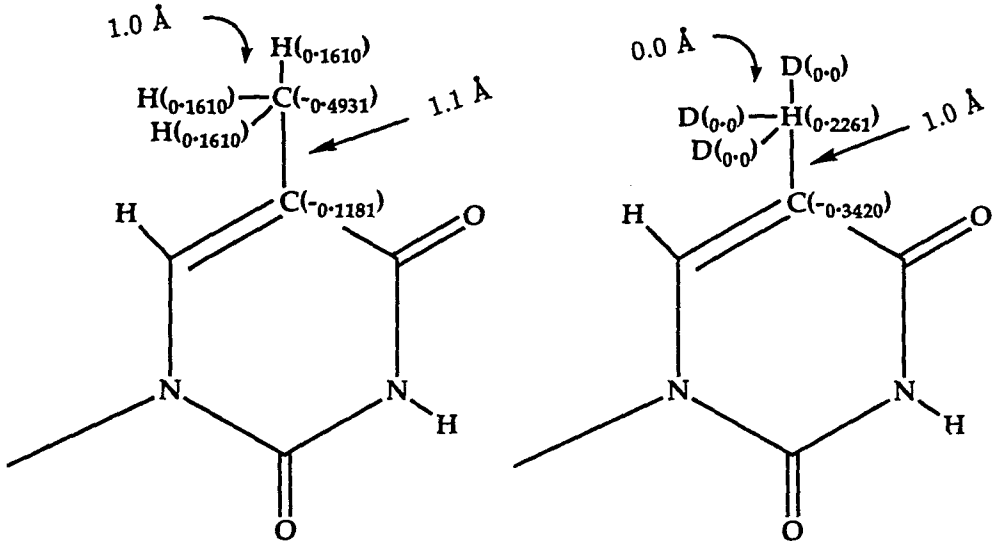


Figure 4: The perturbation conducted for the simulations reported here. "D" represents a dummy atom with no charge or van der Waals parameter. Note bond length scaling.

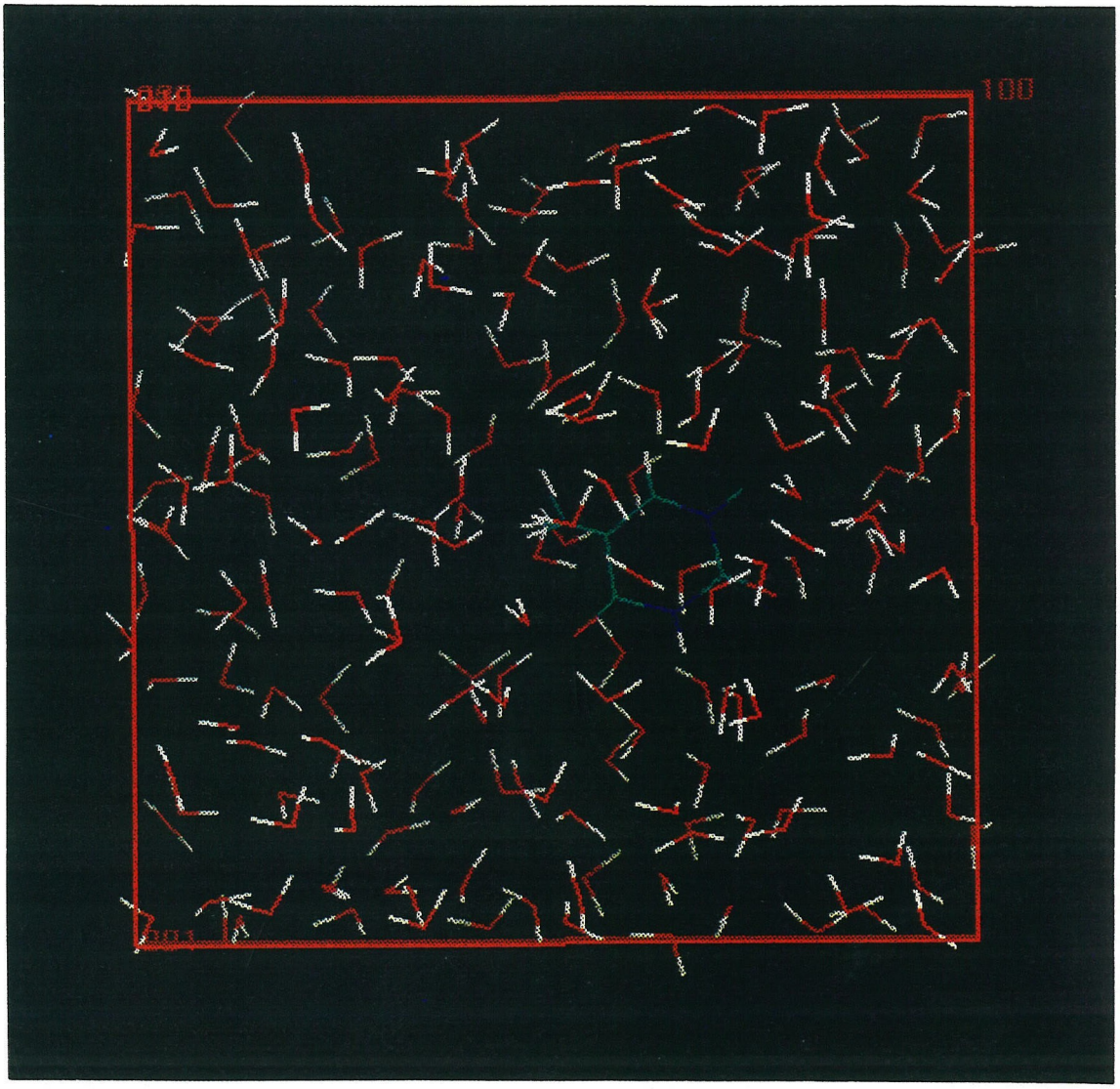


Figure 5: A solvated free thymine base with 214 water molecules in periodic boundary conditions. The box measures 18.92810 Å on a side and contains 4 mmol/l of base.

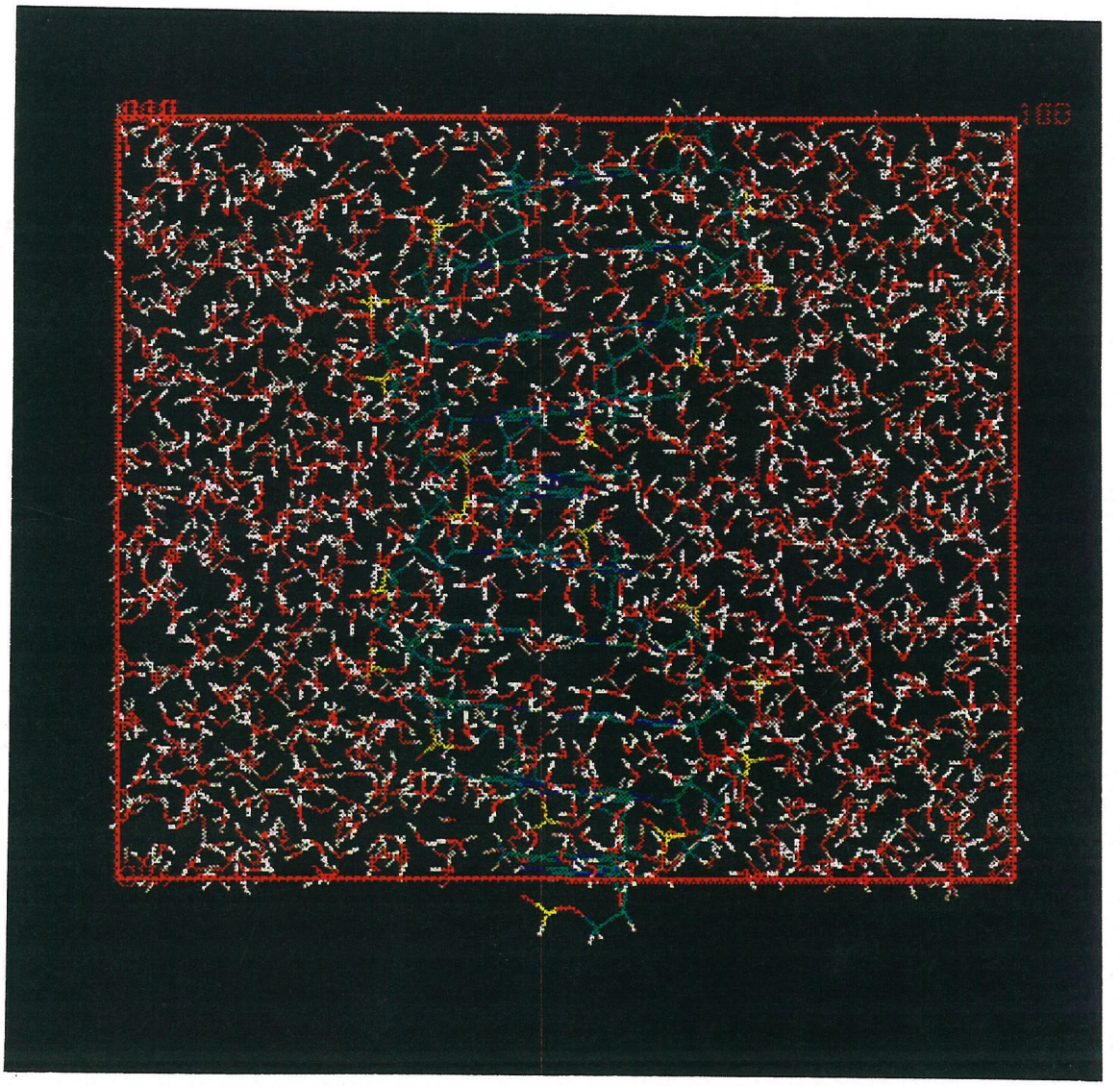


Figure 6: A solvated turn of DNA, with 1572 water molecules in periodic boundary conditions. The box measures 40 Å x 40 Å x 33.8 Å. These dimensions mimic the natural periodicity of DNA and eliminate end effects.

Convergence

Perturbation thermodynamic theory is rigorously true only for averages over all of conformation space. Since any molecular dynamics simulation will necessarily sample only a subset of the infinitely large conformation space available to molecules, particular care must be taken to insure that the average calculated has actually converged onto the correct value. We assayed the convergence of this system by conducting a series of 100 picosecond and longer simulations using modified AMBER parameters to determine how convergence varied as a function of the number of intermediate hybrids. We also investigated how convergence varies as a function of simulation time.

To investigate the effect on convergence of the number of intermediate hybrids, simulations were conducted for 100 picoseconds each in both directions for the free base transformation in 1, 10, 25, 50, 100, 200 and 500 steps (corresponding to zero to 499 intermediate hybrids). One half of the simulation time was devoted to equilibration, and one half to data collection (*e.g.*, for the 25 hybrid case, 2 picoseconds of equilibration was followed by 2 picoseconds of data collection). The results of this investigation, shown in figure 7, indicate that convergence is very well achieved at the 200 hybrid case, and the region from 50-200 hybrids remained relatively well convergent. Longer simulations, up to 200 picoseconds, were no more convergent than 100 picosecond simulations, while 50 picosecond simulations and less were increasingly less convergent. It is interesting to note that simulations involving 500 intermediate hybrids did not converge as well as simulations involving a smaller number of hybrids. We believe that this is due to the extremely short

equilibration time (100 femtoseconds) for each intermediate hybrid not allowing true equilibrium to be established.

Internal Energy Contributions

The contribution of the internal free energy difference between thymine and uracil to the relative free energy of solvation was estimated to be negligible. Because the simulations conducted use a thermodynamic cycle (figure 2) to calculate relative free energies, the difference in internal free energy between the gas phase and the aqueous phase can make a contribution. The internal relative free energy in the gas phase was estimated to be 1.6 kcal/mol, while in aqueous phase it was effectively identical, indicating that this contribution to the relative free energy is negligible. To further test this hypothesis, simulations were conducted with internally fixed and mobile bases and essentially equivalent relative free energies were obtained. All further simulations were conducted on the fixed bases. This approximation has been used (though not explicitly justified as is done here) with limited impact on accuracy in a variety of simulations of the solvation of DNA (Bash *et al.*, 1987; Ferguson *et al.,* 1992). Because the gas phase internal free energy of double stranded DNA cannot be calculated–the double helix is not a stable conformation in the gas phase–this assumption could not be tested for our simulations of helical DNA, but other simulation work has achieved an excellent level of success in describing many features of the solvation of DNA, such as hydration (Poltev *et al.*, 1992) and counter ion enviornment (Mills *et al.*, 1992), using molecules fixed in the B conformation.

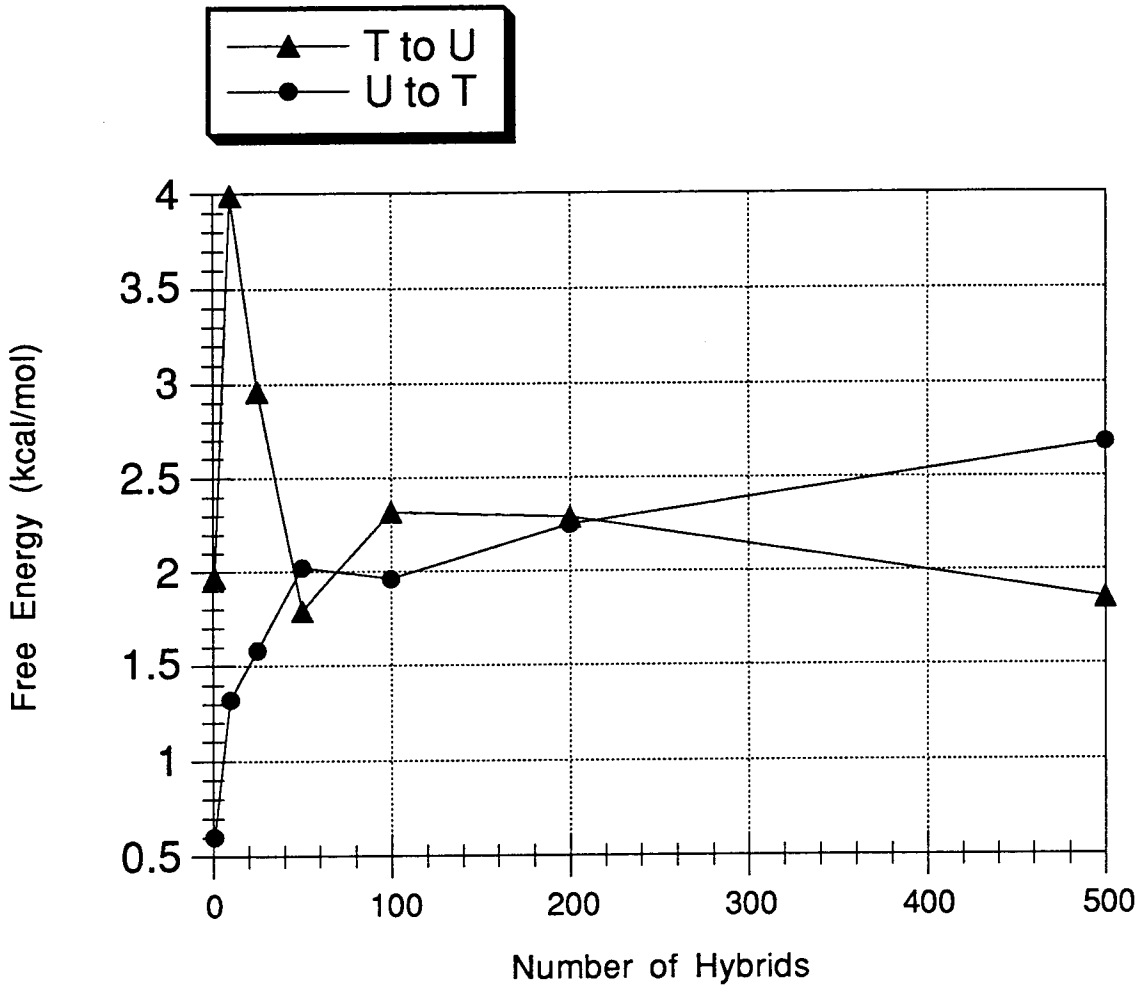


Figure 7: Convergence as a function of the number of intermediate hybrids for 100 picosecond free base perturbation simulations.

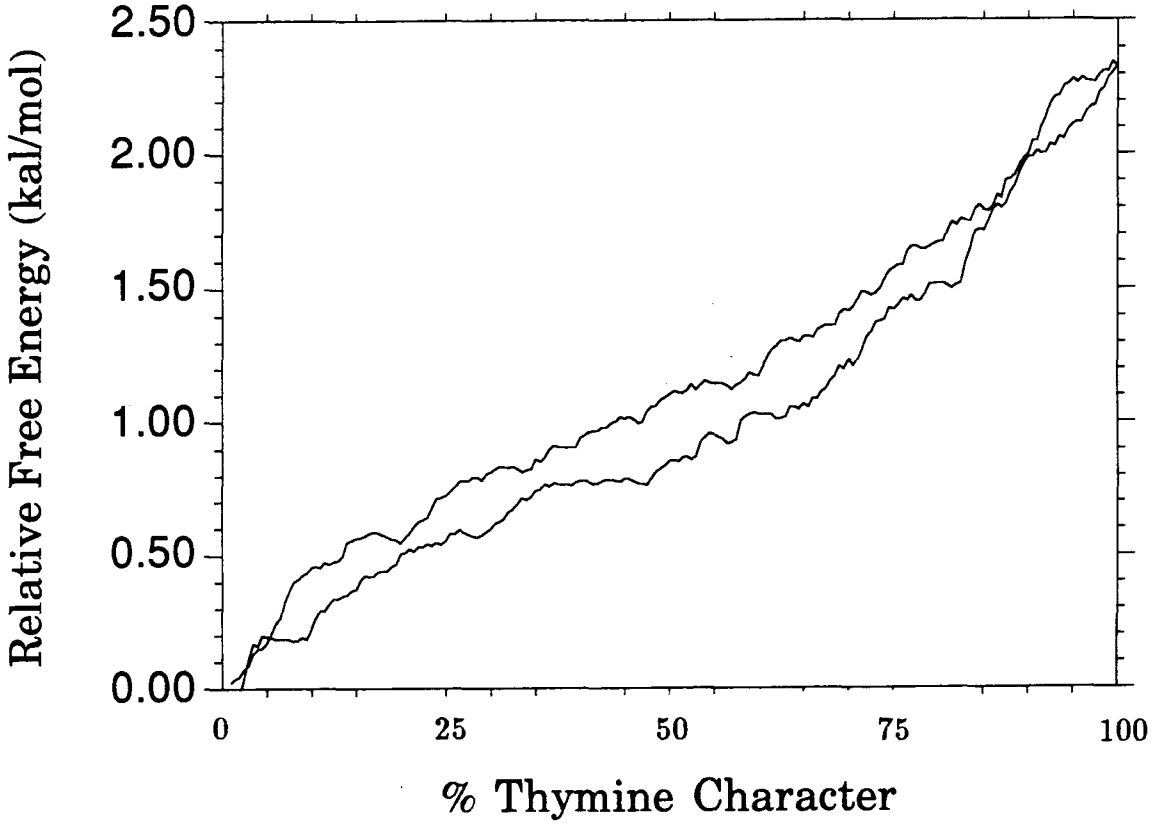


Figure 8: A plot of the free energy profile for the transformation of the free base thymine to uracil and back to thymine. Note the limited hysteresis.

Table 1: The computed relative free energies for each transformation simulated. The change in solvent accessible area is the calculated difference in solvent accessible surface between the two species and was calculated using the BIOGRAF simulation package (BIOGRAF, 1992).

	Transformation	R.F.E.	Δ Area
Free Bases	T→U	-2.36 kcal/mol	24.4 Å^2
	U→T	2.32 kcal/mol	
Helical DNA	T→U	-0.98 kcal/mol	4.0 Å^2
	U→T	0.81 kcal/mol	

Results

The relative free energy of solvation of uracil and thymine was determined to be 2.34 ± 0.03 kcal/mol. The values obtained in each simulation are shown in table 1. This value is much larger than the relative free energy of solvation observed for the thymine methyl group in helical DNA, which is consistent with the much larger solvent exposed surface of the group in the free base. A measure of the hysteresis (and an indication of the completeness of the thermodynamic sampling) is illustrated in figure 8.

In the context of a helix of DNA, the relative free energy of solvation of the thymine methyl group was determined to be 0.90 ± 0.12 kcal/mol. The values obtained in each simulation direction are shown in table 1. The large reduction in the relative free energy of solvation of this group is thought to be due to the significantly reduced solvent accessibility of this group in DNA. The solvent accessible surface area of the methyl group is reduced from 24.4 $Å^2$ in the free base to 4.0 $Å^2$ in double stranded DNA.

Conclusions

We have shown that the primary energetic contributor to the energy of specificity of the thymine methyl group is the cost of solvating the methyl group and not due to van der Waals interactions between hydrophobic elements in the protein and the hydrophobic methyl group.

The 0.90 ± 0.12 kcal/mol estimate of relative free energy of solvation for the thymine methyl group in the context of DNA is in keeping with the observed contribution to selectivity attributed to this functional group in protein-DNA complexes. The discrepancy between this value and the 0.6 - 1.6 kcal/mol range observed in protein-DNA complexes is attributable to both errors in this difficult to measure value, and in a small, but real, van der Waals contribution to the total energy of interaction. Historically it has been assumed that a degree of complementarity (*i.e.*, a hydrophobic surface on each of the interacting faces) is a necessary component for this mechanism of generating sequence specificity. Our results indicate that a substantial energetic contribution to specificity will occur irrespective of the nature

of the recognition surface, as long as methyl group desolvation occurs and there are no prohibitive steric interactions.

Selection against this base (*i.e.* selection of some other base over thymine) requires the formation of prohibitive steric interactions or sequence specific hydrogen bonding. Since the total number of hydrogen bond donors and acceptors is the same in G·C and A·T base pairs the contributions to relative solvation energy of groups other than the 5-methyl group should be limited. We therefore make the following claim: any protein-DNA complex that excludes solvent from a given base pair in the binding site without making sequence specific hydrogen bonds or occlusive steric interactions, will select for A·T (or T·A) base pairs with an energy of selectivity of 0.9 kcal/mol, corresponding to a fivefold increase in the sequence specific K_d.

The simulations we have performed on the free bases have allowed us to estimate the total contribution the thymine methyl group can contribute to recognition of the free thymine nucleotide, which has important implications in the evolution of deoxynucleotide metabolism. In the process of synthesizing dTMP, the corresponding deoxyuracil nucleotide, dUMP, is formed from the reduction of UMP and the deamination of dCMP. The formed dUMP is methylated by thymydilate synthase forming dTMP which is phosphorylated to generate the dTTP precursor need for DNA synthesis. To prevent the inappropriate inclusion of dU into DNA, cellular levels of dUTP must be kept very low (typically less than 1/300 of dTTP concentrations). Our results indicate that ratio cannot be maintained by any mechanism that requires the selective recognition of thymine. Because interactions that select thymine over uracil are limited to desolvation and van der Waals contacts with the

thymine methyl group, at best a 2.8 kcal/mol selection is possible, corresponding to a selectivity of no more than 1/100. The converse recognition, selection of uracil over thymine, can easily result in much higher energies of specificity because repulsive van der Waals interactions (steric clashes) can be significantly stronger than attractive van der Waals interactions This would explain why both prokaryotes and eukaryotes reduce the ratio of dUTP to dTTP by the energetically wasteful process of phosphorylating both dUMP and dTMP and then selectively degrading the dUTP formed rather than by selectively phosporylating only dTMP in the first place.

Unfortunately, while the free energy of solvation of 1-methyl thymine has been estimated from experiment (-9.1-12.7 kcal/mol; Clark *et al.*, 1965) and simulation (-7.9-9.4 kcal/mol; Ferguson *et al.*, 1992; Bash *et al.*, 1987), no measure has been made of the solution free energy of uracil due to the very high solubility of this species. This limits our ability to compare our simulation work with an experimental benchmark to assess the accuracy of the forcefield description that was used.

In this work we have demonstrated that the commonly held view that hydrophobic interactions such as the recognition of a thymine methyl group are primarily dominated by van der Waals interactions to be incorrect. Rather, we have demonstrated that the primary contributor to the energetics of such interactions is the destablization of solvated, uncomplex DNA that contains such elements.

Acknowledgments

The authors wish to thank Dr. Murco N. Ringnalda for performing the pseudospectral-Generalized Valence Bond calculations used in this study, and Dr. Steve Breit of Kendall Square Research for help in optimizing our dynamics routines. Part of the simulation work was conducted at the Cornell Supercomputer center. KWP was supported in part by a National Science Foundation Graduate Fellowship.

References

Aiken, C. R., and Gumport, R. I., (1991), *Methods in Enzymology*, **208**, 433-457

Bash, P. A., Chandra Singh, U., Brown, F. K., Langridge, R. and Kollman, P. A., (1987), *Science*, **235**, 574-576

Beveridge, D. L. and DiCapua, F. M., (1989), *Annu. Rev. Biophys. Biophys. Chem.*, **18**, 431-493

BIOGRAF/POLYGRAF, copyright 1992, Molecular Simulations, Inc. (Pasadena CA)

Bohm, H. J., Ahlrichs, R., Scharf, P., and Schiffer, H., (1984), *J. Chem. Phys.*, **81**, 1389-1395

Caldwell, J. W., and Kollman, P. A., (1986), *Biopolymers*, **25**, 249-266

Caldwell, J. W., Agard, D. A., and Kollman, P. A., (1991), *Proteins*, **10**, 140-148

Clark, L. B., Pesche., G. G., and Tinoco, I., (1965), *J. Phys. Chem.*, **69**, 3615

Dubendorff, J. W., Dehaseth, P. L., Rosendahl, M. S., and Caruthers, M. H., (1987), *J. Biol. Chem.*, **262**, 892-898

Ferguson, D. M., Pearlman, D. A., Swope, W. C., and Kollman, P. A., (1992), *J. Comp. Chem.*, **13**, 362-370

Fleischman, S. H., and Brooks, C. L., (1987), *J. Chem. Phys.*, **87**, 3029-3034

Goddard, W. A., III, (1993), *M. S. C. Technical Note*, **119**, (Materials Simulation Center, California Institute of Technology, Pasadena CA)

Goeddel, D. V., Yansula, D. B., and Caruthers, M. H., (1977), *Nuc. Acids Res.*, **4**, 3038-3055

Gunasekera, A., Ebright, Y. W., and Ebright, R. H., (1992), *J. Biol. Chem.*, **267**, 14713-14720

Hanley, H. J. M., Klein, M., (1972), **76**, *J. Phys. Chem.*, 1743-1747

Jorgensen, W. L., Chandrasekhar, J., Madurea, J. D., Impey, R. W., and Klein, M. L., (1983), *J. Chem Phys.*, **79**, 926-935

Matthews, G. P., and Smith, E. B., (1976), *Mol. Phys.*, **32**, 1719-1725

Mayo, S. L., Olafson, B. D., and Goddard, W. A. III, (1990), *J. Phys. Chem.*, **94**, 8897-8909

Miaskiewicz, K., Osman, R., and Weinstein, H., (1993), *J. Am. Chem. Soc.*, **115**, 1526-1537

Mills, P. A., Rashid, A., and James, T. L., (1992), *Biopolymers*, **32**, 1491-1501

Novoa, J. J., Whangbo, M.-H., and Williams, J. M., (1991), *J. Chem. Phys.*, **94**, 4835-4841

Poltev, V. I., Teplukhin, A. V., and Malenkov, G.G., (1992), *International Journal of Quantum Chem.*, **42**, 1499-1514

Rao, S. N., Singh, C. U., Bash, P. A., and Kollman, P. A., (1987), *Nature*, **328**, 551-554

Ringnalda, M. N., Langlois, J-M., Greeley, B. H., Russo, T. V., Muller, R. P., Marte, B., Won, Y., Donnelly, R. E. Jr., Pollard, W. T., Miller, G. H., Goddard, W. A. III, and Freisner, R. A., (1993), *ps-GVB*, v0.08, Schroedinger, Inc., (Pasadena, CA)

Ryckaert, G. C. and Berendsen, H. J. C., (1977), *J. Comp. Phys.*, **23**, 327

Schamp, Jr., H. W., Mason, E. A., Richardson, A. C. B., and Altman, A., (1958), *Phys. Fluids*, **1**, 329-335

Sun, Y., Spellmeyer, D., Pearlman, D. A., and Kollman, P. A., (1992), *J. Am. Chem. Soc.*, **114**, 6798-6801

Takeda, Y., Sarai, A., and Rivera, V. M., (1989), *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 439-443

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. Jr., and Weiner, P., (1984), *J. Am. Chem. Soc.*, **106**, 765-783

Yagi, K., (1973), In *International Symposium on Oxidases and Related Redox Systems, 2d, St. Jude Children's Research Hospital, 1971*, (University Park Press, Baltimore MD), 217

Chapter 6

Perturbation Thermodynamic Analysis of Netropsin: Sequence Specificity of DNA Binding Driven by Electrostatic Interactions

Abstract

The naturally occurring polypeptide antibiotic netropsin has emerged as a paradigm example of sequence-specific DNA binding and served as the starting point for much of the current efforts at the rational design of small DNA binding ligands. Crystallographic, NMR and molecular mechanics investigations have defined the primary interactions responsible for netropsin:DNA complex formation and have lead to the suggestion that the poly-dA·dT specificity of the drug is due to the close contact observed between the C-2 hydrogen of Adenine and the pyrrole and methylene hydrogens of netropsin that would result in a steric clash in complexes containing the corresponding purine C-2 amino group. The perturbation thermodynamic analysis reported here indicates that this steric interaction is easily accommodated by the complex and that the sequence specificity demonstrated in netropsin:DNA complexes is predominantly driven by electrostatic interactions. Relative solvation energies and steric interactions are demonstrated to play a relatively minor role in the sequence selectivity of the interaction.

Portions of the text of this chapter will compose an article coauthored with William A. Goddard III. It is to be submitted to the *Journal of Biological Chemistry.*

Introduction

The mechanisms of base sequence dependent selectivity and specificity in the interaction of DNA with proteins and low molecular weight ligands is an important issue in biological molecular recognition and has been under intense scrutiny for some time. In order to further our understanding of this important class of interactions, a great deal of spectroscopic, crystallographic and synthetic effort has gone into probing the nature of the sequence specific interactions of the naturally occurring poly-methylpyrrolecarboxamide family of antibiotics.

The antibiotic netropsin, produced by the fungi *Streptomyces netropsi*, is a well studied member of this family. This crescent shaped, peptide linked dipyrrole binds in the minor groove of DNA sites containing four or more successive A·T base pairs (Zimmer and Wahnert, 1987). Footprinting (Fish *et al.*, 1988) and affinity cleavage (Taylor *et al.*, 1984) studies have shown that the molecule binds double stranded DNA as a monomer in the minor groove of stretches of four or more A·T base pairs.

The selectivity of the netropsin:DNA complex has been investigated by a variety of techniques. The association constant for the complex with d(AATT)·d(TTAA) has been determined to be 11 kcal/mol (Marky and Breslauer, 1987) and complex formation on the related sequence d(AAAA)·d(TTTT) has been

determined to be essentially equivalent. The selectivity of this interaction with regard to G·C containing sequences (or closely related species) has been characterized by exploiting the observation that the flourescent 2-amino containing base 2-amino purine is quenched by netropsin binding. By using this technique the binding affinity for a sequence equivalent to d(AGCT)·d(TCGA) was determined to be 6 kcal/mol less favorable than the interaction observed for the d(AATT)·d(TTAA) sequence (Patel *et al.*, 1992) Finally, binding to poly dG·dC was observed to be less favorable than binding to d(AATT)·d(TTAA) by 5 kcal/mol (Marky and Breslauer, 1987). Suprisingly, no studies have been conducted on single base pair substitutions in the binding site of this important class of ligand. The structure of netropsin and some of its significant interactions with the minor groove of DNA are shown in figure 1.

A variety of techniques have been used to illucidate the structure of the netropsin-DNA complex, including crystallography, NMR spectroscopy, and molecular mechanics. Each of these studies has been limited to the investigation of the interactions of netropsin with its high affinity sites, and all have indicated the existence of a close contact between the pyrrole hydrogens and the C-2 hydrogen of Adenine.

Crystallographic studies performed on the netropsin:d(AATT)·d(TTAA) (Kopka *et al*, 1985) and netropsin:d(ATAT)·d(TATA) (Coll *et al.*, 1989) complexes have defined the primary interactions beteen netropsin and poly(dA·dT) DNA. In these structures, a set of bidentate hydrogen bonds were observed between the amide hydrogens of the drug and O-2 and N-3 hydrogen bond acceptors of adjacent thymine and adenine bases. A relatively close contact was also inferred between

the pyrrole and methylene hydrogens of the drug molecule and C-2 hydrogens of adenine. These interactions are illustrated in figure 3.

Complexes of netropsin and several of its high affinity sites have also been investigated using NMR spectroscopy. Early one-dimensional spectra of netropsin complexed with several high affinity sites indicated that complex geometry was essentially identical in both $d(AT)_3 \cdot d(TA)_3$ and $d(A)_6 \cdot d(T)_6$ (Fritzsche and Crothers, 1983). Later nuclear Overhauser effect investigations demonstrated a close contact between the pyrrole hydrogens and the C-2 hydrogen of Adenine (Patel and Shapiro, 1986), in keeping with previous crystallografic investigations.

Molecular mechanics (Caldwell and Kollman, 1986) based studies have been used to determine the structure of netropsin in complex with several of its high affinity sites. These studies, which were conducted using molecular mechanics rather than molecular dynamics (see chapter one for details on the difference between these approaches) and in the absence of solvent generated a model of the complex that was consistent with the conformation observed with crystallographic and NMR based techniques. That such consistency was obtained indicates that the forcefield description of this molecular interaction is accurate. However, because these simulations were based on molecular mechanics and not molecular dynamics, only static enthalpic contributions to the structure were observed, and because the simulations were conducted in the absence of solvent, no solvation contribution to specificity could be determined. A short peturbation thermodynamics investigation of this complex has also been reported, which was in reasonable agreement with the experminetally determined relative free energy of selectivity for the internal sites of

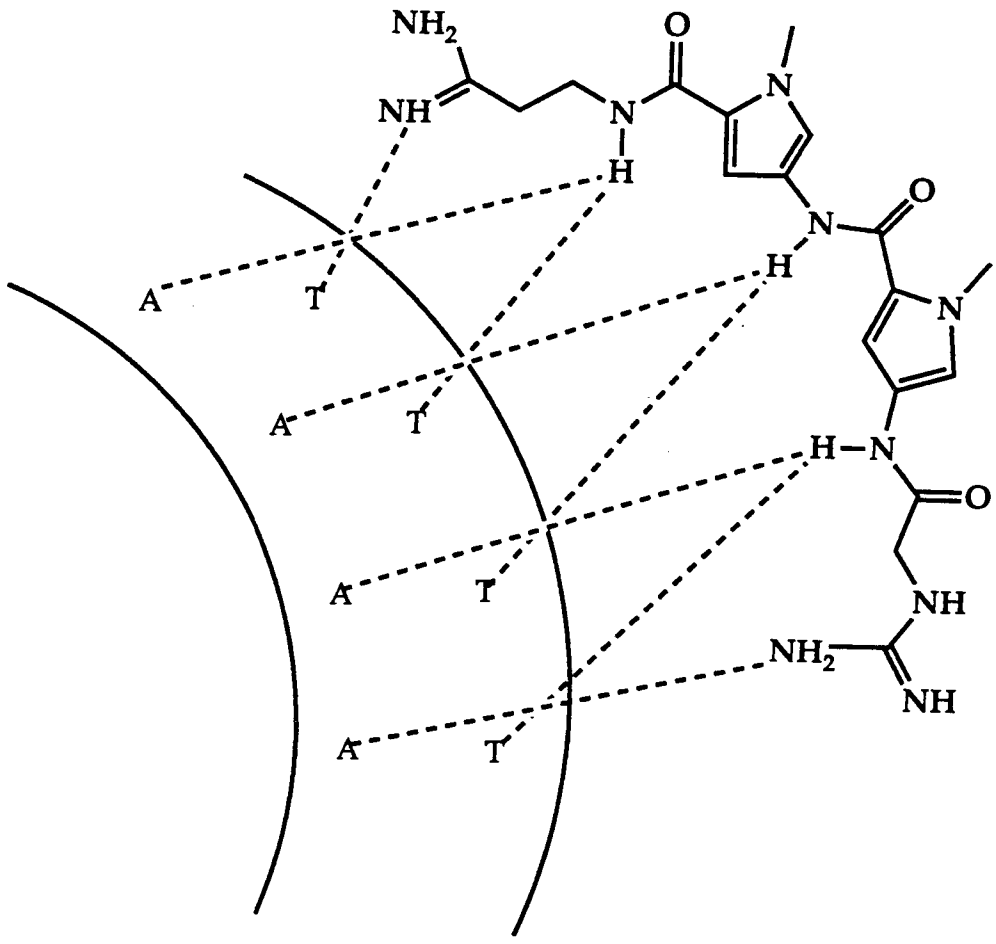


Figure 1: The binding mechanism for Netropsin binding to the sequence d(AAAA)·d(TTTT). "T" represents thymine O2, and "A" the adenine N3 hydrogen bond acceptors, dashed lines represent bidentate hydrogen bonding.

the netropsin-DNA complex, but the simulations were not run in both directions, which is important for estimate true convergence, and no solvation contribution was investigated (Hard and Nilsson, 1992).

A mechanism for the sequence specificity of netropsin-DNA interactions has been postulated from these structural studies. Repeatedly, studies based on crystallography, NMR, and molecular mechanics have indicated that there is a close contact between the pyrrole and methylene hydrogens of netropsin and the Adenine C-2 hydrogen that forms the base of the minor groove (Kopka *et al.*, 1985; Patel, 1982; Coll *et al.*, 1987; Klevit *et al.*, 1986) which would be difficult to accommodate if the static netropsin conformation observed in netropsin:poly-dA·dT complexes were bound to a G·C containing sequence. This has long stood as the proposed mechanism by which the polypyrrole antibiotics demonstrate sequence selectivity. The chemistry of A·T and G·C base pair minor groove character and the proposed close contact are shown in figures 2 and 3.

Despite the mechanistic consistency obtained with such widely diverse experimental techniques, all of the experimental methods used to date to illucidate the mechanism of the selectivity of netropsin have focused on the geometry of its complex with high affinity sites, which could very well lead to their overlooking several important contributions to the selectivity of this class of drug:DNA interactions.

One such overlooked possibility is the role differential solvation may play in the selectivity of the netropsin:DNA complex. If G·C containing sequences are more readily solvated in uncomplexed DNA, then complex induced desolvation of these sequences will be relatively disfavored. While no experimental techniques

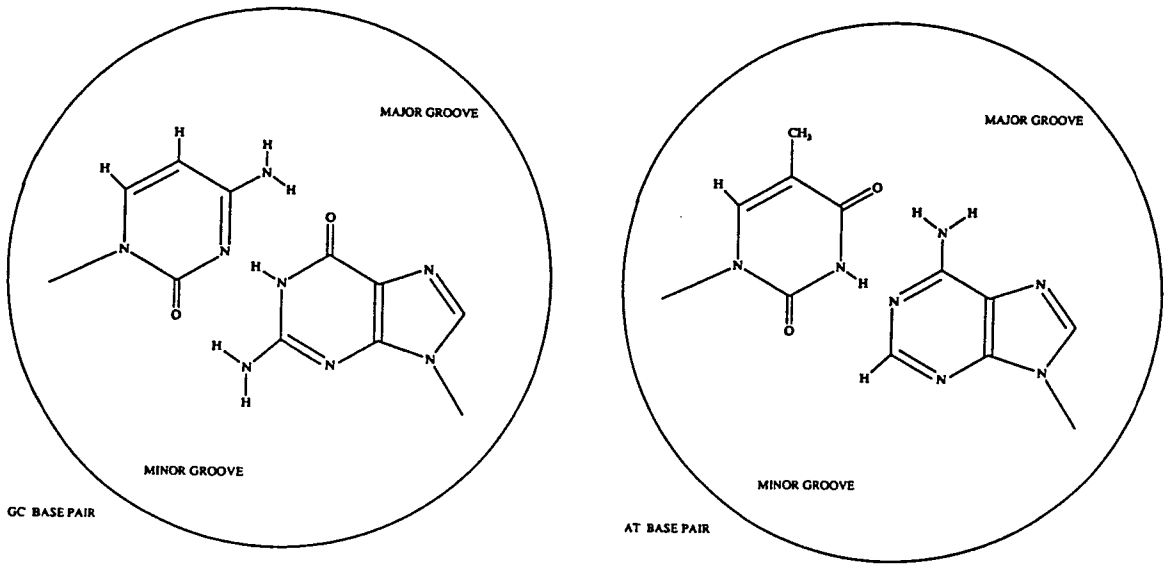


Figure 2: The groove character of A·T and G·C base pairs. Note the guanine 2-amino group protrudes into the minor groove.

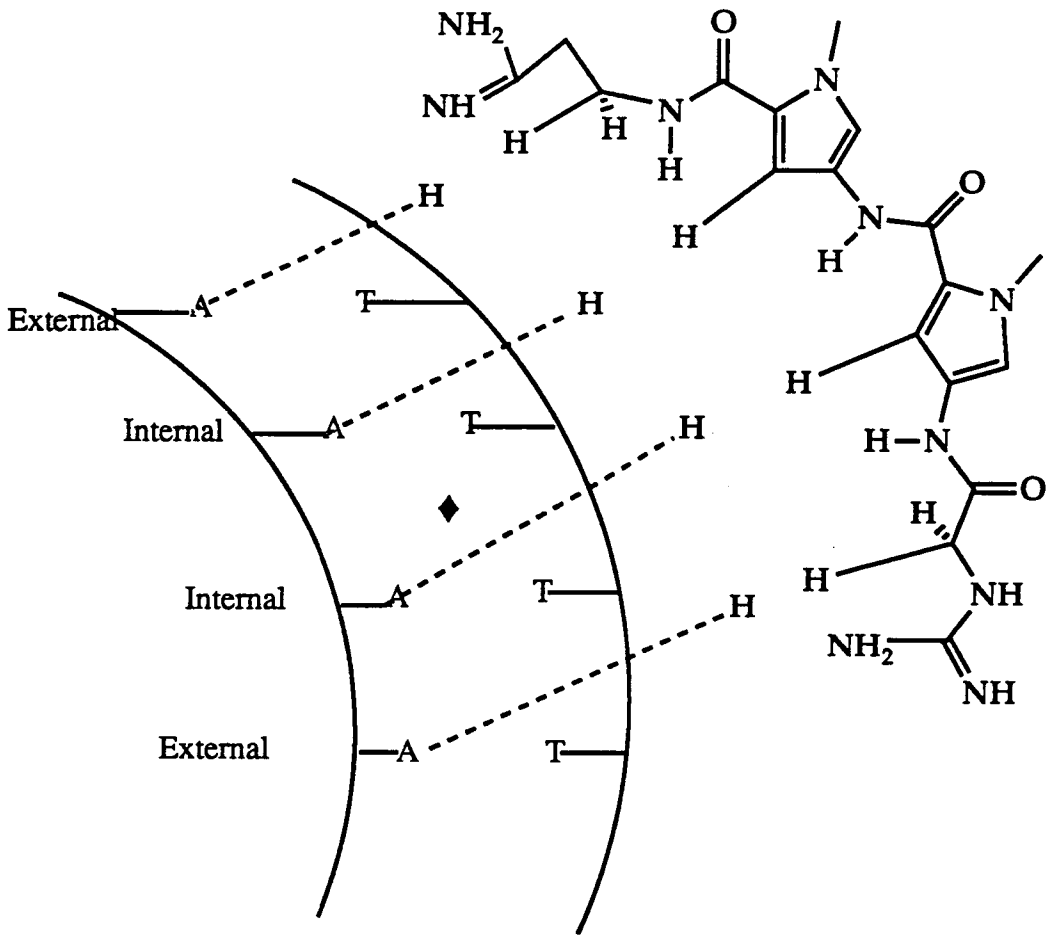


Figure 3: The proposed mechanism for the sequence specificity of the netropsin:DNA complex. Dashed lines represent close steric contacts between various hydrogens on the drug molecule and the adenine C-2 hydrogen. The pseudo-dyad center is marked with a $\diamond$, and internal and external sites are indicated. The interaction geometries of the simulated complex with the the polydA·dT sequence shown here do not significantly differ from those observed in the crystal structure (Kopka *et al.*, 1985).

are presently available to access this contribution to the free energy of complex formation, it is a study particularly suited to simulation.

Because all of the techniques used to date to study netropsin in complex with its cognate sites primarily suited for the determination of inter atomic distances and close atomic packing, a great deal of consideration has gone into the role sterics play in the selectivity of this complex, while almost no consideration has gone into the role of longer range electrostatic interactions. Because the different contributions to a molecular interaction can be easily separated in simulation, this question is easily addressable by computational investigations.

The sequence dependence of minor groove width has also been implicated in the mechanism of specificity for this class of molecular recognition (Fratini *et al.*, 1982). It is difficult to acess the validity of this proposed contribution to selectivity because to date all experimental work has been conducted on complexes with similar, though poorly determined groove widths. Because the simulations reported here were conducted with a constant groove width, this contribution to the sequence selectivity of the netropsin:DNA complex cannot be readily estimated.

Using perturbation thermodynamics, we have undertaken the investigation of this class of DNA binding ligands in order to investigate the role of sterics, solvation, and electrostatics on the sequence specificity of this interaction. Our simulation results indicate that the netropsin:DNA complex is sufficiently mobile as to easily alleviate this postulated steric interaction, and that this interaction is not responsible for the sequence selectivity of this class of DNA binding molecules. Rather, longer range electrostatic interactions the predominant source of the sequence se-

lectivity of this class of DNA binding molecules. Differential solvation effects were noted to make a relatively small but positive contribution to the selectivity of the complex.

Methods

Perturbation thermodynamic molecular dynamics, which was used to calculate the relative free energy of the pertinent molecules involved in complex formation, is described in detail in chapter 4. The thermodynamic cylce used in these investigations is indicated in figure 4. Issues that specifically impact the validity of this work are described in further detail below.

A Forcefield Description of Netropsin and DNA

The accuracy of perturbation thermodynamic analysis is, of course, highly dependent on the accuracy of the parameter set used to describe the species involved. For the simulations reported here netropsin, DNA and water must all be accurately described.

The DREIDING forcefield (Mayo *et al.*, 1990) was used to describe the valence and van der Waals characteristics of the netropsin molecule. This forcefield was choosen because it was designed to be of general utility, and therefore easily and accurately transferable to a description of the netropsin molecule. Because perturbation thermodynamic simulations are sensitive to the united atom approximation

Figure 4: The thermodynamic cycles investigated and the observed free energy differences for the transformations discussed.

(Sun *et al.,* 1992), explicit hydrogens were used. Psuedospectral-Generalized Valence Bond (ps-GVB) calculations were conducted to determine charge distribution for the netropsin molecule. The conformation used for these calculations was a valence force minimized DREIDING structure. We feel the charge distribution used in this work is a significant improvement over earlier work reported in the literature, because all previously reported simulations of netropsin have used charges determined from quantum mechanical calculations on smaller molecules and adapted to the description of netropsin. None the less, our charges are in reasonable agreement with previosly used charge distributions (Caldwell and Kollman, 1986; Hard and Nilsson,1992), and in small test simulations we have conducted led to essentially equivalent results. The charges and DREIDING atom types used in these simulations are shown in figure 5.

The simulations reported here are simulations used to determine the relative free energies of solvation of several speciecs, and thus highly dependent on the use of an accurate water potential. The water forcfield used in this work, TIP3P (Jorgensen *et al.,* 1983), describes water as a rigid triatomic species with charges located on each nucli and a single van der Waals term centered on the oxygen. This water potential was chosen because it exhibits a high degree of accuracy in describing the degree of order experimentally observed in water, and should be well suited to describe the entropic contributions that dominate the solvation energies of hydrophobic groups. This forcefield has been used successfully to conduct simlations of bulk water (Jorgensen *et al.,* 1983) and the solvation of biological macromolecules (*e.g.*, Caldwell *et al.*, 1986; Bash *et al.,* 1987; Rao *et al.,* 1987). In addition to its ability to accurately simulate water and solvation, the forcefield is computationally

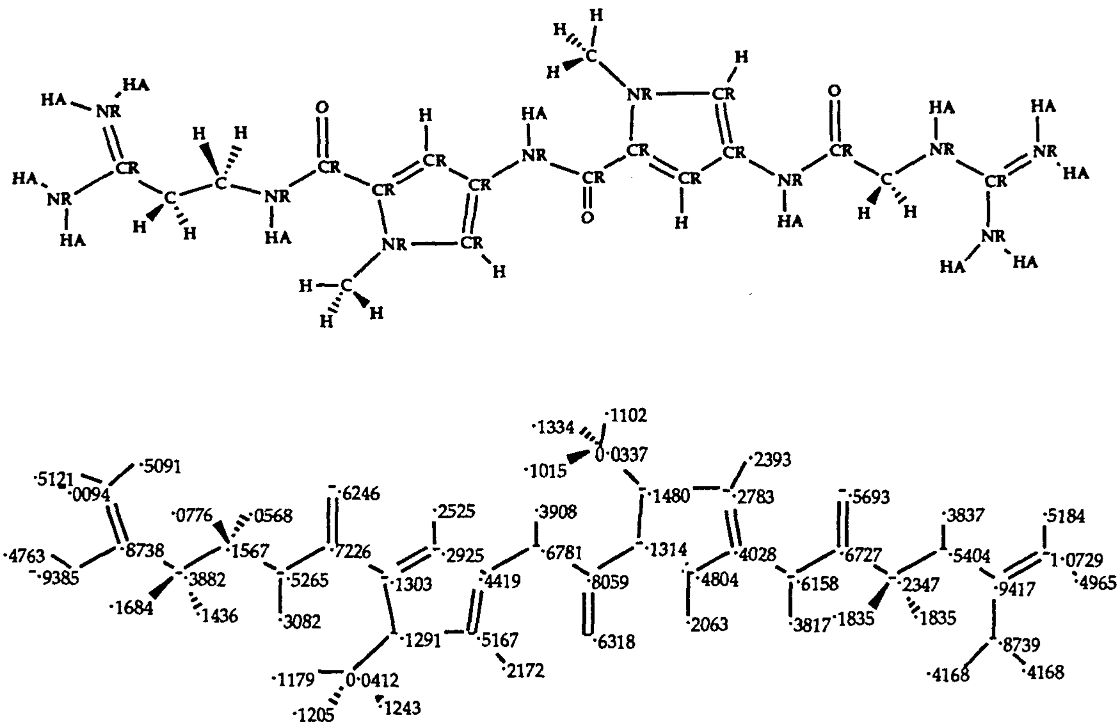


Figure 5: The netropsin molecule with DREIDING forecfield atomtypes (Mayo *et al.*, 1990) and ps-GVB derived charges (Ringnalda *et al.*, 1993). Atoms denoted C represent DREIDING C3 atoms and atoms listed as O denote DREIDING O2 atoms.

simple, allowing for the rapid calculation of conformational energies and thus for the simulation of large ensembles over which to average. TIP3P and other published water forcefields are described more fully in Appendix II.

The parameters used to describe DNA in these simulations were from the AMBER forcefield (Weiner *et al.*, 1984). This forcefield was specifically optimized for use in simulating proteins and nucleic acids and has been used successfully for a variety of simulations of DNA hydration (*e.g.,* Miaskiewicz *et al.,* 1993). Two modifications of the standard AMBER description were made. First, the explicit hydrogen bonding interaction was omitted, because the TIP3P forcefield has been optimized to characterize hydrogen bonding without such an addition by the inclusion of explicit hydrogen charges. The second modification of the forcefield was to the phosphate sp^2 oxygen charges, which were reduced from -0.85 to -0.35 q_e to simulate counterion effects. This was necessary because counterion movent cause large fluctations in the electrostatic enviornment around DNA which is very slow to converge on an average value in simulations we conducted on this system using explicit sodium counter ions. The coordinates of B form DNA were taken as in BIOGRAF (BIOGRAF, 1992).

The perturbations were conducted using A·T and diamino-A·T (daA·T) base pairs to describe the transfrormation of A·T to G·C in the minor groove. The daA·T species is identical to a G·C base pair in the minor groove while A·T like in the major groove, which should simulate the changes of interest in this work while minimizing the size of the total perturbation and thus speeding convergence time.

Several forcefield descriptions of diamino-Adendine (daA) were investigated,

all of which were based on the AMBER parameters for adenine and guanosine, which are shown in figure 6. The first (figure 6a), which directly replaced the C-2 hydrogen of adenine with the 2-amino group of guanine suffered from a net charge, which lead to inappropriate long range interactions with the solvent and the positive charges of netropsin. A second charge set was then generated (figure 6b) by scaling the nitrogen charge by 0.1010 q_e so that the amino group was neutral. While this parameter description lead to qualitativly similar results to those reported here, it suffered from poor convergence, presumably due to the reduced amino group dipole moment, so a four atom perturbation was developed next (figure 6c), in which the C-2 carbon atom was also replaced and the amino nitrogen charge was not scaled. In this simulation, the charge on the guanine amino group was taken directly, while the guanine charge on the guanine 2-carbon was scaled by 0.1010 (to produce a net neutral perturbation). Excellent convergence was obtained with this description, and no net electrostatic monopole interactions were present. Simulations were also conducted by replacing all of the minor groove atoms of A·T with those of G·C (with some minor scaling to maintain neutrality). These results did not significantly differ from the 4 atom perturbation, so the computationally less intensive 4 atom perturbation was used for all of the simulations reported here.

Simulation Methodology

The perturbation thermodynamic molecular dynamics simulations reported here were conducted using a constant temperature molecular dynamics program written for the Kendall Square Research KSR1/64 multicomputer, and described in

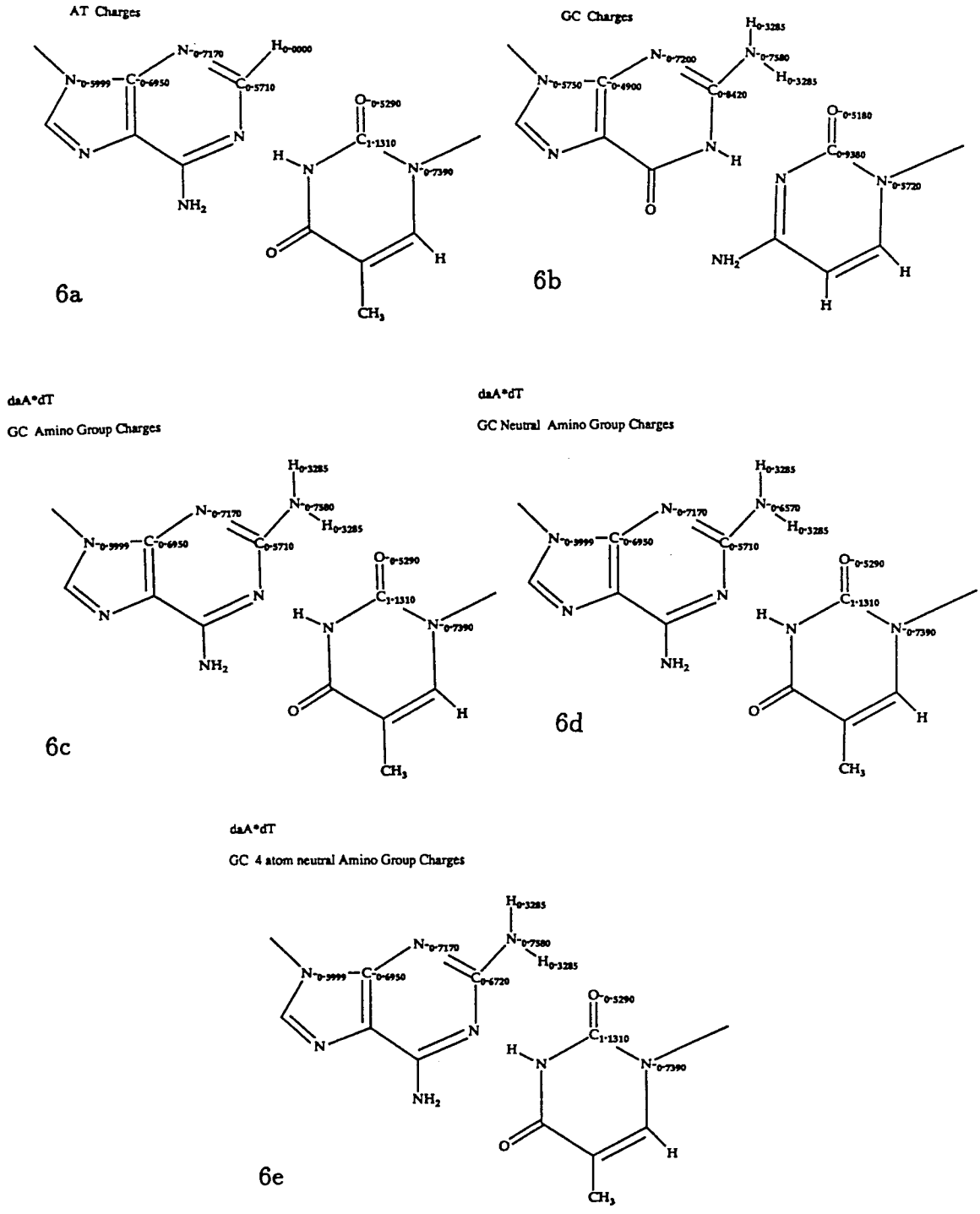


Figure 6: The AMBER charges for A·T and G·C base pairs and the AMBER derived charges for daA·T investigated in this study.

further detail in Appendicies III and IV. To start the simulations, ten base pairs of poly-dA·poly-dT were-fixed in the B conformation (BIOGRAF, 1992) and placed in a periodic boundary conditions box of dimension 40Å x 40Å x 33.8Å. These dimensions mimic the natural periodicity of DNA and create the effect of an infinite helix of DNA, eliminating end effects. Netropsin was manually docked to this sequence using the crystal structure of a netropsin:DNA complex (Kopka *et al.*, 1985) and equilibrated with 25 picoseconds of dynamics. The DNA:netropsin complex and uncomplexed DNA were solvated by removing all water molecules within 1.4 Å of any solute atom from a 1.0 g/ml bath of TIP3P water pre-equilibrated with 100 picoseconds of dynamics and then re-equilibrated with the solute molecule for at least 50 picoseconds. For the two complexes, 1551 and 1572 TIP3P water molecules remained respectively.

All simulations were conducted using a 1 femtosecond timestep and the shake algorithm (Ryckaert and Berendsen, 1977) to maintain rigid water bonds and angles. A cubic spline cutoff was included to scale nonbond interactions over the range zero to 12.0 Å, as described more fully in appendix IV. Simulation work has indicated that this cutoff leads to an excellent fit of phonon dispersion curves and leads to good agreement with the observed heat capacity in simulations of bulk TIP3P water (see appendix II). Pre-equilibration was judged by monitoring heat capacities and was generally complete within 25-50 picoseconds. The ensembles generated were generated at constant temperature and constant volume. Because the volume of the periodic boundary conditions box was held fixed during these simulations, the free energies determined rigrously correspond to Helmholtz free energies, although at the level of accuracy typical for this type of simulation Helmholtz and Gibbs free

energies are expected to be essentially equivalent.

The perturbations were carried out using 100 intermediate hybrids over 25 picoseconds of dynamics. For each hybrid, 0.1 picoseconds equilibration was conducted followed by 0.15 picoseconds of data collection. For all cases, perturbations were conducted independently from both directions and the results averaged. The perturbations were linear in λ. Both van der Waals terms were scaled directly in A and B (*i.e.*, in r_{eq}, not r_{eq}^6) as were charges. Bond lengths were scaled with λ, going to zero for the N-H bond permuted to an H-D (dummy atom) bond. The internal relative free energy contribution to these interactions was assumed to be zero (see chapter 5), and the DNA was held fixed in the B conformation (BIOGRAF, 1992) for the simulations reported here.

The netropsin molecule and the netropsin:DNA complex exhibit a pseudo two fold axis of symmetry as indicated in figure 1. This symmetry reduces the number of types and geometries of interaction that netropsin has with each of the four bases in the recognition site. We have exploited this symmetry in this work by conducting perturbation simulations at only two of the positions in the binding site, which we label internal and external positions, as shown in figure 1. The perturbations were carried out for the sites nearest the guanadinium end of the netropsin molecule.

Geometric analysis were conducted using equilibrated initial structures and sampled every picosecond over 25 picoseconds of non-perturbed dynamics. The poly-dA·dT and the 2-amino containing complexes were each simulated and analyzed independently.

Convergence and Equilibrium

The establishment of equilibrium was determined by an analysis of heat capacities. The degree of convergence was very high as estimated from conducting simulations in both perturbation directions. Fifty picosecond dynamics simulations were also conducted on the uncomplexed DNA and no significant increase in convergence was observed. All perturbation thermodynamic molecular dynamics simulations were determined to be converged in simulations of 100 intermediate hybrids over 25 picoseconds dynamics. Neither the relevant free energy values obtained nor the standard deviation observed were significantly altered in selected longer test simulations.

Results

The relative free energy of solvation of $dA_{10}\cdot dT_{10}$ and $dA_5(dap)A_4\cdot dT_{10}$ were determined to be 0.16 ± 0.04 kcal/mol. Results from both simulation directions are shown in table one. This small solvation effect contributes to the observed sequence selectivity of netropsin complex formation. The relative free energy of solvation of the base pairs has not been determined experimentally, due to the very high absolute solubility of double helical DNA.

The internal position of the complex demonstrated the very significant selectivity. At these positions, netropsin binding to a 2-amino containing sequence was 2.95 ± 0.05 kcal/mol less favorable than binding to the corresponding poly-dA·dT

Table 1: The computed relative free energies for each transformation simulated.

	Transformation	R.F.E.
Uncomplexed DNA	A·T→G·C	-0.19 kcal/mol
	G·C→A·T	0.13 kcal/mol
Complexes		
Internal Site	A·T→G·C	2.91 kcal/mol
	G·C→A·T	-2.99kcal/mol
External Site	A·T→G·C	2.75kcal/mol
	G·C→A·T	-2.70 kcal/mol

sequence, which in contribution with solvation effects produces a relative free energy of binding of 3.11 ± 0.06 kcal/mol over a poly-dA·dT sequence. It is of note that steric interactions between the 2-amino group of this sequence and netropsin contributed no more than a few tenths of a kcal/mol to the interaction energy. The average C-2-pyrrole carbon separation for the poly-dA·dT sequence was 3.5 Å. For the 2-amino containing sequence, this distance had increased to 4.3 Å, which was approximately equally shared by a shift down the minor groove and farther into the solvent.

The external positions of the binding site exhibit a slightly smaller contribution to the overall energetics of complex formation. At these more solvent accessible

positions significant interaction between the amino group and solvent and the amino group and charged elements of netropsin reduce desolvation driven selectivity. With a $\Delta\Delta$ G of binding at these positions of 2.68 $\pm$ 0.11 kcal/mol, the terminal positions contribute 2.84 $\pm$ 0.12 kcal/mol to total selectivity of netropsin. Here too a slight increase in purine C-2 to methylene carbon distance was seen over the poly-dA·dT sequence, amounting to an increase of the average distance of 0.4 Å.

While the relative binding affinity of netropsin for a number of DNA sequences has been determined, no comparisons have yet been made to single position 2-amino group additions, so we are forced to extrapolate from our simulations in order to compare them to experimentally determined specificities. The relative affinity of netropsin for the flourescent 2-amino purine containing sequence d(CTGA(2a-P)TTCAG)$_2$ has been determined (Patel *et al.*, 1992, Marky and Breslauer, 1987). Replacement of both of the internal positions with 2-amino containing bases results in an approximately 3.0 ($\pm$ 0.3, our estimate) kcal/mol reduction in binding affinity for each of the two positions, in excellent agreement with calculated $\Delta\Delta$ G of 3.11 kcal/mol.

Conclusions

It is our conclusion that, based on the observed energetics of the netropsin:DNA interaction, and the small observed shift in netropsin geometry upon complex formation with 2-amino purine containing sequences that the primary mechanism of selectivity of this class of DNA binding molecules is derived from electrostatic interactions with the minor groove of G·C and A·T base pairs and not, as previously

hypothesized, due to unfavorable steric interactions between the 2-amino group and the pyrrole and methylene hydrogens of netropsin. Others have suggested (Fratini *et al.*, 1982; Nelson *et al.*, 1987) that the sequence specificity of this molecular interaction stems from improved van der Waals contacts in the narrower narrow groove of poly-dA·dT. We offer no explicit evidence that this mechanism does not play a role in sequence selectivity, but our results do indicate that at the very least, electrostatic effects are a major contributor to the sequence selectivity of this molecular interaction and steric and solvation considerations play a minor role.

This conclusion is consistent with with several experimental results that are difficult to explain with the steric clash mechanism of sequence specificity.

First, CD experiments have indicated that while the relative affinity of netropsin for 2-amino containing sequences is significantly reduced from poly-dA·dT, the conformation of netropsin is the same in both complexes (Dasgupta *et al.*, 1990). This is consistent with our observation that the observed steric interaction is easily accomodated by the drug molecule without significant conformational change.

Second, efforts at adding a terminal hydrogen bond acceptor to netropsin in the hope of adding G·C selectivity at the end of the netropsin binding site have lead instead to the creation of a molecule that binds as a dimer and demonstrates a high affinity for the sequence 5'-TGACT-3' (Wade *et al.*, 1992). Recent NMR work on this complex (Mrksich *et al.*, 1992) has indicated that the internal 2-amino group of this sequence is hydrogen bonded to the terminal hydrogen bond acceptor group of one of the modified molecules, while the second molecule of the dimer

forms a hydrogen bond to the cytosine carbonyl group in a manner identical to that observed in netropsin:poly-dA·dT complexes. This indicates that 2-amino-pyrrole and 2-amino-methylene steric interactions do not play a significant role in the energetics of complex formation, rather complex formation is dependent on fulfilling the hydrogen bond potential of the guanine amino group.

Based on this theoretical study, we predict that the geometry of netropsin and its homolouges in complex with 2-amino containing sequences, which has not yet been experimentally determined, will be very similar to that of the known netropsin:DNA complexes, exhibiting only a small shift down and out of the minor groove.

Using molecular dynamics to investigate the interaction of the drug netropsin with a variety of DNA binding sites we have demonstrated that the steric interactions seen in experimental observations of the drug complexed with high affinity sites is easily accomodated and does not play a significant role in determining the relative free energy of complex selectivity. We have also demonstrated that relative solvatoin effects play an equally small role. From our work we propose that the selectivity observed for this interaction can be accounted for by the unfavorable electrostatic interactions observed between the 2-amino group of G·C containing sequences and the bound drug molecule. Our observations indicate that this interaction is sufficient to account for the sequence selectivity of complex formation and thus we do not believe that sequence dependent changes in groove width are a significant contributor to selectivity.

Acknowlegements

The authors wish to thank Dr. Murco N. Ringnalda for performing the PS-GVB calculations used in this study. Part of the simulation work was conducted at the Cornell Supercomputer center. We would also like to thank Dr. Steve Breit of Kendall Square Research for help in optimizing our code. KWP was supported in part by a National Science Foundation Graduate Fellowship.

References

Bash, P. A., Chandra Singh, U., Brown, F. K., Langridge, R. and Kollman, P. A., (1987), *Science*, **235**, 574-576

Beveridge, D. L. and DiCapua, F. M., (1989) *Annu. Rev. Biophys. and Biophys. Chem.*, **18**, 431-493

BIOGRAF/POLYGRAF, copyright 1992, (Molecular Simulations, Inc., Pasadena, CA)

Caldwell, J., and Kollman, P., (1986), *Biopolymers*, **25**, 249-266

Caldwell J. W., Agard, D. A., and Kollman, P. A., (1981), *Proteins*, **10**, 140-148

Coll, M., Frederick, C. A., Wang, A. J., and Rich, A., (1987), *Proc. Natl. Acad. Sci. U.S.A.* , **84**, 8385-8389

Dasgupta, D., Howard, F.B., Sasisekharan, V., and Miles, H. T., (1990), *Biopolymers*, **30**, 223-227

Fish, E. L., Lane, M. J., and Vournakis, J. N., (1988), *Biochemistry*, **27**, 6026-6032

Fleischman, S. H., and Brooks, C. L., (1987), *J. Chem. Phys.*, **87**, 3029-3034

Fratini, A. V., Kopka, M. L., Drew, H. R., and Dickerson, R. E., (1982), *J. Biol. Chem.*, **257**, 14686

Fritzsche, H., and Crothers, D. M., (1983), *Studia Biophysica*, **97**, 43-48

Hard, T. and Nilsson, L., (1992), *Nucleosides and Nucleotides*, **11**, 167-173

Jorgensen, W. L., Chandrasekhar, J., Madura, J.L., Impey, R.W., Klein, M., (1983), *J. Chem. Phys.* , **79**, 926

Klevit, R. E., Wemmer, D. E., and Reid, B. R., (1986), *Biochemistry*, **25**, 3296-3303

Kopka, M. L., Yoon, C., Goodsell, D., Pjura, P. and Dickerson, R. E., (1985), *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 1376-1380

Marky, L. A., and Breslauer, K. J., (1987), *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 4359-4363

Mayo, S. L., Olafson, B. D., and Goddard, W. A. III, (1990), *J. Phys. Chem.*, **94**, 8897-8909

Miaskiewicz, K., Osman, R¿, and Weinstein, H., (1993), *J. Am. Chem. Soc.*, **115**, 1526-1537

Mrksich, M., Wade, W. S., Dwyer, T. J., Geirstanger, B. H., Wemmer, D. E., and Dervan, P. B., (1992), *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 7586-7590

Nelson, H. C. M., Finch, J. T., Luisi, B. F., and Klug, A., (1982), *Nature*, **330** , 221

Patel, D. J., (1982), *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 6424-6428

Patel, D. J., and Shapiro, L., (1986), *Biopolymers*, **25**, 707-727

Patel, N., Berglund, H., Nilsson, L., Rigler, R., McLaughlin, L. W., and Graslund, A., (1982), *Eur. J. Biochem.*, **203**, 361-366

Rao, S. N., Singh, C. U., Bash, P. A., and Kollman, P. A., (1987), *Nature*, **328**, 551-554

Ringnalda, M. N., Langlois, J-M., Greeley, B. H., Russo, T. V., Muller, R. P., Marte, B., Won, Y., Donnelly, R. E. Jr., Pollard, W. T., Miller, G. H., Goddard, W. A. III, and Freisner, R. A., (1993), PS-GVB, v0.08, (Schroedinger, Inc., Pasadena CA)

Ryckaert, G. C. and Berendsen, H. J. C., (1977), *J. Comp. Phys.*, **23**, 327

Sun, Y., Spellmeyer, D., Pearlman, D. A., and Kollman, P. A., (1992), *J. Am. Chem. Soc.*, **114**, 6798-6801

Taylor, J. S., Schultz, P. G., and Dervan, P. B., (1984), *Tetrahedron*, **40**, 457-465

Wade, W. S., Mrksich, M., and Dervan, P. B., (1992), *J. Am. Chem. Soc.*, **114**, 7863-7892

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. Jr., and Weiner, P., (1984), *J. Am. Chem. Soc.*, **106**, 765-783

Zimmer, C., and Wahnert, U., (1987), *Prog. Biophys. Mol. Biol.*, **87**, 31-112

Appendix I

Derivation of Perturbation Thermodynamics

The partition function for a given system, A, is given by the definintion:

$$Z_A = \int_0^\infty e^{-H_A\beta} d\tau d\rho$$

where $\beta = 1/\mathrm{kT}$ and H_A is the enthalpy of each conformation of A, and the integeral d τ d ρ is the integral over all positions and momentum of the system. The partition function has the property that:

$$\mathrm{F}_A = \text{-}\mathrm{RTln}\mathrm{Z}_A \ .$$

The relative free energy between two states, A and B, is given by:

$$\mathrm{F}_B \text{ - } \mathrm{F}_A = \text{-}\mathrm{RTln}\left(\frac{Z_B}{Z_A}\right) \ .$$

To relate this to the ensemble average, we start with the definition of Z_B:

$$Z_B = \int_0^\infty e^{-H_B\beta} d\tau d\rho$$

which leads to the equivalent statement:

$$Z_B = \int_0^\infty e^{-H_A\beta} e^{(H_A - H_B)\beta} d\tau d\rho \ .$$

Factoring out the term $\int_0^\infty e^{-H_A\beta} d\tau d\rho$ gives:

$$Z_B = \int_0^\infty e^{-H_A\beta} d\tau d\rho \left(\frac{\int_0^\infty e^{-H_A\beta} e^{(H_A - H_B)\beta} d\tau d\rho}{\int_0^\infty e^{-H_A\beta} d\tau d\rho} \right)$$

which reduces to:

$$Z_B = Z_A \left(\frac{\int_0^\infty e^{-H_A\beta} e^{(H_A - H_B)\beta} d\tau d\rho}{\int_0^\infty e^{-H_A\beta} d\tau d\rho} \right)$$

from which we get an expression for relative free energy:

$$\Delta F = \text{-RTln} \left(\frac{Z_B}{Z_A} \right) = \text{-RTln} \left(\frac{\int_0^\infty e^{-H_A\beta} e^{(H_A - H_B)\beta} d\tau d\rho}{\int_0^\infty e^{-H_A\beta} d\tau d\rho} \right) .$$

This can be simplified by observing that the first exponential in the numerator is a Boltzman weighting factor, that is it is proportional to the probability of state A being in any given conformation. The denominator of the equation is that proportionality constant. Thus, the equation listed above is nothing more than the exponential energy difference between states A and B weighted by the probability of A being in each conformation, which is simply the average of the exponential difference taken over the ensemble of A:

$$\Delta \mathrm{F}_{BA} = \text{-RTln} < e^{(H_B - H_A)\beta} >_A$$

which is the form in which we will utilize this result for our perturbation analysis.

Reference

Zwanzig, R. W., (1956), *J. Chem. Phys.*, **22**, 1420

Appendix II

Water Forcefields

Biochemical reactions are predominantly the reactions of aqueous solutions. Because of the important role water plays as the solvent in so many reactions of interest, a great deal of effort has gone into developing accurate forcefield descriptions of the water potential function.

Most of the important water potentials have been developed using parameters optimized while holding the water bond lengths and angles fixed. This minimizes the optimization to the nonbond descriptors of the molecule:

$$E_{vdW} = \sum_{i} \sum_{j>i} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{C q_i q_j}{r_{ij}}$$

where r_{ij} is the interatomic distance, q_i the charge on atom i, and A_{ij} and B_{ij} the van der Waals parameters that describe the r^6 London dispersion type attraction and the r^{12} electron exchange repulsion term.

One of the first succesful water potentials was the single point charge model (SPC) which represented the water molecule as a collection of three point charges (one at each atom center) and a single set of van der Waals terms centered at the oxygen (Berendsen *et al.*, 1981). This model intitially gained popularity because it reproduced many structural and dynamic properties of bulk water yet the computation of the water-water interaction is particularly simple and fast. It was originally constructed for the simulation of protein hydration and has frequently been used

to simulate the properties of solvated biomeolecules (*e.g.,* Anderson *et al.*, 1986). The potential was developed using a Monte Carlo simulation method to optimize the 4 terms involved (oxygen charge, hydrogen charge, and van der Waals repulsion and attraction terms) to emperically fit the observed thermodynamic properties of water. During the optimization, the H-O-H bond angle was held fixed at 109.47 degrees, and the O-H bond length was fixed at 1.0 Å. No explicit hydrogen bonding potential was used, as the hydrogen bonding properties of water were adequately described by electrostatic dipole interactions. The properties fit for optimization of the parameters included the radial distribution function and the mean potential energy. The best fit was observed with charges of 0.41 and -0.82 (hydrogen and oxygen respectively) and van der Waals terms of 629,400 $Å^{12}$ kcal/mol and 625.5 $Å^{6}$ kcal/mol.

The next significant step beyond SPC was the transferable intermolecular potential model with three points (TIP3P) (Jorgensen *et al.,* 1983). During the same Monte Carlo test of emperical fit bond angles and lengths were fixed at the experimentally determined 104.52 degrees and 0.9752 Å. The derived best fit parameters were oxygen and hydrogen charges of -0.834 and 0.417, and van der Waals terms of 582,000 $Å^{12}$ kcal/mol and 595.0 $Å^{6}$ kcal/mol. This potential has been used for a number of successful simulations of the solvation of biological macromolecules, including carbohydrates (Linert *et al.*, 1992), and protein-DNA complexes (Howard and Kollman, 1992), as well as a variety of thermodynamic analysis listed below. A comparison of results generated with TIP3P and other water potentials is shown in table 1.

Table 1: A comparison of the simulation derived heat capcity, internal energy and heat of vaporization for several water forcefields with experiment (from Jorgensen *et al.,* 1983; Dang and Pettitt, 1987; Jorgensen, 1982; Mills, 1973).

Potential	C_v (kcal/mol·K)	E_i (kcal/mol)	ΔH_{vap}(kcal/mol)
SPC	23.4	-10.18	10.77
TIP3P	16.8	-9.86	10.45
TIP4P	19.3	-10.07	10.66
PFM	18.2	-10.13	na
Expt	18.0	-9.92	10.51

A further refinement of the transferable intramolecular potential was obtained by using a four-center descriptor in which the charge of the oxygen atom was displaced to the molecular center of mass some 0.15 Å from the oxygen (Jorgensen *et al.,* 1983). The same bond lengths and angles used for TIP3P were used for this parameterization, and, as noted above, all bond angles and lengths were fixed at thier equilibrium values. Using a Monte Carlo emperical fit technique, they obtained best fit charges of -1.04 and 0.52, and van der Waals terms (again, cenered on the oxygen)

of 695,000 $Å^{12}$ kcal/mol and 600.0 $Å^6$ kcal/mol. The TIP4P parameter set leads to slightly better radial distribution function for water, but does not significantly improve simulation derived values for heat capacities and heat of vaporization over TIP3P (Jorgensen *et al.*, 1983). It has been sucessfully used to simulate the hydration of oligonucleotides (Subramanian and Beveridge, 1993), pyrroles (Nagy *et al.,* 1993), and ions (Jorgensen *et al.,*, 1989). Unfortunately, the introduction of a center of mass charge term increases the number of nonbond interaction distances that must be calculated, slowing computation time.

For each of the models listed above, the internal degrees of freedom available to the water molecule were eliminated using the shake algorithm (Ryckaert and Berendsen, 1977). To test the accuracy of this the Pettitt flexible modification (Dang and Pettitt, 1987) was developed, which used SPC charges and van der Waals terms but included two bonding parameters (one nonphysical):

$$\frac{1}{2}k_{OH}((r_{OH_1} - r_{OH}^{eq})^2 + (r_{OH_2} - r_{OH}^{eq})^2) + \frac{1}{2}k_{HH}(r_{HH} - r_{HH}^{eq})^2$$

and one angle term:

$$\frac{1}{2}k_{HOH}(\theta_{HOH} - \theta_{eq})^2$$

to describe the internal degrees of freedom of water. These bonding terms were optimized to fit the vibrational spectra of water. An excellent fit was obtained with the bond force constant values:

k_{OH} = 1054.2 kcal/mol/$Å^2$,

k_{HOH} = 79.9 kcal/mol/rad^2

k_{HH} = 79.8 kcal/mol/Å^2.

Simulations using these parameters to describe non-rigid water did not significantly effect the thermodynamic parameters derived from the SPC forcefield, indicating that the rigid molecule approximation is at least as reasonable as this description of water with internal degrees of freedom. A comparison of the calculated internal energies, heat capacities and heat of vaporization derived from these parameters are listed in table 1.

For the solvation simulations reported here, the TIP3P forcefield was used because it exhibits a high degree of accuracy and has been used sucessfully for perturbation thermodynamic analysis of a variety of biological macromolecules (Caldwell *et al.*, 1991; Bash *et al.*, 1987; Rao *et al.*, 1987). In addition to its accuracy, TIP3P is computationally very simple, which makes it possible to conduct extensive simulations from which correspondingly large ensembles can be generated to calculated highly convergent thermodynamic averages.

To test our implimentation of molecular dynamics and the use of this forcefield with spline cutoffs, we have conducted simulations on pure water using our constant temperature molecular dynamics routines and the TIP3P forcefield. These simulations were conducted with 214 TIP3P molecules in an 18.56301 Å cubic box with a spline ramping from 0.0 to 12.0 Å. Simulations were conducted for 30 picoseconds with a 1 femtosecond timestep. From this work, we have obtained a constant volume heat capacity of 16.5 kcal/mol·K (note, for water at 300 K $C_p = 0.995 \cdot C_v$) and an internal energy of -9.73 kcal/mol for TIP3P water, values which are in good

agreement with the experimentally obtained values of 18.0 kcal/mol·K and -9.83 kcal/mol (Mills, 1973).

References

Anderson, A., Carson, M. C., and Hermans, J., (1986), in *Computer Simulation of Chemical and Biochemical Systems*, Beveridge and Jorgensen eds. (New York Acadamy of Science, New York), 51-54

Bash, P. A., Chandra Singh, U., Brown, F. K., Langridge, R., and Kollman, P. A., (1987), *Science*, **235**, 574-576

Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., and Hermans, J., (1981), in *Intermolecular Forces*, B. Pullman ed. (Reidel, Cordrecht, Holland) 331

Caldwell J. W., Agard, D. A., and Kollman, P. A., (1991), *Proteins*, **10**, 140-148

Dang, L. X., and Pettitt, B. M., (1987), *J. Phys. Chem.*, **91**, 3349-3354

Howard, A. E., and Kollman, P. A., (1992), *Protein Sci.*, **1**, 1173-1184

Jorgensen, W. L., (1981), *J. Amer. Chem. Soc.*, **103**, 335

Jorgensen, W. L., (1982), *J. Chem. Phys.*, **77**, 4156

Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, W., and Klein, M., (1983), *J. Chem. Phys.*, **79**, 926-933

Jorgensen, W. L., Blake, J. F., Buckner, J. K., (1989), *Chem. Phys.*, **129**, 193-200

Linert, W., Margl, P., and Renz, F., (1992), *Chem. Phys.*, **161**, 327-338

Nagy, P. I., Durant, G. J., and Smith, D. A., (1993), *J. Am. Chem. Soc.*, **115**, 2912-2922

Mills, J., (1973), *J. Phys. Chem.*, **77**, 685

Rao, S. N., Singh, C. U., Bash, P. A., and Kollman, P. A., (1987), *Nature*, **328**, 551-554

Ryckaert, G. C., and Berendsen, H. J. C., (1977), *J. Comp. Phys.*, **23**, 327

Subramanian, P. S., and Beveridge, D. L., (1993), *J. Theor. Chem.*, **85**, 3-15

Appendix III

An Outline of a Constant Temperature Molecular Dynamics Routine

A) READ .macro file. Input:

1) Coordinate and velocity file name

2) Calculate one energy?

3) Conduct dynamics?

4) Length of dynamics run (picoseconds)

5) Timestep (femtoseconds)

6) Number of steps between outputing coordinate files

7) Coordinate output file root name

8) Nonsolvent atoms?

9) Move atom number

10) Start of nonsolvent atoms

11) End of nonsolvent atoms

12) Number of steps to average heat capacity over

13) Bath temperature

14) Number of steps between temperature scaleing

15) Remarks

16) Parallelization tile size

17) Number of parallel threads

B) READ prt.macro to check for pert therm calculations. Input:

1) Conduct perturbation thermodynamics?

2) Start and end set A

3) Start and end set B

4) Number of steps between sampeling data

5) Number of hybrids

6) Number of samples ignored for equlibration

C) bgfrd:

1) Read biograf format coordinate file (BIOGRAG, 1992)

2) Determine number of atoms

3) atom coordinates

4) atom types-charges

5) PBC parameters

D) check for consistency between .bgf and .macro

E)set up number of bondshake and angle shakes to be considered

F) bondlist:

1) make 1-2, 1-3, and 1-4 bondlists

2) Omit solvent bonds

G) rdpar:

1) Read j.ksr.par parameter file

2) Determine number of each type of parameter

3) Read parameters into parameter arrays

H) aspar:

1) asign parameters to each atom and bond

2) scale nonbond terms

I) set up tile size and thread team for parallel processes

(Kendall Squares Research, 1992)

J) efcall:

1) Zero force arrays

2) call nonbond calculation routines

i) pf: calculates solvent-nonsolvent nonbond interactions

(no bonded interactions here)

ii) p: calculates solvent-solvent nonbond interactions

(excludes bonded atoms)

iii) mm: calculates nonsolvent-nonsolvent nonbond interactions

(includes some bonded atom pairs)

iv) nbcor: subtracts out bonded atom pairs

from nonbond calculations

3) Call Dreiding Valence calculation routines

(Mayo *et al.*, 1990)

i) bndef: calculates nonsolvent bonded interactions

(solvent bonded terms never calculated)

ii) angef: calculates nonsolvent angle interactions

(solvent angle terms never calculated)

iii) toref: calculates torsions (none in solvent)

iv) invef: calculates inversions (none in solvent)

4) Calculate rms force. Print out atoms with forces greater than 100

5) Print energies and rms force

K) Check for dynamics run

L) Check for consistency

M) Print pert therm parameters from prt.macro

N) Dynam:

1)output: Calls internal routines to write
first coordinate and velocity files

2) velrd: Reads initial velocity file

3) fvelscl: scales initial velocities

4) zrocm: Zeros center of mass momentum

5) Enter dynamics loop. Loop from 1 to itr

6) pbndshk: Zeros net forces along bond axis, iteratively

i) for water bonds only

ii) Total force conserved

iii) Convergence at 0.0001 kcal/molÅ

7) pangshk: Zeros net forces in angle plane, iteratively

i) for water angles only

ii) Total force conserved

iii) Convergence at 0.0001 kcal/molÅ

8) strtdyn: calls internal routines

i) s1: updates velocities for step n+0.5 (Verlet, 1967)

ii) s2: calls internal routines

a)pbndvelshk: zero net velocity along bond axis
momentum conserved. For water bonds only
Iterative, convergence at 0.0001 m/sec

b)pangvelshk: zero net velocity in angle bond plane
momentum conserved. For water bonds only

Iterative, convergence at 0.0001 m/sec

c)zrocm. determines and zeros residual cell velocity

iii) ps3: Translate coordinates. Max. displacement defined here

9) ptrnspbc: Translate molecules/atoms back into unit cells

Water molecules translated whole if Oxygen outside of PBC

Non solvent atoms translated independently

10) pbndcrdshk: Removes residual drift in bond length. Iteratively

for waterbonds only

Center of mass position conserved

Convergence at 0.0001 Å

11) pangcrdshk: Removes residual drift in angle. Iteratively

for water angles only

Center of mass position conserved

Convergence at 0.0001 Å

12) velscl: Scales velocities to conform to bath temperature

Called every sclfrq steps

13) efcall: Calls internal energy routines: See above

14) mine: saves coordinates/velocities of lowest energy conformation

15) cprt: calls internal perturbation thermodynamics routines

(Beveridge and DiCapua, 1989)

i) uprt: calculates λ and updates parameters

ii) peprt: calculates energy of set a

iii)swap set a for set b

iv) peprt: calculate energy of set b

v) calculated ΔE and exponetials

vi) print averages

16) spheat: calculates specific heat

17) output: calls internal .bgf and .vel writing routines

Called every "update"th step

18) End of dynamics loop

19) output: calls internal .bgf and .vel writing routines

20) Called to output last configuration

21) mine: swaps loweset energy coordinates and velocities

into c(3,n) and velo(3,n) for outputting

22) output

Called to output lowest energy configuration

23) End of dynamics routine

O) End of program

References

BIOGRAF/POLYGRAF, copyright 1992, (Molecular Simulations, Inc., Pasadena, CA)

Beveridge, D. L. and DiCapua, F. M., (1989), *Annu. Rev. Biophys. and Biophys. Chem.*, **18**, 431-493

Kendall Squares Research, (1992), *Parallel Programing for the KSR*, (Kendall Squares Research, Boston MA)

Mayo, S. L., Olafson, B. D., and Goddard, W. A. III, (1990), *J. Phys. Chem.*, **94**, 8897-8909

Ryckaert, G. C. and Berendsen, H. J. C., (1977), *J. Comp. Phys.*, **23**, 327-340

Verlet, L., (1967), *Phys. Rev.*, **159**, 98

Appendix IV

The Optimization of Molecular Dynamics Routines for Near Linear Speedup on Distributed Memory Parallel Supercomputers

The simulation of biologically relevent molecular species such as proteins, nucleic acids, and small molecules in aqueous solutions is a computationally intensive prospect. The relevant biological macromolecules are very large by simulation standards: a 50 amino acid protein is on the order of 1500 atoms and a single turn of DNA nearly a thousand. Moreover, biochemistry is largely the chemistry of aqueous solutions, so realistic simulations must include explicit solvent molecules to accurately model dynamics and energies. Because of these unfortunate realities, typical simulations of solvated macromolecules must determine the bonding and nonbonding potential for systems of several thousand atoms, a computationally daunting task.

The lion's share of computation time needed to calculate a molecular dynamics simulation of 1000 atoms is expended on calculating the nonbond interactions that occur in the system. Nonbond interactions scale with the square of the number of atoms, thus literally millions on nonbond terms must be considered to determine the energy and forces of a single conformation of this size. In table 1, the number of nonbond terms in the potential function for a variety of systems reported in this work are listed.

Table 1: The number of valence and nonbond interactions for a variety of simulations reported here. Thy(aq) represents a solvated free thymine base, DNA(aq) corresponds to a solvated single turn of helical DNA, and Net:DNA corresponds to the solvated DNA:Netropsin complex. The listed nonbond cutoffs, Min. Im. and Spline 12.0 correspond to full minimal image and a cubic spline with an outer cutoff of 12.0 Å respectively. Nonbond and Bond are the number of nonbond (approximate) and valence (water and netropsin only) interactions considered.

Species	Cutoff	Atoms	Nonbond	Bond
Thy(aq)	Min. Im.	657	215,000	642
Thy(aq)	Spline 12.0	657	63,100	642
DNA(aq)	Min. Im.	5365	11,591,000	4716
DNA(aq)	Spline 12.0	5365	1,242,000	4716
DNA:Net	Min. Im.	5364	11,591,000	4856
DNA:Net	Spline 12.0	5364	1,250,000	4856

There are a variety of techniques for dealing with the complexity of nonbond interations. A traditional approach, called minimal image, ignores the problem and calculates all nonbond interactions in the unit cell, an unfortunately computationally intensive method. Other less computationally daunting routines assume that long range nonbond interactions fall to zero beyond a fixed cutoff (typically 8-9

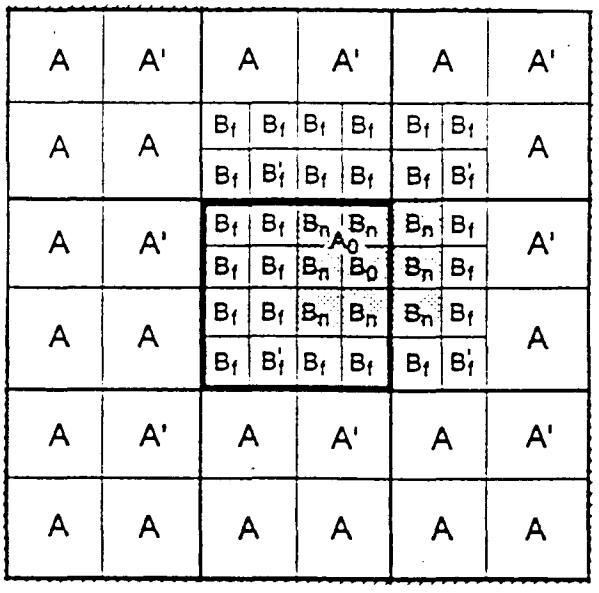


Figure 1: The cell multipole method for calculating nonbond interactions. For atoms in the cell under investigation, interactions with atoms in the nearest neighbor cells are summed in a pairwise fashion. To calculate the contribution of atoms in more distant cells, their contribution to the higher moments of the cell they reside in are calculated (over progressively larger fractions of the total volume) and these monopole, dipole and quadrapolar terms are used to calculate the contribution made to the potential acting on atoms in the first cell (From Ding *et al.*, 1992).

Å and can be ignored. Unfortunately, for highly polar species such as water, this assumption generally does not hold true. A promising order n method of calculating nonbond interactions without cutoffs is the Cell Multipole Method (CMM), which approximates (to a very high degree of accuracy) long range (near neighbor interactions are calculated explicitly) van der Waals and electrostatic interactions by dividing space into a number of discrete cells and summing over these cells to determine monopole, dipole and higher order terms. These moments are then used

to calculate the field generated by the distant region in space at the point under investigation. If space is divided in a geometrically expansive fashion (see figure 1), then this problem scales with n (Ding *et al.*, 1992), leading to significant speedup for the nonbond calculations of systems of more than a few thousand atoms. Unfortunately, while our experience has indicated that the accuracies avalable with this tecnhique are generally on the order of 0.001% in energy and 0.2% in force, CMM tends to lead to poorly convergent averages in perturbation thermodynamic analysis.

To reduce these artificial energy and force fluctuations and the excessive structure (and computational complexity) seen with a minimal image and rigid cutoff approaches, we have used the spline technique in which a scaling of energy (and it's derivative, force) is used to describe the fall off of long range nonbond interactions between two atoms. If the two atoms, i and j, are at a distance, r_{ij}, that is within the spline ramping region (the scaling is often conducted only in a limited shell around the atom), then the energy will be scaled. For a linear spline, we have:

$$E_{ij} = E_{ij}\frac{r_{off}-r_{ij}}{r_{off}-r_{on}} \ .$$

We used the more common cubic spline (BIOGRAF, 1992) which weighs near atom interactions more heavily:

$$E_{ij} = E_{ij}\frac{(r_{off}-r_{ij})^2(r_{off}+2r-3r_{on})}{(r_{off}-r_{on})^3} \ .$$

Note that $E_{ij} = 0$ for $r_{ij} \geq r_{off}$ and energy is not scaled for distances less than r_{on}. Simulation work has indicated that a smooth scaling from r=0 to r=12 Å gives the best fit of phonon dispersion curves in solids (Goddard, 1993) and leads to

excellent agreement with the observed heat capacity in simulations of bulk TIP3P water (see apendix II). The spline cutoff method was used for all of the perturbation thermodynamic simulations reported here.

Another method of speeding up simulations is, of course, to improve the speed of the computer. The perturbation thermodynamics analysis reported here were conducted on a KSR 1/64 multicomputer with a peak through put of 2.4 billion floating point operations per second (Kendall Squares Reasearch, 1992). It achieves this rather astounding speed (estimated to be the 95th fastest computer installed as of 5/93) by coupling 64 very fast processors into a single parallel processor machine by using a unique and very efficient fully distributed memory system.

While the speedups possible with the KSR machine are impressive, they are not trivial to accomplish. The distributed memory architecture forces careful consideration of how a problem is distributed amongst the different processors available. In particular, if two processors try to update a given parameter at the same time (*e.g.*, both try to make an addition to the force on the same atom) then that parameter will not be correctly updated. To avoid this problem, algorithms must carefully be designed so that no two processors ever try to update the same memory location simultaneously.

Each of the many processes that are conducted by a molecular dynamics routine can be optimized for parallel processing on a fully distributed memory supercomputer in its own, unique way. Here we will focus on how we solved the pertinent issuse for the algorithms that calculate the nonbond interactions, as they are the most computationally intesive part of the simulations reported here.

One solution to the problem of updating nonbond forces for each atom in a non-overlapping processor way is to define for each processor a separate force array and sum these arrays at the end of each nonbond calculation. Unfortunately, this method is hampered by enormous amount of interprocessor communication required to update the final sum. We have overcome this problem by breaking nonbond space into non-overlapping "tiles" that iteratively cover all of nonbond space, while during any given iteration, each processor is calculating the forces on a unique subset of all nonbond space. A schematic diagram of how this division is produced is illustrated in figure 2.

A variety of other routines were also parallelized. The calculation of movable-fixed atom interactions were sped up by the fact that the the forces on fixed atoms are not calculated, so each processor can look at the interactions of the same set of fixed atoms with a different subset of the movable atoms simultaneously. To do so, this algorithm was parallelized by simply dividing the movable atoms over the number of processors available.

After efficient parallelization of the nonbond routines is inacted, the shake and Newtonian translation routines become significant in total computational throughput. The shake algorithm can easily be parallelized by dividing an equal number of water molecules to each processor, though the convergence criteria for this iterative process requires that once per iteration the total residual error be calculated by summing over all cells. Because translations are unique to each atom or molecule, the dynamics translation algorithms, including PBC translations back into the unit cell, can be easily parallelized.

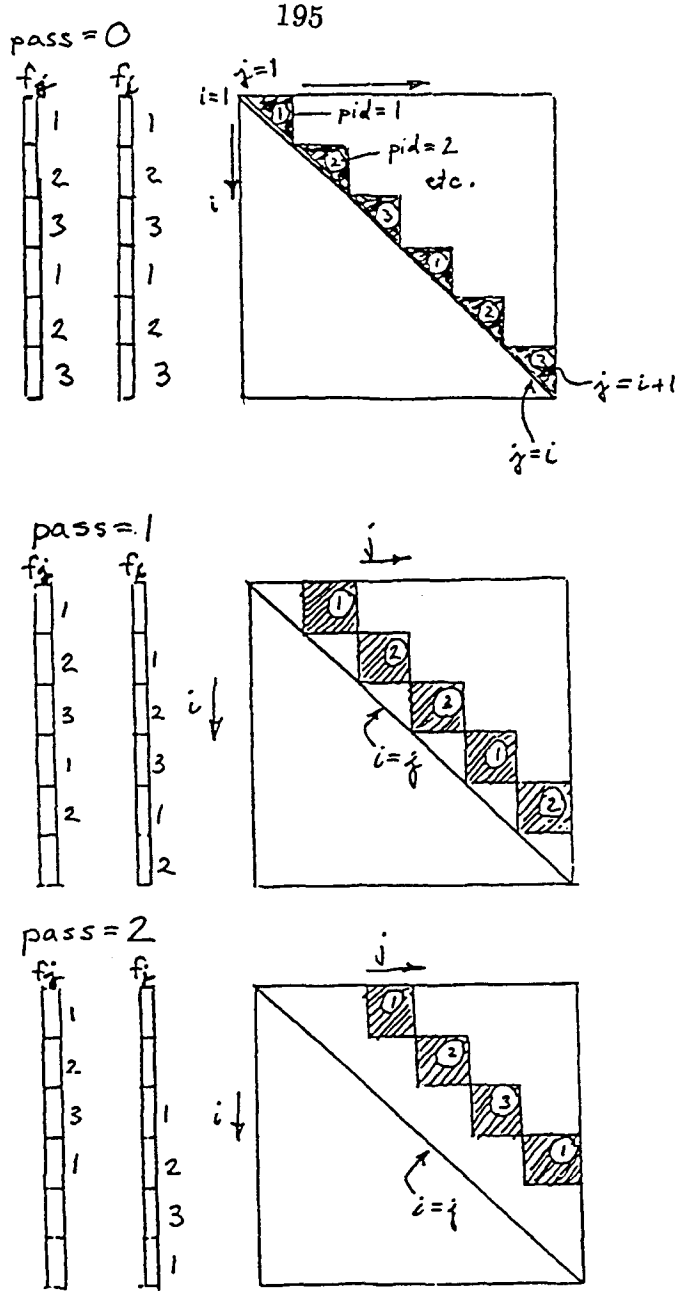


Figure 2: The "tiling" pattern used to iteratively cover nonbond space without allowing any two processors to update the force array element for a given atom at the same time. Since the force on each atom is modified up to two times in each iteration (once on the j axis and once on the i axis) only two dummy arrays are needed, and the final summation of all the dummy arrays to determine total forces is rapid. For this example, 3 processors (pid=1,2,3) are being used, with a tile size of 1/6 the total number of atoms.

Of course, parallelization of an algorithm that runs with poor efficiency on one processor will run with poor efficiency over many; parallelization is not a cure for badly written code. Issues in nonparallel optimization that were important for the work reported here include the significant speed up that was obtained by machine specific modifications of the code. For example, calculation of electrostatic terms requires the calculation of the reciprocal of the interatomic distances:

$$1/r_{ij} = \frac{1}{\sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}} .$$

For the KSR, the calculation of $\frac{1}{\sqrt{r^2}}$ is approximately 50% faster than the calculation of $\sqrt{r^2}$ followed by the determination of 1/r. This simple change leads to a typical speed up of 20% for nonbond calculations on the larger systems reported here. The inclusion of this and other machine specific optimization of the code for the KSR was responsible for a 45% speed up over code optimized for other machines.

For the most computationally intense systems reported here (a solvated turn of double helical DNA) parallelization of the six algorithms listed above (corresponding to 1000 lines out of 5000 total lines of code) is sufficient to achieve a very high level of parallelization efficiency. Figure 3 demonstrates the speed up observed over 1 to 30 cells. Over 30 cells, an effeciency of 66% was obtained.

The total cpu throughput over 30 cells is estimated to be approximately 300 million floating point operations/second (mflops). A complete comparison of the simulation speeds observed on the variety of computers used in this work is shown in table 2. For the test case, which consists of simple dynamics on a single turn of DNA with 1572 water molecules in periodic boundary conditions times ranged from

Table 2: Time per iteration for a dynamics simulation of a hydrated single turn of DNA with 1572 water molecules in full Minimal Image PBC. Machine, clock, and processors represent the machine name and the number of processors used in the simulation. Time represents the cpu time for load sharing machines and clock time for unshared processors, in seconds. The results reported were the average computation time per timestep for a 1.0 picosecond simulation with a 1 femtosecond timestep, rigid waters, and constant temperature scaling.

Machine	Processors	time (seconds)
SGI 4D/380	1	92
HP Apollo 9000	1	35
KSR1/64	1	48
KSR1/64	25	2.5
KSR1/64	50	1.9

48 to 1.9 seconds per iteration. To put these figures in perspective, the simulation of a single turn of hydrated DNA for 100 picoseconds (100,000 steps of dynamics) requires approximately 2.5 days over 30 cells on the KSR, which corresponds to 2.7 months on an SGI 4D/380.

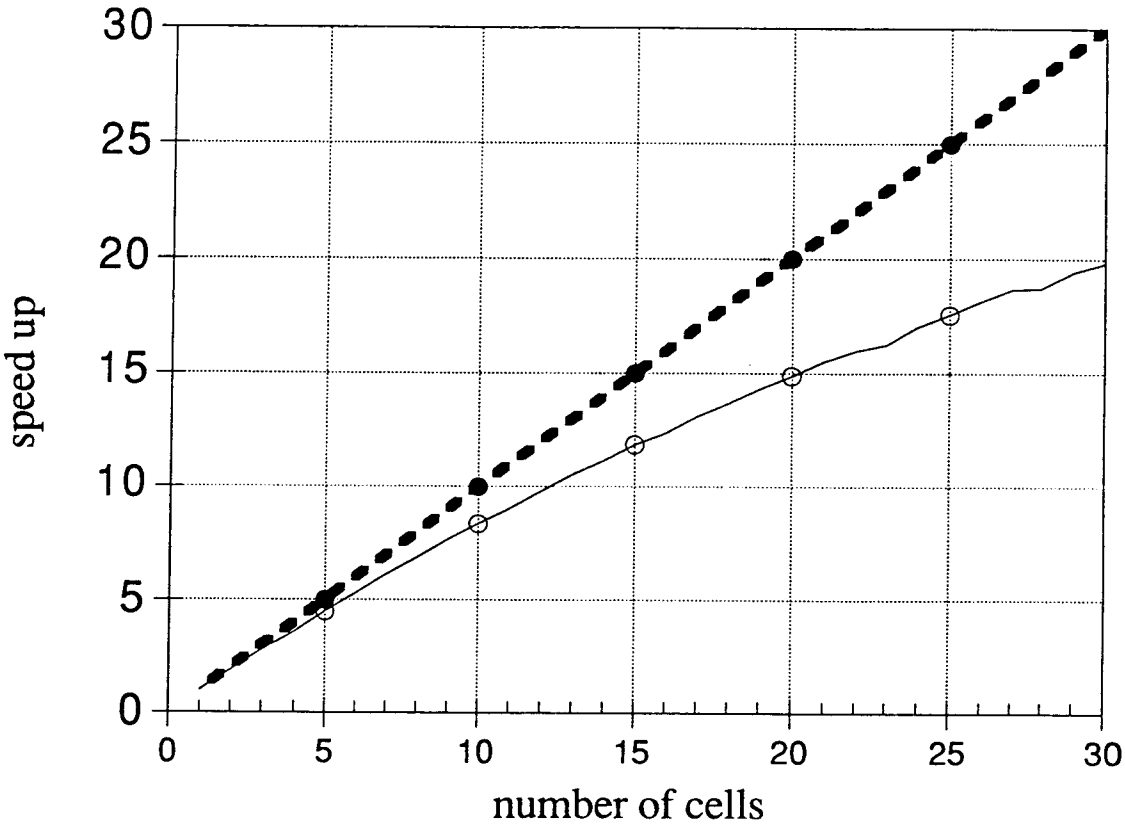


Figure 3: The observed speed up for a simulation of 1572 water molecules and a single turn of helical DNA in periodic boundary conditions and with full minimal image cutoffs. The dashed line represents the theoretical maximum speedup. The observed speedup is 66% of this maximum over 30 nodes of a KSR1/64 supercomputer.

References

BIOGRAF/POLYGRAF (1992) copyright Molecular Simulations, Inc. (Pasadena CA)

Ding, H. Q., Karasawa, N., and Goddard, W. A., III, (1992), *J. Chem. Phys.*, **97**, 4309-4315

Goddard, W. A., III, (1993), *M. S. C. Technical Note*, **119**, (Materials Simulation Center, California Institute of Technology, Pasadena CA)

Kendall Square Research, (1992), *KSR Parallel Programers Guide*, (Kendall Squares Research, Boston MA)