# Development and Validation of Statistical Potential Functions for the Prediction of Protein/Nucleic-Acid Interactions from Structure

Timothy Allen Robertson

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2007

Program Authorized to Offer Degree:
Department of Biochemistry

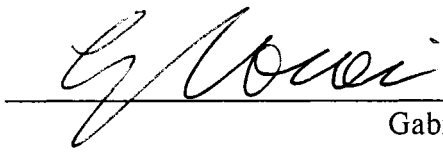
University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Timothy Allen Robertson

and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the final examining committee have been made.

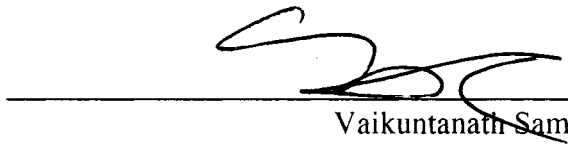
Chair of the Supervisory Committee:



Gabriele Varani

Reading Committee:

Gabriele Varani

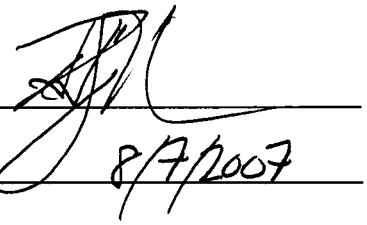


Vaikuntanath Samudrala

Stephen Hauschka

Date: July 30, 2007

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature 

Date 8/7/2007

University of Washington

## Abstract

Development and Validation of Statistical Potential Functions for the Prediction of Protein/Nucleic-Acid Interactions from Structure

Timothy Allen Robertson

Chair of the Supervisory Committee:
Professor Gabriele Varani
Department of Biochemistry

This work outlines the development, validation and application of a series of novel statistical (knowledge-based) potential functions to the prediction of protein/nucleic-acid interactions from structure. Three methods are described: a statistical potential for the evaluation of inter-molecular hydrogen bonds at protein/nucleic-acid interfaces; an all-atom, distance-dependent statistical potential for protein-DNA interactions, based upon the naïve Bayes classifier formalism; and a similar approach, specific to the structural properties of protein-RNA interactions. These three methods are shown to be able to reliably discriminate non-native and near-native structures from native protein/nucleic-acid complexes, and are successfully demonstrated in applications to computational molecule/molecule docking (the prediction of molecular interactions from structure), rational (structure-based) protein design, and the recapitulation of experimentally determined binding energies for mutations to protein/nucleic-acid complexes. Despite their simplicity, these statistical techniques are found to be sensitive to subtle structural and chemical changes at protein/nucleic-acid interfaces, and in several cases, are demonstrated to possess performance characteristics on par with significantly more complicated, physics-based methods. These results suggest that simple, statistical potential functions can serve as a generally useful tool for the computational prediction, design and simulation of protein interactions with nucleic-acid molecules.

## TABLE OF CONTENTS

## LIST OF FIGURES

Figure Number Page

## LIST OF TABLES

## PREFACE

This work comprises four separate documents, of which two have been published previously (in slightly modified forms), and two are submitted or under review for future publication. Section 2 was co-authored by Dr. Yu Chen, Dr. Tanja Kortemme, Dr. David Baker, Dr. Gabriele Varani and myself, and published in a 2004 issue of *Nucleic Acids Research* (1). Section 3 was co-authored by Dr. Gabriele Varani and myself, and was published in a 2007 issue of *Proteins: Structure, Function and Bioinformatics* (2). Section 4 was co-authored by Dr. Suxin Zheng, Dr. Gabriele Varani and myself, while Section 1 was co-authored by Dr. Gabriele Varani and myself, and will be published in 2008 as a chapter in the second edition of the book Structural Bioinformatics (Wiley; J. Gu and P. Bourne, editors). I wish to acknowledge these individuals for their substantial contributions to this work.

## ACKNOWLEDGEMENTS

As much as I would like to pretend that this dissertation represents a completely personal achievement, the reality is that no work of any significance can be accomplished without the support of a community of dedicated, caring people. This manuscript is the culmination of nearly a decade of work, and over that time, I have been the recipient of far too many personal and professional graces from my friends and colleagues to believe that a few pages of acknowledgement will adequately recognize their contributions; I can only hope that they understand the depth of my gratitude for years of guidance, collaboration and friendship.

In particular, I would like to thank the members of my advisory committee: Professors Bruce Clurman, Valerie Daggett, Phil Green, Steve Hauschka, Ram Samudrala and Gabriele Varani. I don't believe that the quality of the food that I provided made up for the marathon length of the meetings that we held, but I know that their patience and guidance have been invaluable to me over the years. Without their considerable insight, I might still be searching for an interesting question to study.

All of the members of the Varani laboratory have been like an extended family to me, but a few have been like siblings: Carsten Detering taught me the essential value of finding joy in little things (such as a good birthday cake) when other forms of happiness are not forthcoming; his unflagging optimism and good nature is an inspiration. Kate Godin was a willing ear and confidant for more gloomy graduate-school stories than anyone has a right to hear, and I wish that I possessed her powers of empathy and caring, so that I might be able to return the favor. Julia DeBaecke was always available for a night out, and her reminders of the benefits of the social world provided me with a much-needed prod away from lonely evenings in front of the computer. Finally, Priti Deka is the little sister that I never had; she has been a constant source of humor and happiness, on even the most difficult days.

Outside of the laboratory, it is no exaggeration to say that I would never have finished this process without the friendship of Chris Saunders. From practically the first day of graduate school, we have commiserated over everything from bugs in code to bad

landlords, and I consider myself exceptionally lucky to have found such a patient, supportive friend. Likewise, I owe Phil Danielson a debt of gratitude; he encouraged me to begin this process, and his friendship and advice have helped to ensure that I finished what I started.

Professionally, Jim Havranek, Carlos Duarte and David Baker have graciously provided access to their own research data, making several of the experiments in this work possible. Carol Rohl, Bill Wedemeyer, Brian Kuhlman and Jerry Tsai were exceptional resources when I was a beginner in the field, and later on, Jeff Gray and Kira Misura were invaluable for their good advice, both professional and personal. Tanja Kortemme was a role model and a mentor, and I will never be able to repay her for the countless hours we spent discussing everything from research results to academic politics. Finally, Suxin Zheng and Yu Chen were dedicated collaborators, and much of the work in this dissertation represents their own tireless efforts and skill.

Of course, absolutely nothing in this work would have been possible without the support and dedication of my advisor, Professor Gabriele Varani. Gab took a risk when he accepted me into his laboratory – I was an unproven student, in a new field of research for his group – and I will forever be grateful for his generosity and kindness. I hope that he is as proud of the work that we have accomplished together, as I am proud of what I have I learned from his instruction.

Finally, I would like to thank my family for their patience and understanding over the years. On far too many occasions, they have accepted that the lifestyle of a graduate student means that branches of the family tree are allowed to wither from neglect, yet they have consistently and enthusiastically supported my goals. Their encouragement and caring have made this difficult process bearable.

# 1 An Overview of the Structure-Based Analysis of Protein/Nucleic-Acid Interactions

The specific recognition of nucleic acid sequences by nucleic acid binding proteins is of critical importance to the biological function of every living species. As a result, the phenomena responsible for this recognition process have long been of interest to biological scientists. Beginning with Seeman *et al.*, research into sequence-specific DNA recognition focused on the search for a "recognition code" — a collection of simple rules that would pair particular amino acids to specific bases (3). However, it was soon realized that any recognition code would be degenerate (4), and a general "code" may not exist at all (5, 6).

The structural mechanisms underlying sequence-specific DNA recognition received renewed attention with the increased availability of high-resolution structures for protein-nucleic acid complexes. Computational studies of these structures have classified their interactions (7), described features of their binding sites (8-10), and the evolutionary conservation of their interface residues (11, 12). However, relatively little effort has been devoted to the application of this structural knowledge to computational models for the prediction of protein-nucleic acid interactions. In contrast, structural information has been used extensively to create potentials for prediction of protein structures (13-19), as well as protein-ligand (20-23), and protein-protein interactions (19, 22, 24, 25).

## 1.1 Motivations and Applications

Though it is obvious that the ability to accurately and reliably predict the sequence-specificity of nucleic-acid-binding proteins from their structures would represent a great intellectual accomplishment, any discussion of the problem is sure to encounter skepticism: what is the practical application of such a method? For example, the structure of a given protein-nucleic-acid complex represents a great deal of time and effort; it is almost certain that structurally characterized complexes have also been experimentally characterized, and that many of their preferred DNA (or RNA) binding

sites are known. Indeed, the development and testing of the methods described here have crucially depended on this fact – the performance of new methods is frequently evaluated by examining their ability to reproduce the known sequence specificities of previously solved protein-nucleic-acid structures (2, 26-30).

Given this view of the problem, the value of any structure-based method must extend from its generality: if a model can predict the preferred DNA or RNA binding sequences of a known structure, it is interesting, but not especially useful; if it can predict the binding sequences of homologous (though structurally uncharacterized) proteins, it has more value; if it can be used to develop entirely novel nucleic-acid-binding proteins (or to otherwise predict the energetic consequences of changes to a complex), it will almost certainly find wide, practical application in future research. These are analogous to the roles of structure-based models in other areas of biology: the prediction of preferred DNA/RNA-binding sequences for a protein has parallels to the problem of structure-based drug design; the prediction of nucleic-acid recognition sequences for homologous protein sequences is an obvious application of protein homology modeling; and the development of novel nucleic-acid-binding proteins is just a special case of the rational protein design problem. Even the obvious application – the structure-based identification of DNA/RNA binding sites for known protein-nucleic-acid complexes – can be useful, because many of the "known" recognition sequences of nucleic-acid-binding proteins are not absolute, or perfectly characterized (31, 32). Structure-based models may one day provide crucial insights into the range and flexibility of recognition sequences for proteins whose nucleic-acid-binding properties are not currently well-known.

## 1.2 Potential Functions for Protein/Nucleic-Acid Interactions

### 1.2.1 Physical Potentials

Because it is now routine to computationally simulate the dynamics of proteins, nucleic acids and other biopolymers, it should not be surprising that many attempts to predict the energies and specificity of protein-nucleic-acid interactions are extensions of the same techniques. In particular, the molecular dynamics (MD) simulation of nucleic-

acid molecules and protein-nucleic-acid complexes was routine long before the first efforts to computationally predict the DNA/RNA recognition sequences of proteins from structure (33-41), and MD force fields were amongst the first potential functions to be applied to the problem of predicting the energetic properties of sequence-specific, protein-nucleic-acid interactions (42-47).

The prediction of nucleic-acid-binding sequence specificity is a different, larger problem than predicting the dynamics or binding energy of a single protein-nucleic-acid complex. In one of the earliest attempts to apply a molecular dynamics force field to this problem, Pichierri *et al.* used the AMBER package to derive free energy, enthalpy and entropy "maps" of base-amino-acid interactions (48), a study followed by similar efforts from other groups (49, 50). By sampling protein side-chain conformations at grid points surrounding canonical nucleotide structures, these efforts were able to demonstrate energetically favorable conformations for the interaction of particular amino-acid/base-pair interactions. Somewhat later, Thayer and Beveridge applied their group's ongoing MD simulation research into sequence-dependent DNA deformation to the prediction of binding sites for the catabolite activator protein (CAP) (51). They demonstrated a hybrid approach, wherein a hidden Markov model (HMM) was trained using both binding-sequence data and nucleotide roll/tilt data obtained from MD simulation of DNA molecules, and found that the structural information tended to improve the quality of binding-site predictions. Thus, both direct and indirect recognition mechanisms have been explored using MD potentials.

Ultimately, molecular dynamics simulations are not well-suited to the prediction of cognate binding sequences for nucleic-acid binding proteins. There is no MD analogue to the process of computational base "mutation", and DNA/RNA sequence changes must be approximated through time-consuming simulations of the bound and unbound forms of the sequence variants (*i.e.* simulation of the thermodynamic cycle for every possible combination of nucleotide mutation); this process is computationally prohibitive for all but the smallest protein-nucleic acid complexes. As such, many applications of physics-based potentials to the problem of binding-site scanning have used molecular mechanical

techniques, wherein the intent is not to follow conformational changes over time, but to directly compute the free energies of interaction between protein and nucleic-acid molecules. For example, Paillard and Lavery used an energy minimization strategy, based on the AMBER force field, to predict the binding free energies and optimal binding sites for a set of 18 DNA-binding proteins (52). In the process, they demonstrated that the recognition of DNA sequences by proteins depends variably (*i.e.* in a complex-dependent manner) on both protein-DNA interactions, as well as the sequence-specific energy of DNA deformation.

Although the AMBER force field is quite commonly applied to the simulation of protein-nucleic-acid interactions, the CHARMM package has also been successfully used, particularly for MD simulations of the sequence-dependent flexibility of protein-bound DNA molecules (*i.e.* "indirect recognition" phenomena); given the overall similarity of these methods, their common application to the problem is not surprising (38, 39, 41, 43, 53). However, in an example of a hybrid physical/statistical potential function, Havranek *et al.* have extended the ROSETTA protein design potential to protein-DNA systems (28), while Chen *et al.* applied the method to the prediction of protein-RNA interactions (1). The ROSETTA potential function incorporates many terms common to physics-based potentials (*e.g.* the Lennard-Jones model of the Van der Waals force), but also makes use of a unique statistical model of hydrogen bonding geometry that appears to confer greater sensitivity to the method (54).

### 1.2.2 Statistical Potentials

In contrast to these examples of complex, physics-inspired potential functions, some of the earliest efforts to predict sequence-specific protein-nucleic-acid interactions from structure involved the use of simple, knowledge-based (a.k.a. "statistical") potentials, which make use of the database of known structures to derive probability-based scores that can be used to predict protein-nucleic-acid interaction energies. While there is a great deal of variation between methods in this category, all share a common theme: the database of *known* protein-nucleic-acid structures is assumed to adequately represent the

likely distributions of particular "features" that can be observed in *any* real biological structure; examples of such "features" include inter-atomic distances, torsion angles, bond lengths, and the spatial distributions of atoms and residues.

From a biophysical perspective, knowledge-based potential functions are rooted in the assumption that individual atomic structures represent low-energy molecular conformations, reflecting the optimal, real-world contributions of many different microscopic forces. If this assumption holds, then a sufficiently large (*i.e.* infinitely large) database of randomly sampled structures would capture the physically realistic range of any particular structural feature. Moreover, these features would be expected to occur in direct proportion to their energies, with high energy features observed far less frequently than low-energy features. The basic concept is straightforward: the more a given structure "resembles" the database of known structures (which are presumably correct), the "better" that structure is likely to be (55-57).

Broadly speaking, most statistical potential functions follow a simple formula: the structural feature of interest is quantified, this measurement is divided into bins (creating a histogram), and the "training set" (*i.e.* structures chosen from the Protein Data Bank specifically to represent the class of molecules being scored) is examined to see how the feature of interest is distributed. For example, if intermolecular atomic distances are being measured across a protein-DNA interface, it might make sense to divide the continuous range of realistic values (*e.g.* from 3Å to 10Å) into 10 bins with 1Å widths. If a diverse training set of protein-DNA structures were then examined to count the number of intermolecular contacts that fell within these bins, a histogram would result, and a simple score for an atom-atom pair could be created by taking the logarithm of the likelihood of each distance bin:

$$s(i,j,d_{ij}) = -\log\left(\frac{f_{observed}(i,j,d_{ij})}{f_{expected}(i,j,d_{ij})}\right) \qquad (1.1)$$

Here, i and j represent atom types on opposite sides of the protein-DNA interface, $d_{ij}$ is the distance between them, and the numerator is the observed frequency of pairs between atom types i and j (separated by distance $d_{ij}$) in the training set. The denominator, in

contrast, represents the ideal (expected) frequency for this same value, and is commonly referred to as the "reference" state for the potential function.

Though the reference state greatly impacts score performance (55), its choice is somewhat arbitrary, and it is often impossible to know what this function should be *a priori*. A naïve implementation might assume that all distances are equally likely for any given atom pair (*i.e.*, that the reference state is a uniform distribution), whereas a more realistic model might assume that the distribution of distances for any particular atom pair is reasonably well-approximated by the overall distribution of distances for *all* atom-atom pairs. The choice is usually justified empirically, and as a result, a large research literature has focused on the reference state, investigating its impact on score performance (2, 13, 14, 18-20, 22, 23, 30, 58).

Again, this is only a simple illustration of statistical potentials. In practice, potentials for protein-nucleic-acid interactions are more complicated, and can be broadly grouped into two categories, based on the structural features that they consider: orientation-dependent potentials (which exploit three-dimensional spatial and angular distributions), and distance-dependent potentials (which use one-dimensional data, such as inter-atomic and residue-residue distances). There are a small number of methods that do not fit cleanly into either category; these methods have thus far been targeted to specific classes of problems, such as the prediction of sequence targets for particular families of nucleic-acid binding domains.

#### 1.2.2.1 Orientation-Dependent Statistical Potentials

Perhaps the earliest example of the application of a knowledge-based potential to protein-nucleic-acid interactions is the work of Kono and Sarai, who used the geometric regularity of the DNA double helix to create local "reference frames" for the nucleotides, which were then used to count the number of amino-acid alpha carbons in three-dimensional spatial bins surrounding the nucleotide bases (26). By converting these counts to frequencies, and using these frequencies to estimate the likelihood of new protein-nucleic-acid complex structures (generated by computationally "threading" new

base sequences onto existing complexes), they were able to successfully predict the cognate DNA recognition sequences of a number of transcription factors. The success of their work is largely due to the non-uniform distribution of amino acids about the DNA bases (e.g. Lysine contacts to Guanine tend to be clustered in the major groove of B-form DNA helices) (9), though the power of the method is striking, given its simplicity.

More recently, Chen *et al.* developed an orientation-dependent, hydrogen-bonding function for protein-RNA interactions, and demonstrated its utility for predicting the amino acid sequences of RNA-binding proteins (1). Unlike Kono and Sarai, this work used a complex, multi-term potential function that involved a mixture of statistical and biophysical terms, but clearly demonstrated that a database of protein-nucleic-acid structures could be used to derive a statistical potential that measures the quality of hydrogen bonds across the protein-RNA interface. This was achieved by observing four structural features of hydrogen bonds (the donor-acceptor bond length, the bond angles at the donor and the acceptor atoms, and the planarity of the bond), and using the structural database to find the angular and distance distributions of these features (54). As in Kono and Sarai, these distributions were converted to frequencies, and used to infer the likelihood of new protein-RNA structures (in this case, generated by substituting new amino acids onto the backbones of known RNA-binding proteins), in the first demonstration of a method for the computational design of RNA-binding proteins.

#### 1.2.2.2 Distance-Dependent Statistical Potentials

Both of the previous examples used three-dimensional features of the protein-nucleic-acid interface to develop potential functions. More recently, several groups have demonstrated that even simple, one-dimensional data can be sufficient to identify native-like protein-DNA and protein-RNA structures. In particular, the inter-residue and inter-atomic distances observed in the database of protein-nucleic-acid structures appears to contain sufficient information to identify cognate DNA and RNA recognition sequences, to discriminate native complex structures from sets of near-native "decoys", and even to estimate the experimentally-determined energetics of protein-nucleic-acid binding.

Despite their simplicity, these "distance-dependent" methods are proving as powerful as the far more complicated, physically-inspired potentials discussed previously, and represent a rapidly growing area of research in the field.

A simple example of this distance-dependent approach was provided by Liu *et al.*, who developed a potential function based on the observed separations between the beta carbons of amino acid residues and the geometric centers of nucleotide triplets (as well as doublets and single nucleotides) in protein-DNA structures (59). Liu *et al.* showed that their approach could successfully predict the binding energies and cognate recognition sequences of a large number of DNA-binding proteins, and theorized that the use of nucleotide triplets allowed for the capture of higher-order interactions, despite the use of one-dimensional feature data.

In a more complex example, Zhang *et al.* developed a statistical potential function based on the observed inter-molecular atomic distances in protein-DNA structures (22). This method mapped all of the protein and DNA atoms in protein-DNA complex structures to 19 chemically derived atom types, and determined a unique inter-molecular distance potential for each of the 361 ($=19^2$) possible atom type pairs. Despite training this method on a set of protein-ligand complexes (instead of protein-nucleic-acid complexes), Zhang *et al.* demonstrated that their potential produced scores that correlate well with the experimentally determined dissociation ($K_D$) constants for a collection of 45 protein-DNA complexes.

Most recently, two groups have independently demonstrated that all-atom, distance dependent statistical potentials for protein-DNA interactions can achieve great predictive power. Both potentials used only the inter-molecular distances at the protein-nucleic-acid interface to achieve remarkably accurate predictions: Donald *et al.* showed that a quasi-chemical potential (wherein the reference state is determined by the relative frequencies of the different atom types) can accurately predict cognate DNA recognition sequences, and $\Delta\Delta G$ values for a large number of experimentally characterized protein-DNA complexes (30); Robertson *et al.* demonstrated that an all-atom, distance-dependent potential (wherein a unique distance-dependent score was defined for each of nearly

14,000 inter-molecular, atom-type pairs) can achieve cognate binding-sequence discrimination performance on par with physical potential functions (2).

#### 1.2.2.3 Other Statistical Models

In addition to the orientation- and distance-dependent potentials discussed above, a few groups have explored the use of methods that don't fit cleanly into either category. For example, Ge *et al.* developed a knowledge-based method for predicting the sequence-specific binding energies of polyamide molecules to DNA double helices, based on the observed positions of water molecules and amino acid atoms in known protein-DNA structures (60). Their method clustered the atoms (either water or amino acid) observed to hydrogen-bond to DNA bases in their structure training set, and used these structural clusters to determine 3D ellipsoid regions of likely drug-DNA hydrogen-bonding interactions.

Finally, in an example of a prediction method targeted to a specific family of DNA-binding proteins, Kaplan *et al.* developed a knowledge-based potential for $Cys_2His_2$ zinc-finger proteins, and used it to scan for transcription factor binding sites in the Drosophila genome (61). The method is neither general, nor is it a true structure-based potential (it does not make direct use of structure data), but because it exploits the well-understood regularity of the $Cys_2His_2$ family of zinc-finger proteins to construct a probability model of DNA recognition, it nonetheless represents an interesting application of structural knowledge to the prediction of transcription-factor binding sites.

## 1.3 Incorporating Protein/Nucleic-Acid Interface Flexibility

An important, largely unsolved problem inherent to the structure-based prediction of protein-nucleic-acid interactions lies in the treatment of molecular flexibility. Although some simulation methods (such as molecular dynamics or gradient-based minimization techniques) allow for the representation of small motions at the protein-nucleic-acid interface, there is good reason to believe that the molecular movements involved in nucleic-acid sequence recognition are larger than can be accurately modeled using these

methods alone (62). For this reason, several groups have applied methods incorporating crude protein flexibility models to the problem of protein-DNA and protein-RNA interface prediction. For example, Endres *et al.* used the AMBER force field, in combination with a dead-end-elimination (DEE) algorithm for protein side-chain packing, and demonstrated that this method could be used to predict the consensus binding sequence of the Zif268 zinc-finger transcription factor (27). However, their results also suggested that the use of rotamer packing may have negatively impacted the performance of their method, leaving in question the value of the flexible model.

Meanwhile, when Havranek *et al.* applied the ROSETTA protein-design algorithm (63) to protein-DNA complexes, they demonstrated that their Monte-Carlo-based approach to side-chain rotamer packing could be used to predict native-like protein sequences from the structures of protein-DNA complexes (28) (indeed, they later successfully used the method to guide a limited re-design of the I-MsoI homing endonuclease protein (64)). But when Morozov *et al.* used the same approach to predict $\Delta\Delta G$ values for a large set of experimentally characterized mutations to protein-DNA complexes, they demonstrated that the performance of the method was actually negatively impacted by the introduction of protein flexibility – when side-chain packing was enabled, the ROSETTA potential did no better than a control method, which simply counted the number of intermolecular contacts observed at the protein-DNA interface; an attempt to incorporate a limited model of DNA flexibility into their software was found to further degrade the performance of the approach (29).

To date, no group has successfully demonstrated a method that improves the prediction of sequence-specific, protein-nucleic-acid interactions through the incorporation of molecular flexibility. Certainly, some techniques (such as computational protein design) are made possible through the judicious application of algorithms for the prediction of protein side-chain conformations, but even these relatively conservative methods can achieve only limited accuracy (65). Thus, the modeling of molecular flexibility represents a fertile area of future research, if our current potentials are to be reliably applied to computationally intensive problems, such as the structure-based

annotation of genome sequence, or the rational design of nucleic-acid binding proteins with entirely novel sequence specificities.

## 1.4 Current Applications

### 1.4.1 Molecular Docking

The structure-based prediction of protein-nucleic-acid interactions is a young field, and there are still relatively few examples of practical applications – particularly those with large-scale, experimentally-validated results. Nevertheless, research in the field is growing rapidly, and existing publications already offer promise that the structure-based analysis and prediction of protein-nucleic-acid interactions will become a valuable tool for biologists seeking new sources of insight and guidance for experimental design.

One straightforward application of structural analysis that has been successfully demonstrated is the prediction of protein-DNA and protein-RNA interactions through the computational docking of protein and nucleic acid structures. Computational docking is a well-established technique in the study of protein-protein and protein-small-molecule interactions, and because it is relatively easy to conduct rigid-body binding simulations of molecular structures, it was one of the earliest tests applied to protein-nucleic-acid systems. In particular, nearly a decade ago, Knegtel *et al.* applied their Monte-Carlo-based docking method (MONTY) to the prediction of protein-DNA interactions. Their research incorporated both protein side-chain (66) and DNA flexibility (67), but tested their methods on only a few protein-DNA complexes. Later, Aloy *et al.* used the innovative fast global search algorithm of FTDock to predict protein-DNA interactions for a larger number of rigid-body structures (68), and subsequently extended the method to incorporate protein side-chain flexibility at the protein-DNA interface (69). Most recently, Robertson and Varani used the FTDock rigid-body method to validate their knowledge-based potential for protein-DNA interactions, and showed that their score was able to significantly improve upon the FTDock results (2). Chen *et al.* also achieved good docking decoy discrimination performance for protein-RNA complexes using the ROSETTA docking method (Monte-Carlo-based, rigid-body docking, coupled with

protein side-chain flexibility (70)), in combination with their statistical hydrogen bonding potential function (1). Finally, in one of the most advanced examples of protein-nucleic-acid docking to date, Van Djik *et al.* used their HADDOCK method to conduct fully flexible docking simulations of three protein-DNA complexes (71). By alternating rounds of rigid-body docking with computationally-intensive MD simulations of the best-scoring docking decoys, they were able to formulate reasonably accurate models of protein-DNA complexes starting from structures of their unbound components.

Despite these early successes, there are still practical limitations to the prediction of protein-nucleic-acid interactions using docking simulations. In particular, only a few attempts have been made to dock unbound structures of protein and DNA molecules, with mixed results; again, our understanding of molecular motion limits our ability to accurately model protein-nucleic-acid interactions. The flexible nature of DNA and RNA molecules, the prevalence of induced fit, and the importance of indirect recognition in protein-nucleic-acid binding mean that any generally useful docking method for protein-DNA or protein-RNA interactions will have to incorporate a robust model of molecular flexibility.

### 1.4.2 Structure Analysis and Refinement

Structural biologists have long been using computational analyses of protein-DNA and protein-RNA structures to develop and explore hypotheses about the mechanisms of protein-nucleic-acid recognition that are difficult (or impossible) to address experimentally. For example, a number of papers have explored the use of computational models to quantitatively investigate the importance of direct and indirect recognition in sequence-specific protein-nucleic-acid interactions. Direct recognition (recognition due to atomic interactions across the protein-nucleic-acid interface) can in theory be distinguished from indirect readout (recognition of larger structural features of the nucleic acid) by quantifying the relative contributions of inter- and intra-molecular interactions to protein-DNA binding free energy. Paillard and Lavery (52) and Gromiha *et al.* (72) have both performed large-scale analyses of this sort using different potential

functions, and found that the contributions of direct and indirect recognition vary from structure to structure. Previously, Steffen *et al.* showed that they could partially predict the preferred binding sites of the integration host factor (IHF) protein by examining the energies of bending for different DNA sequences (73, 74). A number of other structure-based analyses have investigated this question for different protein-DNA complexes (46, 75-79). This is an excellent example of a question that is very difficult to investigate using experimental techniques; the ability to computationally model protein-nucleic-acid complexes allows for scientific investigation that would be otherwise impractical.

For statistical potential functions in particular, another interesting, largely unexplored application lies in the refinement of protein-nucleic-acid structures. Traditionally, both crystallographic and NMR structures have been refined using physics-based potentials, but more recently, Kuszewski *et al.* has developed a statistical potential function (DELPHIC) that describes the relative orientations of di-nucleotide steps. They have used this potential to refine several NMR structures of DNA and RNA molecules (80, 81), and demonstrated that the potential was able to improve the quality of the refined structures, and that the use of a database-derived potential did not hinder their ability to obtain correct refinement for non-canonical nucleic acid structures. These results suggest that statistical potentials may one day be widely used for the refinement of molecular structures.

### 1.4.3 Structure-Based Genome Annotation

One of the most compelling applications of structure-based methods for the prediction of protein-nucleic-acid interactions is the also one of the most obvious: if it is possible to accurately predict the sequence-recognition preferences of a protein from its structure, it should be possible to use that structure to predict the binding sites of the protein within genomic sequence data. In fact, the idea of using structures of protein-nucleic-acid complexes to annotate genomes for transcription factor (and other) binding sites is so compelling that nearly every publication in the field has incorporated some sort of binding-sequence scanning experiment as a methodological test. For example, Kono

and Sarai used their statistical potential to predict the recognition sequences of 25 different DNA-binding proteins (to varying degrees of success), and also demonstrated the method's ability to identify five of the six known binding motifs of the MAT a1/α2 protein in the upstream region of a known target gene (26). Liu *et al.* used their distance-dependent statistical potential to find the known binding sites of the cyclic AMP regulatory protein (CRP) in the *E. coli* genome (59). Paillard and Lavery used their physics-based potential to reproduce the consensus binding sites of multiple protein-DNA complexes (52), while Morozov *et al.* used their method to predict position weight matrices for a number of DNA-binding proteins (29), and Robertson and Varani used their statistical potential function to recapitulate the cognate recognition sequences of 52 different DNA-binding proteins (2). There are still many details to consider (for example, most of these tests did not consider interface water molecules in their analyses, and few incorporated any type of interface flexibility) but it is nonetheless clear that the structure-based genome annotation is rapidly becoming a practical application. Moreover, some research is beginning to tackle the difficult problem of predicting nucleic-acid binding sequences from protein homology models: Morozov and Siggia recently employed a simple contact-counting method to predict the binding preferences of 57 *S. cerevisiae* transcription factor proteins of previously unknown structure (82). This approach represents a leap forward in the application of structural models to biological research, by expanding the number of interesting targets well beyond those proteins with high-resolution structures.

### 1.4.4 Rational Protein Design

If a structure-based model of a DNA- or RNA-binding protein is accurate enough to be used to predict the cognate nucleic-acid recognition sequences for that protein, then it might be possible to use the model to solve the converse problem – the prediction of a protein sequence that will optimally bind a given nucleic acid sequence. This is simply a restatement of the protein design problem, as applied to nucleic-acid-binding proteins. And while there has been an enormous amount of research dedicated to the structure-

based design of certain classes of nucleic-acid binding proteins (83) (particularly the zinc-finger proteins, which are a family of modular domains that can be combined to produce DNA-recognition molecules with a desired sequence-specificity (84-95)), there are still no general-purpose methods that can successfully design protein sequences that will fold into conformations with a desired DNA- or RNA-binding function. Given that designed zinc-finger proteins have already been used as therapeutic compounds for human diseases (94, 96), there are clearly many potential applications for any method that can solve this more general problem.

Recently, a few groups have been making progress toward this goal, using structure-based, protein-design algorithms. For example, Havranek *et al.* used the ROSETTA protein design software to recapitulate the native amino acid sequences of DNA-binding proteins given their protein backbone structures, as well as the structures of their bound DNA molecules (28); Chen *et al.* successfully applied the same approach to RNA-binding proteins (1). These results demonstrate that the ROSETTA potential can largely (though not perfectly) identify the native amino-acid sequences of protein-nucleic-acid complexes as optimal for their bound structures. This is a necessary – though insufficient – condition for any structure-based protein design algorithm.

To date, the ROSETTA software has been used to re-design two different nucleic-acid-binding proteins: Dobson *et al.* used the method to design a variant of the U1A RNA-binding protein (97), while Ashworth *et al.* conducted a more limited design of the I-MsoI homing endonuclease (a DNA-binding protein) (64). Dobson *et al.* demonstrated that the completely re-designed U1A molecule (with approximately 30% sequence identity to the wild-type protein) was able to fold into a native-like structure, but that unfortunately, the RNA-binding function of the protein was abolished. Ashworth *et al.* used a more conservative design strategy for the I-MsoI protein-DNA complex, making only two amino acid mutations at the protein-nucleic-acid interface, and demonstrated that the re-designed protein bound specifically to a DNA sequence with a single guanine/cytosine base-pair transversion (relative to the cognate recognition sequence). Thus, although early tests of this method have been promising, there is still a great deal of

work to be done before it is possible to reliably design nucleic-acid-binding proteins of all classes, with arbitrary sequence specificity.

Both of these attempts at structure-based protein design relied exclusively upon structural information (coupled with the energetic analysis provided by a physical potential function) to make their predictions. However, a few groups have demonstrated a phenomenon that may be useful for a more directed type of protein design: significant correlations can be observed between patterns of protein sequence evolution within families of nucleic-acid-binding proteins, and the evolutionary patterns of the nucleic-acid recognition sequences of those same proteins (98-100). In particular, Raviscioni *et al.* demonstrated that for 12 different families of transcription factors, the evolutionarily most important protein residues of the families tended to interact with the most conserved base pairs of their DNA recognition sequences, and subsequently used this information to experimentally alter the sequence specificity of a zinc-finger transcription factor (100). Thus, it appears that this so-called "evolutionary trace" methodology (101) can be useful to guide the rational design of nucleic-acid-binding proteins, both by highlighting the most relevant protein residues for nucleic-acid recognition, and by suggesting protein mutations that will lead to desired changes in sequence specificity. Coupled with physics-based or statistical models of protein-DNA interactions, evolutionary trace methods may prove to be a powerful tool for the rational, structure-based design of DNA- and RNA-binding proteins.

## 1.5 Critical Challenges and Future Work

Thanks to steady increases in computer power, the size of the structural database, and our improved understanding of protein, DNA and RNA biochemistry, the structure-based modeling and prediction of protein-nucleic-acid interactions is rapidly becoming a realistic prospect. The field has experienced impressive advancements within the last decade: ten years ago, research into protein-nucleic-acid recognition was dominated by the search for a non-existent "recognition code", yet today we may choose between multiple successful methods to predict, design and model protein-DNA and protein-RNA

interactions at an atomic level. Nevertheless, a great deal of work remains, if the promise of this research is to be fulfilled. In particular, there are three major areas of research that must be advanced, if the modeling of protein-nucleic-acid interactions is to become a widely used tool: the field must improve its ability to simulate molecular flexibility of protein, DNA and RNA molecules; must account for missing details in models of the molecular interface (such as water, ions, *etc.*); and improve the accuracy of the potential functions used to conduct these simulations.

As noted above, the best current methods for simulating protein-nucleic-acid interactions can capture only small amounts of molecular flexibility, such as protein side-chain rearrangements, or (in the case of molecular dynamics methods) small-scale molecular motions. However, even these limited techniques take prohibitive amounts of computer time, and thus far have been unable to significantly improve simulation results in most circumstances. When one considers that protein-DNA and (particularly) protein-RNA interactions are often dominated by large-scale conformational rearrangements of both protein and nucleic acid molecules, the need for improved conformational simulation techniques is all the more acute. Clearly, new methods need to be developed to accurately sample or simulate the motions of protein and nucleic-acid molecules as they interact. Moreover, these new methods must be computationally tractable for large-scale simulations, if they are to be usefully applied to problems such as structure-based genome annotation.

In addition to limitations in the ability of current techniques to simulate molecular motion, it is clear that many of today's models are missing important details of the protein-nucleic-acid interface – for example, few of the methods discussed here explicitly consider the role of interface water molecules or metal ions at the protein-nucleic-acid interface, despite the fact that nearly every known protein-DNA and protein-RNA complex depends on the precise positioning of these molecules to mediate sequence specificity and enzyme function. The associated questions are numerous: how do water molecules position themselves in the protein-nucleic acid interface? What are the enthalpic, entropic and free-energy consequences of water- or ion-mediated interactions

with nucleic acid molecules? Can interactions with highly conserved interface water molecules be displaced through the careful re-design of hydrogen-bonding networks? And yet, this is just one area for which our understanding of the details of protein-nucleic-acid recognition is poor; from the importance of cation-pi interactions, to the role of conformational entropy in binding free energy, there are many other important, unanswered questions remaining to be explored.

Finally, as we advance our understanding of the dynamics and details of protein-nucleic-acid recognition, we will almost certainly uncover deficiencies in our understanding of the forces and interactions involved in these processes. For now, it appears that even the simplest statistical potentials perform as well as complicated molecular dynamics force fields in many situations, but will this continue to be true as simulations increase in complexity and detail? Will more complicated models of electrostatics, solvent effects, and hydrogen bonding be necessary to accurately model detailed interactions at molecular interfaces? Or, is it possible that database-derived potentials can implicitly capture these effects as well as the best physics-based methods?

Ultimately, the structure-based modeling of protein-DNA and protein-RNA interactions is an vibrant, interesting field, with many interesting and relevant scientific implications; nevertheless, it is not without real challenges. Within the last decade, the field has witnessed great advancements in understanding the mechanisms responsible for sequence-specific, protein/nucleic-acid recognition, and a number of practical applications have been demonstrated, despite the use of relatively simple techniques. Yet there are many unsolved problems in this area, and the promise exists for significant advances in the future. As these open questions are addressed, it seems probable that the computational prediction of protein/nucleic-acid interactions will emerge as an important, widely used application of structural bioinformatics technology.

# 2 A Statistical Hydrogen-Bonding Potential for Protein-RNA Complexes Predicts Specific Contacts and Discriminates Decoys

## 2.1 Introduction

The interaction of DNA- and RNA-binding proteins with nucleic acids plays a critical role in gene expression and its regulation. If we had available proteins that could control these interactions at will, it would be possible to interfere with gene expression pathways and gain a much better understanding of the architecture of gene expression networks. Combinatorial methods such as phage display have been used to engineer DNA-binding proteins with altered specificity, but these methods have limitations (102, 103) and have met with only limited success when they have been applied to RNA-binding proteins (104, 105). Given a better understanding of the principles of nucleic acid recognition, it might be possible to use rational approaches to design new RNA- and DNA-binding proteins. By establishing a design cycle involving both computational design and experimental validation, we would also be able to examine the molecular origin of recognition. Thus, the development of a physical models capable of reliably quantifying the molecular interactions responsible for affinity and specificity between proteins and nucleic acids is critical to the development of computational tools to design new RNA-binding proteins.

The recent spectacular increase in the number of structures of protein-nucleic acid complexes provides unprecedented opportunities. A number of authors have analyzed protein-nucleic acid interfaces computationally using visualization and statistical tools analogous to those used with proteins (5, 8, 9, 11, 106-115). In these important studies, common interaction patterns between amino acids and nucleotides were reported. The relative roles of packing, hydrogen bonding and electrostatic interactions in molecular recognition were described as well. In some cases, it was possible to attribute interaction propensities (*e.g.* Arginine-Guanine *etc.*) to specific patterns of hydrogen bonding and electrostatic interactions (9, 107). However, no systematic attempt has been made to

correlate these geometrical preferences with quantitative estimates of the relative contribution of each interaction to the total free energy of binding. Computational studies of protein-nucleic acid interactions remain very few when compared to the body of theoretical and experimental work dedicated to understanding interactions within protein cores and at protein-protein interfaces, and to redesigning new protein structures and interfaces (16, 63, 116-126). In other words, the knowledge encoded in the ever-growing database of protein-nucleic acid structures remains to be exploited in the quantitative dissection of energetic features responsible for affinity and specificity and in the development of predictive tools to be used in specificity redesign.

The strong orientational character of hydrogen-bonding interactions (127) makes them particularly important in determining the specificity of protein recognition and folding (128). Protein-nucleic acid interfaces are significantly more polar compared to most protein-protein interfaces and to protein cores (108): interactions involving ion pairs and hydrogen bonds should play a correspondingly greater role in dictating interaction specificity between proteins and nucleic acids (8, 106, 107, 115). However, the quantitative description of the geometric features of hydrogen bonding interactions from first principles is not straightforward. The direction of the lone electron pair cannot simply be assumed by the hybridization of the acceptor, because hydrogen bond formation may perturb the hybridization state of the acceptor atom (129, 130). In fact, most current force fields used in molecular dynamics simulations describe hydrogen bonds through a combination of Coulomb electrostatics and Lennard-Jones interactions with refined atomic charges, and thus lack explicit directionality (35, 131, 132). Furthermore, bulk physical models cannot easily capture differences in entropy costs associated with freezing exposed and buried side-chains or solvent-dependent effects.

An attractive approach to the description of hydrogen bonding interactions relies on the statistical examination of hydrogen bonds observed in high-resolution crystal structures (133-135). The statistical preferences observed experimentally can then be converted into a mean field potential by inverting Boltzmann statistics (136). The mean-field potentials relate the probabilities of occurrence of atom-atom interactions in a

database to the energies of these interactions (13, 55, 136-138) and implicitly incorporate environmental effects such as solvation and side-chain entropy. While theoretical limitations of this approach have been described (139), the technique has been shown to be quite effective, in practice (54, 140). Indeed, the physical basis for such a potential has been demonstrated by its striking correspondence, at least for protein side-chains, with quantum mechanical calculations of hydrogen-bonded dimers (141).

This chapter describes the development of a hybrid statistical/physical model for protein-RNA interfaces. The model is based on physical potentials which describe packing, solvation and electrostatics, and on a distance- and orientation-dependent hydrogen-bonding potential developed from the statistical analysis of hydrogen bonds observed in the high-resolution structures of protein/nucleic-acid complexes. The predictive power of the atomic model is demonstrated through its ability to recover the native amino acid sequence of a set of diverse protein-RNA interfaces. Finally, a scoring function based on the new hydrogen bonding potential is used to successfully discriminate native protein-RNA complexes from a large set of non-native decoy structures.

## 2.2 Methods

### 2.2.1 Construction of a Protein/Nucleic-Acid Structure Database

Protein-DNA and Protein-RNA structures were downloaded from the Protein Data Bank (PDB) (142). Only X-ray crystal structures with a resolution of 2.5Å or better and a crystallographic R factor of 0.25 or better were included in the statistical analysis. The database contained 42 protein-RNA and 125 protein-DNA complexes as of December, 2003. However, the protein-RNA complexes include the 50S ribosome structure comprising 2RNA and 28 individual polypeptide chains (143); therefore, the dataset effectively contains nearly 70 independent protein-RNA structures. For crystals with multiple complexes in a unit cell, only one representative structure was included. The database was checked with BLAST or MACAW to remove redundant protein structures

(more than 30% sequence homology), but homologous proteins were retained when bound to DNA or RNA sequences that were significantly different.

### 2.2.2 Analysis of Hydrogen Bonding Geometry

Hydrogen atoms are generally not included in the coordinates derived from the crystal diffraction data. Thus, polar hydrogen atoms were added when the position of the hydrogen itself is clearly defined by the chemistry of the donor atom. For proteins, protons were added to all backbone amides and to the Trp indole, His imidazole, Asn and Gln amides and Arg guanidinium groups. For nucleic acids, imino and amino protons were added. The bond length between proton and donor was set to 1.01Å for N-H bonds (as established by CHARMM27) (38, 39, 41). Angles were defined using the same method as used by HBPLUS (144), with the exception of the protein backbone amide protons, where the angles of C-N-H and Cα-N-H were set to be equal (the difference is only 4° in HBPLUS). It is difficult to define the orientations of Asn, Gln and His side-chains in X-ray crystal structures at resolutions greater than about 1Å, and therefore, incorrect placement (flipping) is possible in these structures; this problem was not corrected in the present work, as it would require assumptions about hydrogen bonding energies. The protonation state of His was assumed to be the most common Nε2 protonation state (144). No attempt was made to add polar hydrogens to the hydroxyl groups of Ser, Thr and Tyr, to the amino group of Lys, or to the RNA 2' hydroxyl. These hydrogens are not observed explicitly in models derived from X-ray diffraction studies and cannot be located in an unbiased way in the absence of neutron diffraction data. Because of these omissions, the distributions of hydrogen bonding interactions among different amino acids and nucleotides differ somewhat from previous studies (8, 9, 106).

Four geometric parameters were used to describe hydrogen bond geometry (Figure 2.1), as previously described (54): i) the distance between the hydrogen and acceptor atoms ($\delta_{HA}$); ii) the angle at the hydrogen atom ($\Theta$); iii) the angle at the acceptor atom ($\Psi$); iv) the dihedral angle (X) corresponding to rotation around the acceptor-acceptor base bond. For the dihedral angle around the phosphate oxygen (O1P) and phosphorous,

the reference atom (R) was chosen as the second phosphate oxygen (O2P); therefore, the plane defined by O1P-P-O2P defines 'planarity' for phosphate oxygen acceptors. A pre-defined cutoff range (1.4Å to 2.6Å) was set for distance between hydrogen and acceptor atoms, while an upper limit of 4Å was chosen for the donor-to-acceptor heavy-atom distance; no pre-condition was applied for the three angular parameters describing hydrogen bond formation. In the analysis of geometric preferences, bin sizes of 0.1Å and 10° were assigned to describe distance ($\delta_{HA}$) and angular distributions ($\Theta$,$\Psi$,X), respectively. After counting the number of observed hydrogen-bonding contacts in each bin, raw counts were corrected for the different volume elements encompassed by the bins to ensure that the number of observations in each bin is representative of the density of points and is not affected by the different bin sizes (135). Adjustments of sin($\Theta$) and sin($\Psi$) were applied to achieve this correction, but no such adjustment was applied to the X angle, because the volume elements considered for the dihedral angle are of equal size. A distance correction ($\delta_{HA}^2$) was also applied.

### 2.2.3 Construction of a Potential Function for Hydrogen-Bonding Interactions

One of the goals of this study was to generate a self-consistent model for description of proteins, nucleic acid and their complexes for design purposes. Therefore, the parameters chosen to describe hydrogen bonds in nucleic acids (Figure 2.1) were equivalent to those used to describe hydrogen bonds in proteins (54). We defined an orientation-dependent hydrogen-bonding potential comprising a distance-dependent energy term ($E(\delta_{HA})$) and three angular dependent energy components: $E(\Theta)$ (the angle at the hydrogen atom), $E(\Psi)$ (the angle at the acceptor atom) and $E$(X) (the dihedral angle of the hydrogen bond). The hydrogen bonding potential was derived based on inverting Boltzmann statistics to convert observed frequency distributions into a potential of mean force. This is possible if the assumption is made that the total energy of a system frozen in very low energy states is the sum of independent contributions (54, 133):

$$E(p) = -kT\ln\left(\frac{f_{pdb}(p)}{f_{random}(p)}\right) \qquad (2.1)$$

Where $f_{pdb}(p)$ is the frequency at which a geometric parameter $p$ is observed in a certain bin in the dataset and $f_{random}(p)$ is a reference frequency value assuming an unbiased distribution in all bins. The hydrogen bond energy ($E_{HB}$) was then derived from the linear combination of the four distance and orientation terms under the assumption that they are independent:

$$E_{HB} = E(\delta_{HA}) + E(\theta) + E(\Psi) + E(X) \qquad (2.2)$$

The free energy of protein-RNA interactions was modeled as the linear combination of physical and knowledge-based potentials describing: 1. van der Waals interactions (the attractive part of a Lennard-Jones potential ($E_{LJattr}$) and a distance-dependent repulsive term ($E_{LJrep}$)); 2. The orientation-dependent hydrogen bond potential ($E_{HB}$); 3. The implicit solvation free energy ($G_{sol}$)(123); 4. The backbone-dependent rotamer probability ( $E_{rot}(aa,\phi,\varphi)$)(56); 5. The amino-acid-type (aa) dependent backbone $\phi,\varphi$ probabilities ( $E_{bb}(aa \mid \phi,\varphi)$); and 6. The amino-acid-type dependent reference energies ( $E_{ref}(aa)$):

$$\begin{aligned} G = {} & W_{LJattr} \cdot E_{LJattr} + W_{LJrep} \cdot E_{LJrep} + W_{HB} \cdot E_{HB} + W_{sol} \cdot E_{sol} + \\ & W_{bb} \cdot E_{bb}(aa \mid \phi,\varphi) + W_{rot} \cdot E_{rot}(aa,\phi,\varphi) + \sum_{aa=1}^{20} n_{aa} E_{ref}(aa) \end{aligned} \qquad (2.3)$$

Two types of orientation-dependent hydrogen bonding potentials were used: one is based on previous study for hydrogen bonds between amino acids (54), and the other is directly derived from current result for hydrogen bonds between amino acids and nucleic acids. When the hydrogen bonding potential was supplemented with a Coulombic model of charge-charge interactions, a linear distance-dependent dielectric constant was used

and partial charges were taken from the CHARMM19 parameter set for proteins (131), and from CHARMM27 for RNA (39).

The weights W of the different components of the model were obtained by requiring the energy function to optimally reproduce the native amino acid sequences of known protein-RNA interfaces. For this task, we used a training set of 25 protein-RNA complexes, including the ribosome (1JJ2.pdb). Amino acid-dependent reference energies (that approximate the free energy of the unfolded reference state) were also obtained in the same fitting procedure. The remaining 17 complexes were set aside as a test set. During the fitting procedure, each of the components of the energy function for all protein side-chain rotamers at each interface position were computed assuming a constant environment of all other amino acids in their native conformation. The weights were then optimized using a conjugate gradient method to maximize the probability of the native amino acid type at each position. The rotamer library was taken from Dunbrack (145), with the addition of small deviations (10-20°) of the $\chi_1$ and $\chi_2$ angles for buried residues. All RNA atoms were fixed except for the 2' hydroxyl, whose position was optimized using a rotamer approach (146) according to the hydrogen bonding network. The protein-RNA interface was defined according to distance cut-off values between the C1' of nucleic acids and Cβ of amino acids. Depending on the size of the amino acid side-chain, the distance cut-off value varies from 10 to 15Å.

### 2.2.4 Creation of Decoy Sets

A set of 2000 decoys for each of five representative protein-RNA complexes were generated using the ROSETTA docking routine by rigid-body perturbations of the relative position and orientation of the two partners in the protein-RNA complexes (147). Both the protein and RNA molecules were treated as a rigid bodies during docking. However, interfacial amino acid side chains were repacked and minimized using a backbone-dependent rotamer packing algorithm after rigid-body docking (70). Decoys were scored and compared to the native structure based on the hydrogen bonding scoring function derived directly from the statistical analysis of protein-nucleic acid hydrogen

bond geometries (equation 1) without any weight. For infrequent distance and angular values, the score was set to 0 and no penalties were applied.

## 2.3 Results

### 2.3.1 Construction of a Protein-RNA Structure Database

As of spring, 2004, the Protein Database Bank (PDB) contains approximately 166 high-resolution (< 2.5Å) protein/nucleic-acid crystal structures. Accounting for the complexity of the ribosomal structure (28 protein and 2 RNA molecules), the total number of independent structures included in the database is close to 200. The database contains 3445 distinct hydrogen bonds involving protein and DNA or RNA. Phosphate oxygen atoms provide the largest number of hydrogen bond acceptors (53%), while the amino groups of amino acid side-chains (1167) and backbone amides (672) are the most common donors. This number certainly under-represents the total number of hydrogen bonding interactions between amino acid side-chains and nucleic acids, since $sp^3$ hydrogen bond donors (Ser, Thr and Tyr hydroxyl groups, and Lys $NH_3$) were excluded from the analysis because their hydrogens cannot be positioned explicitly without assumptions about hydrogen bonding energies.

The current structural database is too small to analyze each possible pair of hydrogen bond donor and acceptor types while generating statistically significant results. Therefore, different types of donor and acceptors were grouped together according to their structural and chemical similarity. Subtle differences between related hydrogen-bonding groups are inevitably lost (e.g. all base nitrogens were classified as a single atom type), but smooth distributions could be generated for most acceptor/donor pairs, suggesting that the statistical sample is large enough to yield reliable results. In choosing how to partition hydrogen bonds, we followed criteria similar to those used in analogous studies of proteins to ensure consistency in the description of hydrogen bonding interactions (54). Five types of hydrogen bond donors and acceptors were defined for nucleic acids, and five for proteins, based on whether the atom belongs to the protein and nucleic acid backbone or side-chain and on the hybridization state of the acceptor (Table

2.1). Separate statistics were collected for protein side-chain acceptors of $sp^2$ and $sp^3$ hybridization to take into account different electronic distributions around the acceptor atoms. Phosphate and ribose oxygen atoms were separated as well because of their different hybridization states. This classification partitions all hydrogen bonds between proteins and nucleic acids into 11 different classes (Table 2.2).

### 2.3.2 An Analysis of Hydrogen Bonds at Protein/Nucleic-Acid Interfaces

We have analyzed the distance and angular distributions for hydrogen bond donor and acceptor pairs observed in protein-nucleic acid complexes according to the four geometrical parameters defined in Figure 2.1:

#### 2.3.2.1 Hydrogen Bond Distance ($\delta_{HA}$)

The maxima in the distance distributions between protons and hydrogen-bond acceptors are generally centered between 1.8Å and 2.0Å, with small differences observed between different classes of hydrogen bonds. However, the breadth of the distributions differ when particular donor-acceptor pairs are examined. Interactions between phosphate oxygen atoms and both protein backbone and side-chain amide protons are the most common polar contacts at protein-nucleic acid interfaces (8, 106-108). For both sets, we observe well-defined maxima in the distributions, as would be expected for interactions with strong hydrogen bonding (as opposed to purely electrostatic) character (Figure 2.2a). The distance distribution for protein backbone amide interactions with phosphate oxygen atoms (data not shown) is centered over a narrower range than interactions with side-chain $NH/NH_2$ groups (Figure 2.2a), probably reflecting the structural constraints imposed by protein backbone secondary structure. Although hydrogen bonds to base nitrogen atoms are not numerous, the relatively sharp distance distribution observed suggests that these interactions are energetically very favorable and geometrically highly constrained (Figure 2.2b). Interactions between base NH groups and protein backbone carbonyl oxygen atoms have a relatively sharp distance distribution (data not shown), suggesting that the steric and structural constraints imposed by the

protein backbone and the RNA bases result in few energetically favorable interaction geometries.

#### 2.3.2.2 Hydrogen Bond Donor Angle (Θ)

This angle measures the linearity of the hydrogen bond: if a hydrogen bond were perfectly linear, its value would be 180°. As expected, hydrogen bonds between nucleic acids and proteins are almost always very close to linear; the distributions generally have maxima in the Θ angle range between 160° and 180°. Interactions between the RNA backbone phosphate oxygen atoms and the protein backbone amides display particularly strong linearity (data not shown), while interactions with side-chain NH groups have broader distributions with a maximum slightly removed from the linear value (Figure 2.2a). Interactions between base nitrogen atoms and protein backbone amides have nearly perfect linear distributions (data not shown), but broader spreads are observed for contacts between base nitrogen atoms and protein side-chain NH groups (Figure 2.2b). The distributions for hydrogen bonds between base oxygen atoms and protein backbone amides and side-chain NH groups have their maxima skewed to values slightly smaller than linear; the distribution is particularly broad for interactions involving the protein side-chains (Figure 2.2c).

#### 2.3.2.3 Hydrogen Bond Acceptor Angle (Ψ)

The acceptor angles for interactions between phosphate oxygen atoms and protein donors are centered at 120°, with a broad distribution, especially for interactions involving protein side-chains (Figure 2.2a). Hydrogen bonds between nucleic acid base nitrogen atoms and the protein side-chain NH groups (Figure 2.2b) have Ψ distributions resembling those observed for interactions involving protein side-chains (54). In contrast, hydrogen bonds to base oxygen atoms have broader distributions, skewed to values much closer to linear, particularly for hydrogen bonds involving protein side-chain NH groups, where a nearly flat distribution is observed between 120° and 180° (Figure 2.2c).

#### 2.3.2.4 The dihedral angle (X)

This angle measures the planarity of the hydrogen bond. A value of 0° (or ±180°) occurs when the proton is located in the plane defined by the acceptor, acceptor base and reference atom (Figure 2.1). Protein backbone and side-chain amide groups make strongly planar interactions with base nitrogen acceptors. This preference is more significantly pronounced for protein side-chains (Figure 2.3a) because of a larger statistical sample, but it is also clear for the protein backbone (data not shown). This observation places strong constraints on the direction of the hydrogen bonds between amino acids and the RNA/DNA bases. The planar preference for base carbonyl oxygen atoms is not as marked as for the ring nitrogen atoms, but still clearly observable. While interactions between nucleic acid base NH and protein backbone carbonyl oxygen atoms are devoid of any statistical preference (data not shown), weak but clear preferences for a planar arrangement are observed for interactions between nucleic acid NH and $NH_2$ donors and $sp^2$-hybridized acceptors on the side-chains of proteins (Figure 2.3b). Hydrogen bonds involving phosphate oxygen atoms tend to be planar when paired with amino acid backbone amides, but are less clearly so when paired with side-chain donors (Figure 2.3c,d).

### 2.3.3 Construction of a Knowledge-Based Hydrogen-Bonding Potential Function

The potential of mean force describing hydrogen-bonding interactions at protein/nucleic-acid interfaces was derived from the reverse Boltzmann relationship by taking the negative logarithm of the observed frequency distributions for each hydrogen bond acceptor-donor pair. The total hydrogen bonding potential is composed of a linear combination of the distance-dependent energy term ($E(\delta_{HA})$) and the three angle-dependent energy components ($E(\Theta)$, $E(\Psi)$, $E(X)$) (see Methods).

Figure 2.4 shows the result of this calculation for hydrogen bonds between base N and protein side chain NH groups. Clear minima appear in the energy profiles reflecting the strong distance and directional preferences, as observed in the database of high-resolution crystal structures. In other words, the strong distance- and orientation-

dependence of the hydrogen bond places significant energetic restrictions on the relative positions of the donor and acceptor atoms at protein-nucleic acid interfaces.

### 2.3.4 Prediction of Protein Sequences at Protein-RNA Interfaces

Two tests were carried out to demonstrate the importance of the distance- and orientation-dependent effects of hydrogen bonding at the protein-RNA interface, and to validate the ability of the statistical model to capture these effects. The first test probed the ability of the potential to recover the native protein sequence at a protein-RNA interface. This test is based on the assumption that the substitution of the residues at a given protein-RNA interface with non-native amino acids generally results in an increase in free energy compared with the naturally occurring sequence. This assumption is consistent with mutation studies, which shows that amino acid changes are most often destabilizing (148, 149). In order to assess the importance of the hydrogen bonding potential, we repeated this test, first by eliminating the orientation-dependent component of the potential, and finally by replacing the hydrogen-bonding potential with a Coulomb potential, using a linear, distance-dependent dielectric constant.

Of course, the complete energy function used to score protein-RNA complexes includes van der Waals interactions, solvation, amino acid rotamer and backbone conformational energies, in addition to the statistical-based hydrogen bonding potential. In deriving the complete energy model, we used 25 protein-RNA complexes (a total of 1500 amino acid positions) to obtain weights for these terms, and set aside 17 independent structures (a total of 425 amino acid positions) to execute the amino acid recovery test. The weights for the energy terms in each of the experiments were re-optimized (see Methods) for each test (complete hydrogen bonding function; no orientation-dependent component; no hydrogen bonding potential).

The results of the test are shown in Figure 2.5, Figure 2.6 and Figure 2.7, where we report how often the native amino acids were found to be energetically most favorable. The overall recovery rate is 44%, which compares well with what is observed on single domain proteins (52% for buried positions, and 26% for all positions) and protein-DNA

interfaces (43%) using similar experiments (28, 54). The lower recovery rate (compared with tests on protein cores) is to be expected, because the identities of the native amino acids at protein-RNA interfaces (like protein-protein interfaces) are not determined solely by energetic considerations, but also by functional and solubility constraints. The complete hydrogen-bonding potential identifies the native amino acid most often as the energetically most favorable replacement for almost all charged (A, D, K, R), polar (N, Q, S, T) (Figure 2.5) and polar aromatic (H, W, Y) amino acids (Figure 2.6). The exceptions are Lys, Gln, Thr and Trp. Although the hydrogen-bonding potential significantly improves the recognition of native amino acids for polar and charged residues, the overall prediction accuracy for these residue classes remains worse than for hydrophobic amino acids (A, I, L, V, F, M, G and P) which are all predicted with the highest frequency (Figure 2.7).

Significantly worse results are observed in nearly every case when the orientation-dependent component of the hydrogen-bonding potential is removed; the model performs even worse when the hydrogen bonding term is substituted with a purely electrostatic description of the interaction. Replacing hydrogen-bonding interactions with a purely Coulombic term gives the worst recovery rate in all cases except for Glu and Thr, and combining both the hydrogen bonding and electrostatic potentials only slightly improves the overall performance of the total energy function in recovering the native sequence (data not shown).

Results for individual amino acids are revealing. Arg has the highest recovery frequency (over 79%) among all 19 amino acids and is also preferred to the native amino acid for Lys, Gln, Thr and Tyr. This is consistent with the high occurrence of Arg (over 15%) at the protein-nucleic acid interfaces (8, 107). Interestingly, Lys was not recovered most frequently when the full hydrogen-bonding potential was used, but was found most often when the angular terms of the hydrogen bonds were switched off. This is probably due to the limited conformational sampling of the rotamer approach, which makes it difficult for long polar amino acids to find optimal hydrogen-bonding geometries (Lys was also omitted from the hydrogen bond geometry analysis, because its polar hydrogen

atoms cannot be placed without assumptions about hydrogen bonding energies). Similar results have also been reported in a protein-protein interface study (54). Thr is most often recovered when a purely electrostatic potential is used, but not when the hydrogen-bonding potential is used instead. For polar, aromatic amino acids, Trp is less frequently recovered than Tyr, which has the second highest frequency of recovery (66%). The high recovery rate for Tyr is presumably due to the hydrogen bonding properties of its hydroxyl groups, in addition to its ability to form stacking interactions. While these interactions are not explicitly modeled, steric constraints implicit to the Lennard-Jones term are likely to recapture at least some aspects of the base-amino acid stacking interactions observed in many protein-RNA complexes. Trp is present in only a very small number of cases (11 positions) in our test set, and is the only polar aromatic amino acid not selected correctly with high frequency of recovery. We note that its large aromatic ring could potentially introduce steric clashes if conformational space is not adequately sampled, and observe that when the aromatic proton radius was reduced from 1.2 to 0.7Å, the recovery of native Trp was increased (by reducing Tyr occurrence), and at the same time, the recovery of native Tyr also increases.

### 2.3.5 Decoy Discrimination in Protein-RNA Docking

As a second test, we assessed the ability of the new hydrogen bonding function to discriminate native from non-native protein-RNA structures (Figure 2.8). This test is based on the assumption that native protein-RNA interfaces (like protein-protein interfaces), are generally electrostatically optimized when compared to alternative binding conformations (54, 150-152). We selected five protein-RNA complexes, and generated 2000 decoy structures covering a range of RMSD values from below 1Å to over 20Å. The five structures were chosen according to their sizes (less than 200 amino acids and RNA lengths between 8 and 29 nucleotides), crystallographic resolution (1CVJ: 2.60Å; 1EC6: 2.40Å; 1FXL: 1.80Å; 1JID: 1.80Å; 1URN: 1.90Å), and characteristics of their interfaces – four complexes represent single-strand RNA interacting with one or two RRM's, and 1JID provides an example of a protein bound to the major grove of an

irregular helix RNA (Table 2.3). Together, these represent the major known interaction modes between protein and RNA. In order to produce docking decoys, small perturbations (translation and rotation) were applied to the native complex structures, to obtain both near-native decoys, as well as decoys with larger RMSD values. Protein backbone conformations in all structures were kept fixed, but protein side-chain conformations were modeled using a standard rotamer library to allow for the rearrangement of side-chains at the protein-RNA interface during binding. All RNA molecules were kept in their native conformations throughout the docking process.

Figure 2.8 graphically shows these results, while Table 2.3 shows the Z-score values measuring the discrimination of the native structures from all other decoy conformations. Again, we compared the full hydrogen-bonding potential with the performance of a Columbic potential with a linear distance-dependent dielectric constant. In all cases, the hydrogen bonding potential successfully discriminated the native structures, with a lowest Z-score of 2.70 (where success is defined as a Z-score > 1). The hydrogen-bonding potential performs much better than the Columbic model, especially in the low RMSD range (up to 3Å) where correct and incorrect structures are most difficult to discriminate. When the angular terms of the hydrogen bonding potential are removed, the Z-score values are only slightly affected, but three out of five native structures are not discriminated well from the rest of the decoys. The results of this test strongly suggest that native protein-RNA complexes maximize the number of hydrogen bonding interactions in the interface, and that hydrogen bond geometry plays a significant role in the affinity and specificity of protein-RNA interactions.

We expected the statistical model to perform best when recognition was primarily of single stranded nucleotides (as compared to more structured RNA molecules). Consistent with this, the shape of the score distribution at low RMSD values was not as distinct for the complex involving a structured RNA (1JID), compared to the other four decoys sets, though the Z-score values remain very high (9.12). In this structure, the protein binds to the major groove and tetraloop of a helical RNA, with few direct protein-base contacts. A complex network of highly ordered water molecules is also present in this protein-

RNA interface. The presence of water molecules is certain to affect the accuracy of the hydrogen bonding potential, since these are not modeled with the currently described methods, and are discarded in the docking process.

## 2.4 Discussion

There has been a remarkable, recent increase in the number of RNA-protein structures: we now know the structures of most if not all major RNA-binding protein families and how they bind RNA (153-155). However, for even the best-studied case (the RNA recognition motif, or "RRM"), the molecular basis of specificity in protein-RNA recognition remains far from clear (153, 156). A fruitful approach to understanding the molecular determinants of protein-protein interactions has been the establishment of computational tools to redesign the specificity of protein-protein interactions (118, 157-159). Clearly, we cannot claim to understand a physical phenomenon unless we are capable of making valid predictions based on our understanding; the computational redesign of proteins and protein-protein interfaces provides such a test. Thus, an aim of the present work is to establish comparable tools for the study of protein-RNA recognition.

### 2.4.1 The Statistical Potential Is an Effective Tool for Identifying Native-Like Protein-RNA Interactions

Our starting point was the statistical analysis of hydrogen bond geometries at the interfaces between proteins and nucleic acids present in high-resolution crystal structures. Recent manuscripts have analyzed statistical properties of protein-RNA interfaces and provided insight into features of recognition, such as amino acid preferences in the interactions with certain bases, and macroscopic characteristics such as polarity and average size (8, 9, 11, 106-109). However, the scope of these studies has until now failed to include the quantitative analysis of the energetic features responsible for specificity and affinity in protein-RNA recognition, nor have these analyses led to the development of a testable model of protein-nucleic acid interfaces with predictive power. Here, we

use statistical analysis to establish a distance- and orientation-dependent hydrogen-bonding potential function that is fully compatible (and indeed, inspired by), a successful model of hydrogen bonding in protein cores and at protein-protein interfaces (54). We have demonstrated that this potential provides a quantitative tool to analyze protein-RNA interfaces by conducting two independent tests: 1) we have successfully used the method to recover native amino acid sequences at protein-RNA interfaces, and 2) we have used the method to discriminate native structures of protein-RNA complexes from a very large set of docking decoys.

This statistical, hydrogen-bonding potential recovers native amino acids at protein-RNA interfaces approximately 44% of the time when included in a complete physical model of protein-RNA interfaces that contains terms describing steric interactions and solvation. This result is comparable to similar studies conducted with single-domain proteins and protein-DNA complexes (28, 54), which also used an orientation-dependent hydrogen bond potential, based on the geometries of hydrogen bonds observed in protein crystal structures. The success of current model is particularly encouraging when one considers that protein-RNA interfaces are generally more structurally diverse than double-stranded protein-DNA interfaces (where the interactions are mainly through the major or minor grooves of a regular helix). Our success in recovering polar and aromatic-polar amino acids is compromised when the angular terms of the hydrogen-bonding potential are abolished (i.e. the hydrogen bond is assumed to be radially symmetric) or when a Coulombic potential is used instead of the hydrogen bonding potential. As was observed for proteins, even if van der Waals and other components of the model are retained (and these terms undoubtedly provide geometric restriction to the possible range of intermolecular interactions), they are insufficient to discriminate native amino acid sequences from random mutations (54).

The current hydrogen-bonding potential also discriminates native protein-RNA structures from large sets of decoys prepared by a small-perturbation method, and greatly outperforms a purely Coulombic model, especially in the low-RMSD range. The description of the hydrogen bonding potential through statistical approach, as well as its

directionality and explicit placement of polar hydrogen atoms, provide major advantages over a purely Coulombic description of interactions. The decoy studies in this work also demonstrate that in the high-RMSD range (up to 15Å), the distance-dependent Coulombic potential has a better score-RMSD linear relationship than does the hydrogen-bonding potential (which is mainly effective below 3.0Å). Thus, in future protein-RNA (DNA) docking studies, one might consider combining these two potentials: in the initial binding-surface search, an electrostatic potential can be used to guide the two unbound partners; an orientation-dependent hydrogen bonding potential could be applied to refine the bound conformation.

Curiously, in the interface native amino-acid recovery test, we did not obtain higher recovery rates when Coulombic potential was used in combination with the hydrogen-bonding potential. This is probably because the backbones of both the RNA and the protein molecules have been fixed, and only the individual amino-acid side-chain conformations were allowed to change. The hydrogen-bonding potential has a greater discriminatory power in this near-native case.

### 2.4.2 Important Features of Protein/RNA Intermolecular Hydrogen Bonds

What features of the hydrogen bond between proteins and nucleic acids are most significant? We believe that there are several important characteristics that contribute to the performance of this method:

#### 2.4.2.1 Hydrogen Bonds Between Protein and RNA/DNA Atoms Are Constrained Over Narrow Distance and Angular Values

Hydrogen bonds involving the nucleic acid bases are undoubtedly an important source of specificity in protein-RNA/DNA recognition (8, 107). However, they are much fewer than the contacts with the backbone phosphates; in RNA, most interactions involve the protein backbone (107). The sharp distance and orientation preferences observed in the present study reveal very narrow minima in the potential function subtending these interactions. They are both energetically and geometrically constrained.

### 2.4.2.2 Hydrogen Bonds Involving the Nucleic-Acid Bases Have a Very Strong Preference for Planarity

The planarity of hydrogen bonds involving the nucleic acid bases is particularly stunning (Figure 2.3). By way of comparison, contacts between protein side-chains only display mild planar preference for $sp^2$ hybridized acceptors. Backbone contacts in proteins deviate significantly from planarity with maxima in the distribution near -120° for α-helices, -100° for irregular structures, with a bimodal distribution centered around -130°, and a broad peak near 0° for β-sheet structures (54). The very strong preference for planarity of hydrogen bonds with nucleobases (Figure 2.3) may reflect the electron distributions of the planar ring systems, as well as steric constraints of the interaction with the conjugated bases. Whatever its origin, this observation places very significant constraints on the type of intermolecular hydrogen bonds between proteins and nucleic acids. This observation may also have implications for drug design – many existing drugs contain hetero-aromatic rings, including nucleosides, which are likely to share hydrogen-bonding characteristics with the nucleic acid bases.

### 2.4.2.3 Hydrogen-Bonding Interactions with Phosphate Groups Have Strong and Distinct Angular and Distance Preferences

In contrast to the distance and angular distributions reported here, purely electrostatic interactions would generate distributions that increase monotonically with distance and would not be directional (54, 107). Phosphate oxygen acceptors provide the majority of intermolecular hydrogen-bonding interactions between proteins and nucleic acids. Very often, these interactions involve basic side-chains such as Arg or Lys (8, 9). While their contribution to affinity has long being recognized as very important, their contribution to specificity (indirect recognition) has been more difficult to dissect. The observation of clear distance and orientation constraints indicates that only certain structural arrangements are conducive to favorable interactions between nucleic acid phosphates and proteins. This is probably the major reason why a purely Coulombic model performs

less satisfactorily compared to an orientation-dependent hydrogen bonding model derived from existing protein-nucleic acid structures.

### 2.4.3 Hydrogen Bonding and Direct/Indirect Recognition

By comparison with hydrogen bonds within proteins and at protein-protein interfaces, much narrower planar angular ranges are generally accessible to nucleic acid atoms, particularly for interactions involving the bases (Figure 2.3). Clearly, formation of hydrogen bonds at protein/nucleic-acid interfaces places strong orientation constraints on the relative placement of hydrogen bonding atom pairs. These preferences define the kind of interactions that are energetically favorable between nucleic acid and proteins. Interactions involving the bases are especially directional and tightly constrained geometrically. Direct recognition of RNA and DNA functional groups, even by the protein backbone (as is very commonly observed in RNA-protein interactions) (54), is a highly effective way to achieve specific recognition because of these strong geometric constraints. Interactions involving phosphate oxygen atoms are most favorable within relatively narrow distance ranges and are remarkably directional. By controlling the spatial location of phosphate groups and therefore dictating which interactions between the phosphates and protein side-chains are energetically favorable (or even feasible), nucleic acid structure contributes to the indirect recognition of a nucleic acid sequence.

The present work introduces and validates a set of computational tools for the design of nucleic-acid-binding proteins with altered specificity. Such proteins would provide valuable new probes for biological interactions and, potentially, new therapeutic agents. Combinatorial methods such as phage display are certainly effective for at least some classes of nucleic acid binding proteins (102, 103, 160, 161), however, it would be highly advantageous to be able to alter the specificity of existing nucleic acid binding proteins in a predictive way, using design algorithms that have become increasingly powerful in the design of proteins and protein-protein interfaces(118-122). The physical model presented here is capable of energetically quantifying subtle molecular interactions between proteins and nucleic acids based on a full, atomic representation, and incorporating both

physical and statistical components. In the future, we hope to improve the model by allowing for RNA flexibility, and by incorporating more complicated electrostatic effects (such as cation-$\pi$ interactions). We hope that the method will find wide application to the rational, structure-based design of nucleic-acid binding proteins.

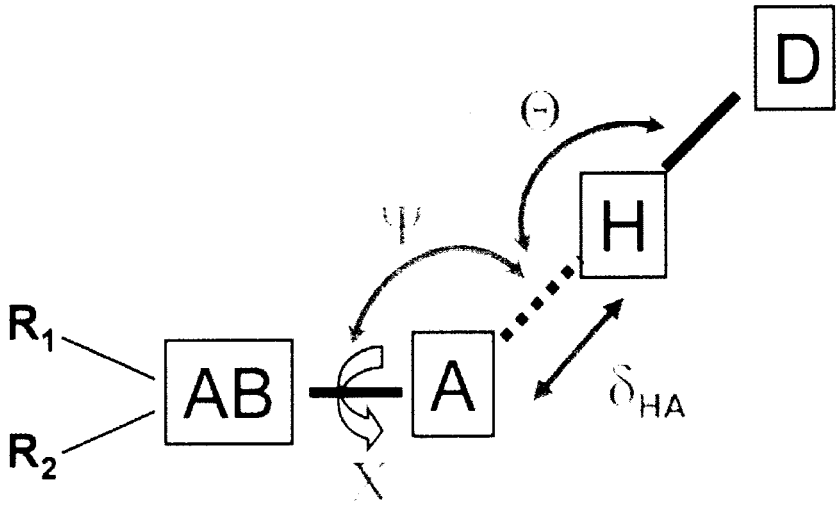



**Figure 2.1: The geometric parameters used to describe hydrogen-bond geometry.** Here, $\delta_{HA}$ represents the distance between the hydrogen and acceptor atoms; $\Theta$, the angle at the hydrogen atom; $\Psi$, the angle at the acceptor atom; and X represents the dihedral angle given by rotation around the acceptor-acceptor base bond. For $sp^2$ hybridized acceptors, X is a measure of the planarity of the hydrogen bond. Atoms are named according to their roles in the hydrogen bond: A, acceptor atom; D, donor atom; H, hydrogen atom; AB, acceptor base; $R_1$, $R_2$, reference atoms bound to the acceptor base.

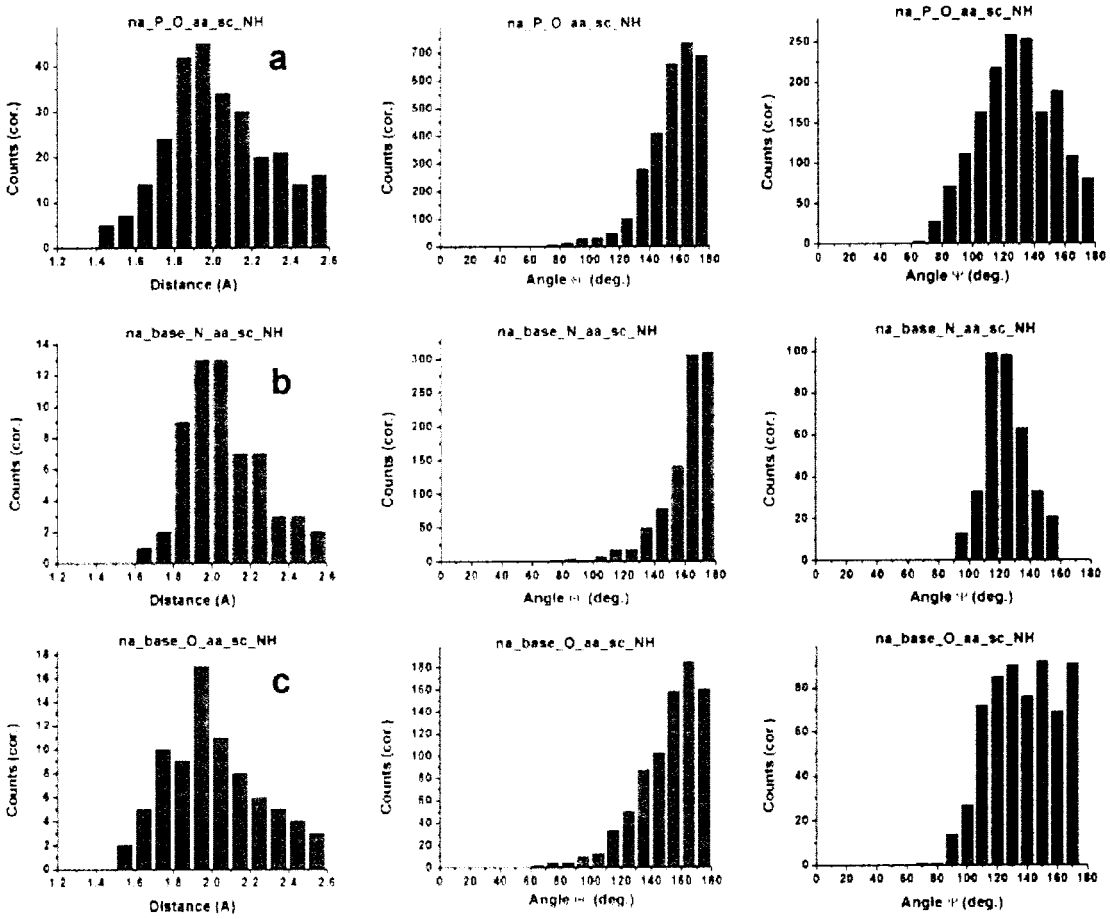



**Figure 2.2: Distance ($\delta_{HA}$), bond angle ($\theta$) and acceptor angle ($\Psi$) distributions for selected hydrogen bonds across protein/nucleic-acid interfaces**

a) Phosphate oxygen to protein side-chain $NH/NH_2$ groups, b) Base nitrogen to protein side-chain $NH/NH_2$ groups, c) Base oxygen to protein side-chain $NH/NH_2$ groups.

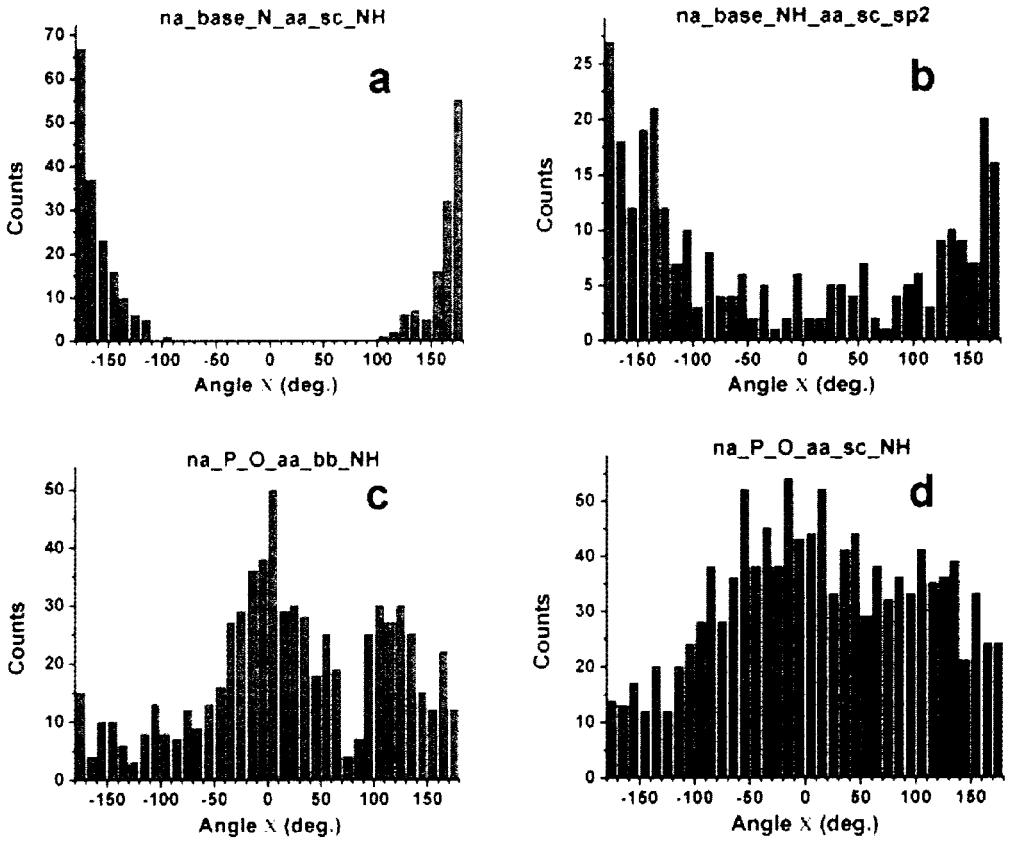



**Figure 2.3: Dihedral angle (X) distributions for selected hydrogen bonds across the protein/nucleic-acid interface**

a) Base nitrogen to protein side-chain $NH/NH_2$ donors, b) Base $NH/NH_2$ to protein side-chain $sp^2$-hybridized acceptors, c) Phosphate oxygen to protein backbone amide, d) phosphate oxygen to protein side-chain $NH/NH_2$

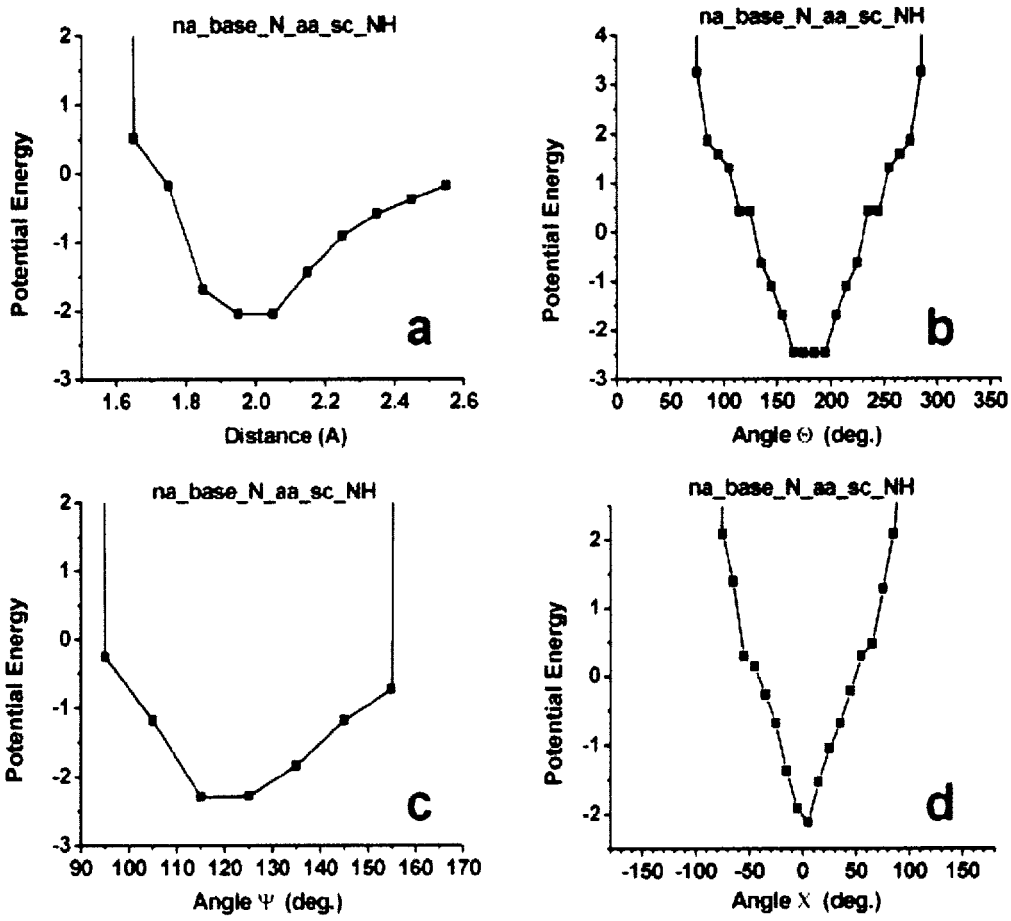



**Figure 2.4: Hydrogen bonding potential for interactions between base nitrogen acceptor atoms and protein side-chain NH/NH2 donors**

a) Distance $\delta_{HA}$, b) angle $\Theta$, c) angle $\Psi$, d) angle X

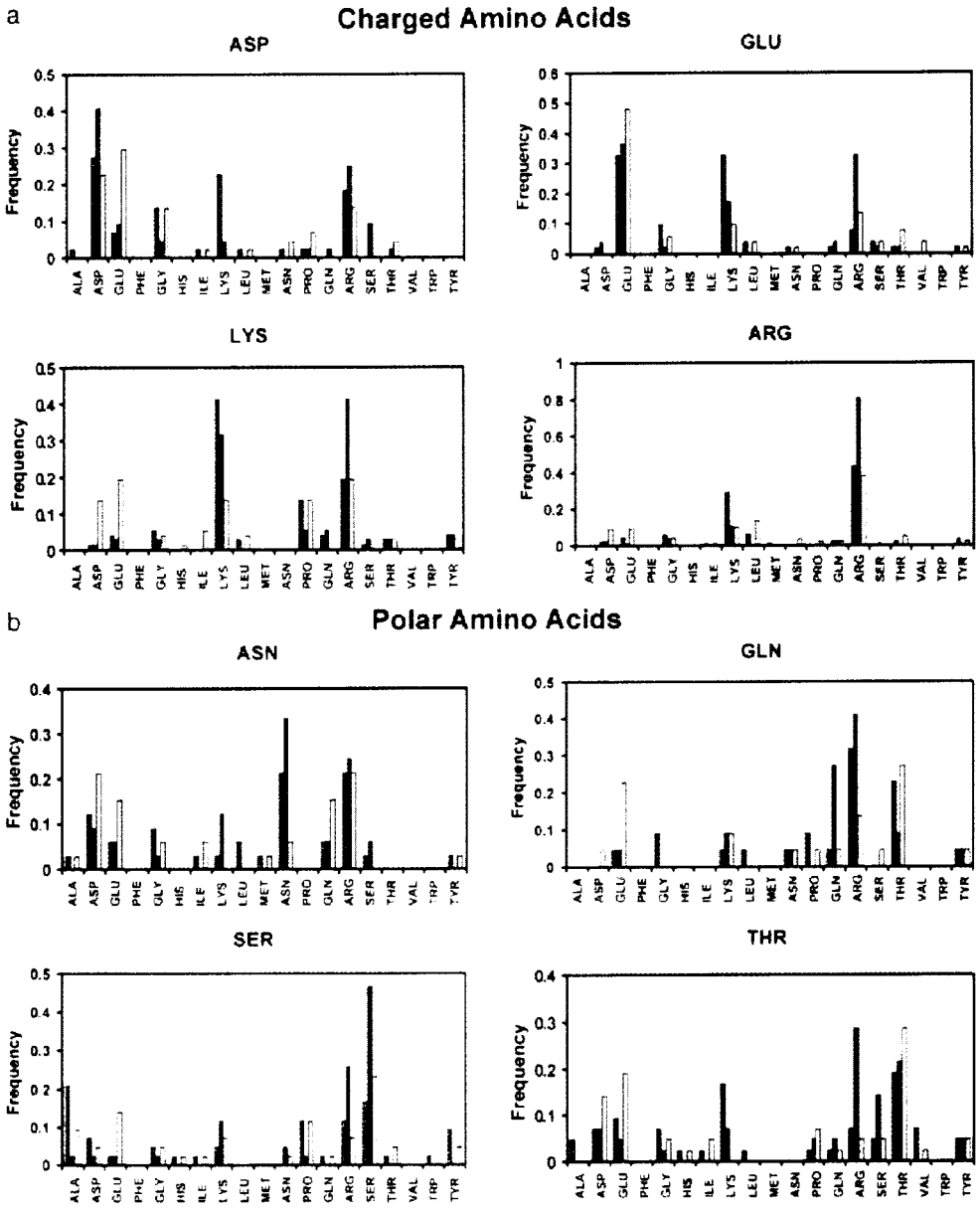



**Figure 2.5: Native protein sequence recovery for charged and polar amino acids**
Bars show how often native amino acid types (except Cys) are found to be energetically most favorable at each interface position probed. Panel a) Charged amino acids; b) polar amino acids. Different energy functions are used to test the substitution profile: red bars, the full energy function; light blue bars, energy function with the angular terms of the statistical hydrogen-bonding term disabled; yellow bars, hydrogen-bonding term is replaced by Coulomb electrostatics.

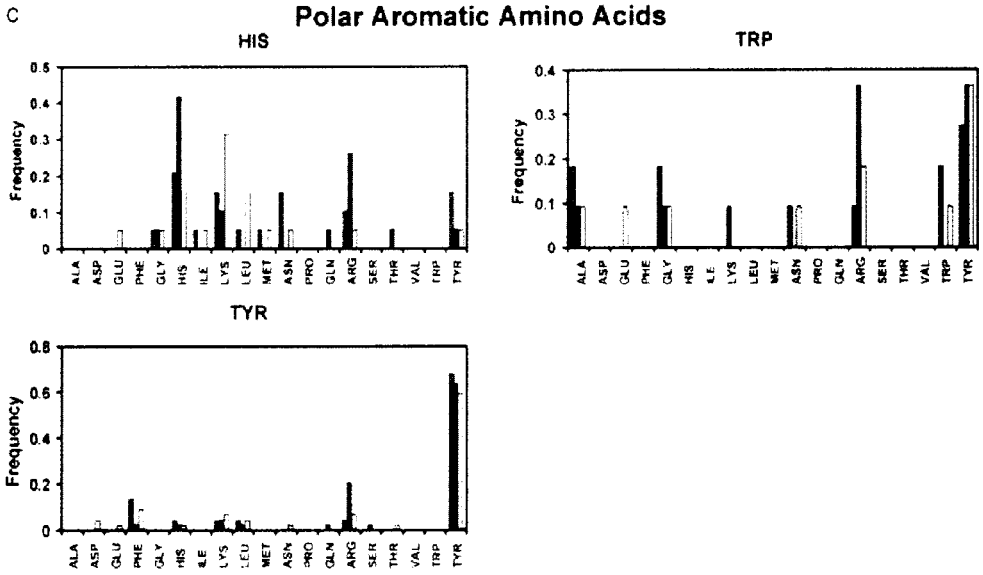



**Figure 2.6: Native protein sequence recovery for polar aromatic amino acids**

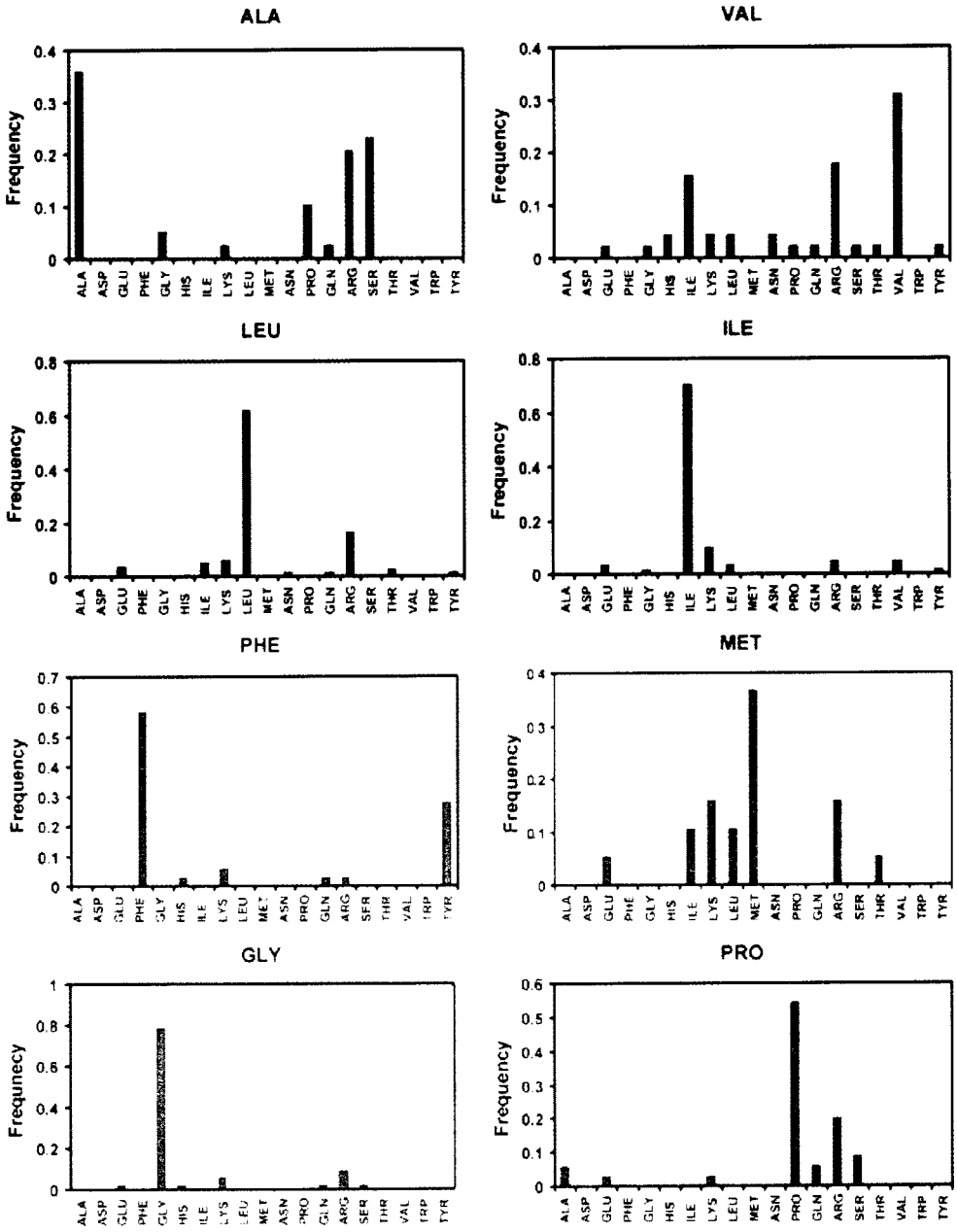



**Figure 2.7: Native protein sequence recovery for hydrophobic amino acids**

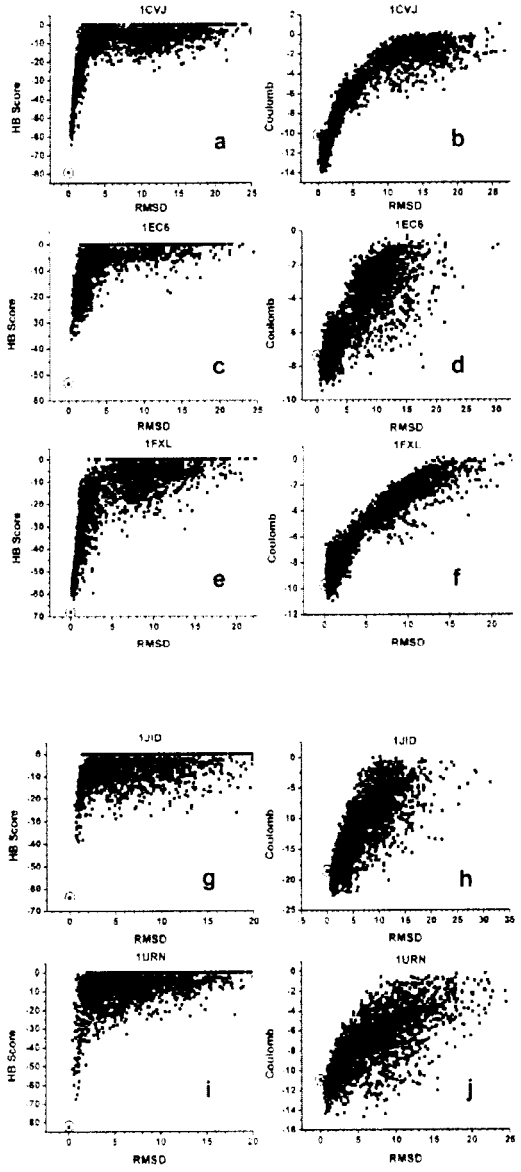



**Figure 2.8: Discrimination of docking decoys with the hydrogen-bonding potential.** Panels show score vs. RMSD scatterplot for 2000 docking decoys, where score is determined using a potential function containing a statistical hydrogen-bonding term (panels a,c,e,g,i), or a similar potential using only a Colulomb electrostatics model (b,d,f,h,j). In all panels, the point representing the native structure is plotted in red. Panels a, b) 1CVJ; c, d) 1EC6; e, f) 1FXL; g, h) 1JID; i, j) 1URN

**Table 2.1: Classification of hydrogen-bonding donors and acceptors**

| | Proteins | | | Nucleic Acids | | |
|---|---|---|---|---|---|---|
| Donors | aa_sc_NH | His | Nε2 | na_NH | A | N6 |
| | | Trp | Nε1 | | C | N4 |
| | | Asn | Nδ2 | | | |
| | | Gln | Nε2 | | G | N1, N2 |
| | | Arg | Nε, Nη1, Nη2 | | T/U | N3 |
| | aa_bb_NH | Backbone Amide (NH) | | | | |
| Acceptors | aa_sc_sp$^2$ | Asp | Oδ1, Oδ2 | na_base_O | C | O2 |
| | | Glu | Oε1, Oε2 | | G | O6 |
| | | His | Nδ1 | | | |
| | | Asn | Oδ1 | | T/U | O2,O4 |
| | | Gln | Oε1 | | | |
| | aa_sc_sp$^3$ | Ser | Oγ | na_base_N | A | N1, N3, N7 |
| | | Thr | Oγ1 | | C | N3 |
| | | Tyr | OH | | G | N3, N7 |
| | aa_bb_O | Backbone Carbonyl (O) | | na_P | O1P, O2P | |
| | | | | na_O | O5*, O4* O3*, O2* | |

**Table 2.2: Observed numbers of hydrogen bonds at protein/nucleic-acid interfaces**

| | | Nucleic Acid Atom Types | | | | |
|---|---|---|---|---|---|---|
| | | Donors | Acceptors | | | |
| | | na_NH | na_P | na_O | na_base_O | na_base_N |
| Protein Atom Types | aa_sc_sp$^2$ | 284 | | | | |
| | aa_sc_sp$^3$ | 51 | | | | |
| | aa_bb_O | 174 | | | | |
| | aa_sc_NH | | 1167 | 238 | 352 | 289 |
| | aa_bb_NH | | 672 | 80 | 94 | 44 |

**Table 2.3: Native Z-scores for the protein/RNA docking decoy sets**

| PDB ID | RNA Binding Mode | Native Z-Score | |
|---|---|---|---|
| | | Coulomb[a] | HB[b] |
| 1CVJ | Single-stranded RNA interacting with two RRM domains | 1.19 | 5.11 |
| 1EC6 | Protein loop interacting with single-stranded RNA | 1.09 | 6.53 |
| 1FXL | Single-stranded RNA interacting with two RRM domains | 1.55 | 2.70 |
| 1JID | Protein interaction with double-stranded RNA major groove and tetraloop | 1.36 | 9.12 |
| 1URN | Single-stranded RNA interacting with an RRM domain | 1.35 | 8.39 |

[a] ROSETTA potential + Coulomb electrostatic model
[b] ROSETTA potential + statistical hydrogen-bonding potential

# 3 An All-Atom, Distance-Dependent Scoring Function for the Prediction of Protein-DNA Interactions from Structure

## 3.1 Introduction

The specific recognition of nucleic acid sequences by nucleic acid binding proteins is of critical importance to the biological function of every living species. As a result, the phenomena responsible for this recognition process have long been of interest to biological scientists. Beginning with Seeman *et al.* (3), research into sequence-specific DNA recognition focused on the search for a "recognition code" — a collection of simple rules that would pair particular amino acids to specific bases. However, it was soon realized that any recognition code would be degenerate (4), and a general code may not exist at all (5, 6).

The structural mechanisms underlying sequence-specific DNA recognition received renewed attention with the increased availability of high-resolution structures for protein-nucleic acid complexes. Computational studies of these structures have classified their interactions (7), described features of their binding sites (8-10) and the evolutionary conservation of their interface residues (11, 12, 99). However, relatively little effort has been devoted to the application of this structural knowledge to computational models for the prediction of protein-nucleic acid interactions. In contrast, structural information has been used extensively to create potentials for prediction of protein structures (13-19, 55), as well as protein-ligand (20-23), and protein-protein interactions (19, 22, 24, 25).

A few attempts have been made to apply "physical" potential functions to the structure-based prediction of the DNA sequences that bind a given protein structure (27, 29, 52, 53), and to the prediction of protein sequences that will bind to a given DNA sequence (28). However, these complex models do not appear to significantly outperform even the simplest statistical methods that have been applied to this problem. For example, Kono and Sarai used the distributions of protein $C_\alpha$ atoms about the nucleic acid bases to derive a potential that could discriminate the native DNA-binding sequences of a diverse

set of DNA-binding proteins (26). More recently, Liu *et al.* predicted protein-DNA binding energies using a distance-based statistical potential describing residue-base interactions (59), while Zhang *et al.* introduced a pseudo-atom statistical potential for the same purpose (22). Despite their greatly simplified views of protein-nucleic acid recognition (none of these statistical methods explicitly considers electrostatics, molecular flexibility, indirect readout, or other factors widely believed to be important for protein-nucleic acid recognition) these methods have achieved some predictive success.

There is some evidence that the performance of these statistical potentials may be limited by the simplicity of their models. In particular, their resolution — the level of atomic detail that they capture — may significantly impact their discrimination capabilities. Indeed, work from our own group has demonstrated that the detailed statistical modeling of hydrogen bond geometry dramatically improves the ability of a physical potential function to predict sequence-specific protein-RNA interactions (1). This observation has led us to hypothesize that increased detail in the description of protein-nucleic acid complexes will improve the predictive power of statistical potential functions. A widely-held belief that the existing structural database is insufficient to train an all-atom statistical potential that is specific to protein-nucleic acid complexes is a concern. However, an extensive literature has addressed the problem of sparse training data for statistical potentials in proteins (13, 14, 18, 20, 21), and it is reasonable to believe that many of these methods are also applicable to the prediction of protein-nucleic acid interactions.

In this work, we report the development of a simple, distance-dependent, all-atom statistical potential function for the prediction of sequence-specific, protein-DNA interactions. This potential is able to reliably discriminate native-like complexes from structures of non-native decoys, and it consistently outperforms lower resolution functions (22) in decoy discrimination tests, strongly suggesting that the increased atomic detail of our function results in its improved performance. We further show that our all-atom potential derives most of its discriminatory power from the information contained in short-range atomic interactions, and that it performs competitively with a physical

potential function (28) in a test of structure-based binding sequence identification using refined decoy structures. These results suggest that high-resolution statistical potentials can achieve decoy-discrimination capabilities on par with more complex physical models of protein-nucleic acid interactions.

## 3.2 Methods

### 3.2.1 Structure Set Selection

Crystal structures of protein-DNA complexes with resolution better than 2.5Å were selected from the Nucleic Acid Database (162). Complexes containing single-stranded DNA were discarded, as were complexes containing chemically-modified nucleotides, or DNA structures that significantly deviated from ideal, double-helical geometry (i.e. flipped bases, bulges, loops, higher-order helices and structures with DNA axial bends of more than $90^{o}$). Structures of significant protein sequence homology were filtered by conducting pairwise sequence alignments using BLAST (163), and discarding the lower-resolution structure of all pairs with BLAST expectation values of less than 0.001. Structures containing significant regions of missing crystallographic density at the protein-DNA interface were also discarded. The composition of the final structure set is available as supplementary material.

For all remaining structures, regions of extra crystallographic density were removed by selecting coordinates with highest occupancy, or in the case of equal-occupancy conformations, by selecting the first conformation listed. If multiple symmetric complexes were found in a structure, only one complex was considered. Finally, water molecules, metal ions, hetero-atoms and other non-protein, non-DNA molecules were discarded, and the net-bending ($\Gamma$) angles of the base steps in each structure were calculated using the 3DNA package (164).

### 3.2.2 Distance-Dependent Statistical Potentials

#### 3.2.2.1 General Form

A total of four distance-dependent conditional potential functions were tested for their ability to identify native protein-DNA complex structures. All of these functions can be expressed in terms of the conditional probability formalism described by Samudrala and Moult (14). Briefly, this formalism describes a form of naïve Bayes' classifier (165), which attempts to assign a candidate molecular system into a "correct" (native-like) or "incorrect" class, based upon properties observed within the set of atomic contacts in the complex.

The potentials used in this work all attempt to classify protein-nucleic acid complexes based on the set of atomic distances, $D = \{d_{ij}\}$, observed at the molecular interface. A principal assumption is that the free energy of the protein-nucleic acid complex, $G$, may be estimated by the probability of "correctness" given by the potential function:

$$G \approx -\ln P(C \mid D) = -\sum_{i}^{N_P}\sum_{j}^{N_D} \ln P(C, t_i, t_j \mid d_{ij}) \qquad (3.1)$$

Here, $d_{ij}$ represents the distance between two interface atoms, and $t_i$ and $t_j$ are the chemical types of these atoms, as determined by a mapping function, $\tau(x)$, described below. $N_P$ and $N_D$ represent the number of protein and DNA atoms in the complex, respectively. A simplifying assumption inherent to the naïve Bayes' classifier is that the atomic distances in a given structure are independently distributed, and therefore, the probability of "correctness" of a complex is assumed to be the joint probability of "correctness" of every atomic contact in the protein-DNA interface.

The probability of an individual atomic contact is expressed in terms of the likelihood of observing a separation $d_{ij}$ between atoms of type $t_i$ and $t_j$ in a native-like protein-DNA complex:

$$P(C,t_i,t_j \mid d_{ij}) = P(C,t_i,t_j) \cdot \frac{P(d_{ij} \mid C,t_i,t_j)}{P(d_{ij})} \qquad (3.2)$$

The Bayesian prior $P(C, t_i,t_j)$ represents the *à priori* probability of observing a "correct" protein-DNA contact between atoms i and j. Because of the difficulty in determining the value of this constant, it is treated as 1 for this study (however, we note that this omission results in a score that is not a true assessment of probability, but rather, a normalized likelihood of classification). The differences between the potentials used in this work lie in the form of the other components of this equation – the likelihood function, $P(d_{ij} \mid C,t_i,t_j)$, the marginal probability, $P(d_{ij})$, and the definition of the mapping function, $\tau(x)$, used to determine the atomic types $t_i$ and $t_j$.

### 3.2.2.2 Likelihood Function

For all of the tested potentials, the likelihood of $d_{ij}$ is estimated by the frequency with which atoms of types $t_i$ and $t_j$ are observed to be separated by a distance less than or equal to $d_{ij}$ in the training set of known protein-DNA structures:

$$P(d_{ij} \mid C,t_i,t_j) \approx f(d_{ij} \mid t_i,t_j) = \frac{N_{obs}(d_{ij},t_i,t_j)}{\sum_{d_{ij}} N_{obs}(d_{ij},t_i,t_j)} \qquad (3.3)$$

where $N_{obs}(d_{ij},t_i,t_j)$ represents the number of contacts observed between two atoms of type $t_i$ and $t_j$, that are separated a distance of $d_{ij}$ in the training set. This function maps the continuous value $d_{ij}$ to a set of discrete distance bins, $\{b_0,b_1,\ldots,b_n\}$, where the number of bins and their distance cutoff values, $\{d_{b_0},d_{b_1},\ldots,d_{b_n}\}$, are free parameters of the model. A count is assigned to distance bin $b_i$ if $d_{b_{i-1}} \le d_{ij} < d_{b_i}$. Atoms separated by distances greater than or equal to $d_{b_n}$ are not counted.

### 3.2.2.3 Marginal Distribution of Atomic Distances

For those potentials using the DFIRE reference state of Zhang *et al.* (22), the denominator of equation 2.2 is defined as:

$$P(d_{ij})_{DFIRE} = \left(\frac{d_{ij}}{d_{b_n}}\right)^{1.61}\left(\frac{\Delta(d_{ij})}{\Delta(d_{b_n})}\right)N_{obs}(d_{b_n},t_i,t_j) \qquad (3.4)$$

where $\Delta(d_{ij})$ is the width of the bin encompassing distance $d_{ij}$, and $N_{obs}(d_{b_n},t_i,t_j)$ and $\Delta(d_{b_n})$ are the number of counts in the final distance bin, and the width of that bin, respectively. For a derivation of this formula, refer to work of Zhou *et al.* (18).

For all other potentials, the marginal probability of intermolecular atomic distances is assumed to be approximated by the frequency of separation $d_{ij}$ between any two atom types in the training set of known protein-DNA complexes:

$$P(d_{ij}) \approx f(d_{ij}) = \frac{\sum_{t_i}\sum_{t_j} N_{obs}(d_{ij},t_i,t_j)}{N_c} \qquad (3.5)$$

where $N_c$ is the total number of contacts observed between interface atoms of all types, at all distances, in the training set.

### 3.2.2.4 Atomic Type Mapping

Two mapping functions, $\tau(x)$, were used to determine the atom types for the statistical potentials tested. For all-atom potentials, every protein and nucleic acid heavy atom type was considered in a residue-specific manner (e.g. alanine $C_\alpha$ was a different atom type than leucine $C_\alpha$, and adenine C1' was a different atom type than guanine C1'). For reduced-atom potentials, the atom types defined by Zhang *et al.* (22) were used, with the exception that all non-protein, non-DNA atom types were discarded.

### 3.2.2.5 Low-Counts Correction

For all potentials that did not use the DFIRE reference state, a low-counts correction was performed using the method of Sippl (13). Briefly, the frequencies calculated using equation 2.3 were modified according to the expression:

$$f(d_{ij},t_i,t_j)_{corrected} \equiv \frac{f(d_{ij}) + \sigma N_{obs}(d_{ij},t_i,t_j)\cdot f(d_{ij},t_i,t_j)}{1+\sigma N_{obs}(d_{ij},t_i,t_j)} \quad (3.6)$$

Thus, the corrected frequencies are close to $f(d_{ij})$ when few counts are observed in the structure database, but approach $f(d_{ij},t_i,t_j)$ as $N_{obs}(d_{ij},t_i,t_j)$ becomes large. The value of $\sigma$ (set to 1/50 in this work as per Sippl (13)) ensures that the terms have equal weight when $N_{obs}(d_{ij},t_i,t_j) = 1/\sigma = 50$.

### 3.2.3 Docking Protein-DNA Complexes

Protein-DNA complex decoy structures were produced using the FTDock rigid-body docking package, as described by Aloy *et al.* (68). Briefly, the protein and DNA structures from each training complex were separated, and used to conduct a global search of possible intermolecular conformations, optimizing a scoring function that selects for molecular surface complementarity. The larger molecule of each protein-DNA system was held fixed at the system origin, and the smaller molecule from the complex allowed to move independently. Default FTDock scoring and search parameters were used for all simulations, with electrostatic screening enabled. Seven of the protein-DNA complexes in the training set were too large to dock given available computational resources. The 10,000 top-scoring decoys from each global docking simulation were retained.

In order to create near-native docking decoy structures for a given complex, a modified FTDock software package was used to sample rotations in $1^o$ increments within $40^o$ of the rotations specified for the three best (lowest-RMSD) decoy structures found using the standard FTDock docking algorithm for that complex. At most, 2000 top-scoring decoys were retained for every complex.

For all docking decoy structures, the $C_\alpha$ RMSD to the native complex (after superimposing the native and decoy DNA structures) was computed for each decoy, as well as the percentage of correct contacts (%CC), using the method described by Aloy *et al.* (68).

### 3.2.4 Decoy Scoring

The 2,000 lowest-RMSD decoys for every docked complex were scored using the distance-dependent potential functions described above. Two variants were computed for each score – a "fair" score that omitted native atomic contacts from the training set (i.e., using contact data from 51 structures), and an "unfair" score that included contacts from all 52 training complexes. Additionally, in order to assess the potentials' sensitivity to a reduction in training data, experiments were performed using contact data assembled from subsets of the training complexes. Five randomly-selected sets of contact data were assembled for each of four (80%, 85%, 90% and 95%) subsets of the structure training set.

The ability of the potentials to discriminate the native complex from docking decoys was determined by computing the Z-score of the native structure relative to all scored decoys (in some cases, the Z-score of the best decoy in the set was also calculated, in order to facilitate comparisons with the FTDock scoring function). Near-native decoy enrichment was determined as the frequency ratio:

$$E = \frac{f(good \mid top)}{f(good)} \qquad (3.7)$$

where $f(good)$ is the frequency of "good" decoys in the decoy set (and "good" is defined as specified in the text), and $f(good|top)$ is the frequency of "good" decoys in the best-scoring 1% of decoys in the decoy set. All other statistical analysis techniques were performed using standard methods.

### 3.2.5 DNA-Binding Site Mutation

The ability of the distance-dependent potentials to identify sequence-specific DNA binding was assessed by evaluating the "fair" potentials for the native complex, as well as for a library of complexes representing non-cognate DNA binding sites. Alternate DNA sequences were constructed for every complex in the training set by swapping idealized bases onto the backbones of the native DNA structures, preserving the native $\chi$ angles, and rotating the mutated bases as appropriate to produce ideal bond angles at the glycosidic nitrogen atoms.

## 3.3 Results

### 3.3.1 Development of an All-Atom Statistical Potential Function

We have developed an all-atom, distance-dependent statistical potential function based on the naïve Bayes classifier (165). In this model, the likelihood of "correctness" of a protein-DNA complex is determined from the set of atomic distances observed across the molecular interface. The form of the potential is similar to that described by Samudrala and Moult for protein structure prediction (14), but intermolecular distances are used in place of intramolecular distances to score each structure.

This potential expresses the likelihood of observing a set of intermolecular atomic distances between the protein and DNA molecules of a complex, and assumes that this likelihood approximates the binding free energy for the molecules. The naïve Bayes classifier is based on the assumption that all features used for classification are independently distributed, and therefore, it is a computationally-efficient, pairwise-additive score. In order to assess the origin of the discriminatory power, we have tested several variants of the score: the all-atom function, a reduced-atom version of the potential (where atoms are grouped according to a small set of chemically-related atom types), as well as reduced-atom and all-atom variants using the reference state proposed by Zhang *et al.* (22).

### 3.3.2 Discriminating Native Protein/Nucleic-Acid Complexes from Decoy Structures Generated with Rigid-Body Docking Methods

Ideally, a scoring method capable of recognizing correctly bound, protein-DNA complexes should also be able to discriminate a native complex from structures that deviate significantly from it. Thus, we tested the ability of our score to discriminate native protein-DNA complexes from a pool of "decoy" complex structures generated using the FTDock rigid-body docking algorithm (Figure 3.1) (68). Here, 2000 docking decoys generated for each of 45 protein-DNA complexes in the training set (seven complexes were too large to dock given available computer resources) were scored with our statistical potential, and these scores were used to rank the decoy complexes, as well as their corresponding native structures.

In order to optimize the free parameters of the scoring function, we constructed four different scores, and compared them over a range of parameters. The tested scores differ in two critical areas: atomic resolution, and choice of reference state. Half of the scores use an all-atom formalism, while half use a reduced-atom representation, wherein protein and DNA atoms are mapped onto 19 pseudo-chemical atom types. Furthermore, half of the scores use the "uniform" reference state of Samudrala *et al.* (14), while half use the "DFIRE" reference state described by Zhang *et al.* (22). For all scores, the chosen parameter set is represented by three numbers, where the first number is the cutoff for the first distance bin, the second is the cutoff for the last bin (the overall distance cutoff), and the third represents the width of each remaining distance bin (after the first). Thus, the parameter triple 3/10/1 reflects a scoring function that considers all atomic contacts separated by less than 10Å, and groups these contacts into eight bins beginning with a 3Å bin, followed by seven 1Å bins.

The all-atom, distance-dependent scoring function produces the greatest overall separation between the scores of decoys and their corresponding native structures. The average native Z-score for this method is -6.8 for the best parameterization (Figure 3.1a, 5/10/1), while the equivalent reduced-atom function results in a mean native Z-score of -4.8 for its best parameterization (Figure 3.1b, 3/10/1). Introducing the DFIRE reference

state to either of these scores reduces their overall performance, producing best average Z-scores of -3.8 (Figure 3.1c, 4/20/4) for the all-atom DFIRE function, and -4.6 (Figure 3.1d, 4/20/4) for the reduced-atom DFIRE function. These values are significantly different from one another at the 99% confidence level (Welch's t-test), with the exception of the best mean Z-scores for the all-atom DFIRE and the reduced-atom DFIRE functions, for which equality cannot be rejected with sufficient confidence ($p = 0.08$, Welch's t-test).

The performance of the all-atom and reduced-atom functions depends on the choice of distance cutoff used to define a contact. Both of these functions perform best when the maximum distance cutoff is 10Å (Figure 3.1a, b; 3/10/1, 4/10/1, 5/10/1), and exhibit a consistent decline in performance as this cutoff value is increased (Figure 3.1a, b; 3/10/1, 3/15/1, 3/20/1). In addition, increasing the width of the initial distance bin has a slight negative impact on the performance of the reduced-atom potential, lowering the magnitude of the mean native Z-score as the width of the bin is increased (Figure 3.1b, 3/10/1, 4/10/1, 5/10/1). Using a one-way analysis of variance (ANOVA), we find both of these trends to be significant at the 95% confidence level. In contrast, addition of the DFIRE reference state to the all-atom and reduced-atom functions virtually eliminates the impact of parameter choice on the scores' performance. Neither the all-atom DFIRE (Figure 3.1c) nor the reduced-atom DFIRE (Figure 3.1d) potentials display significant changes in mean native Z-score in response to changes in bin width or distance cutoff (as determined using one-way ANOVA, $p \geq 0.05$).

### 3.3.3 Discriminating Native Structures from Near-Native Docking Decoys

The previous experiments have demonstrated the all-atom function's ability to discriminate near-native complexes from a set of mixed-quality decoy structures. However, a more stringent test of performance is provided by examining the function's ability to discriminate between near-native decoy complexes. The results of such a test are shown in Figure 3.2. In this experiment, the FTDock protocol was modified to produce decoy structures that were slight perturbations on the best docking decoys

produced in a normal FTDock run. This technique resulted in a set of "refined" docking decoys for each complex, each set containing at least 175 decoys closer than 5Å RMSD to their respective native structures (approximately 500 decoys, on average), and 10 decoys closer than 1Å RMSD (approximately 40 decoys, on average). These decoys were scored using the two best-performing techniques from the previous experiment (all-atom, 5/10/1 and reduced-atom, 3/10/1), as well as the score produced by FTDock.

In this experiment, the differences between the all-atom and reduced-atom scores are less pronounced, though both have significantly more discriminatory power than the FTDock score alone. a shows the Z-Score of the lowest-RMSD (best) decoy for each refined decoy set (in lieu of the native Z-score, as FTDock does not produce a score for the native structure). The observed mean Z-scores are significantly different for the all-atom, reduced-atom and FTDock methods, ($p < .001$, Welch's t-test), with the reduced-atom score providing the greatest discrimination (Z = -2.08), followed by the all-atom score (Z = -1.64), and the FTDock score (Z=-0.48). When score-RMSD correlation is considered (b), the mean correlation coefficient for the all-atom score ($r = 0.64$) is insignificantly ($p = 0.15$) greater than that of the reduced-atom score ($r = 0.57$), while both are significantly ($p < 0.001$) greater than the mean correlation observed using the FTDock method ($r = 0.32$).

Using the metric of native-like decoy enrichment (Figure 3.2c), the all-atom score has a significant advantage over the reduced-atom score. Here, "enrichment" is calculated as the percentage of decoy structures with ≥95% correct intermolecular contacts (as defined by Aloy *et al.* (68)), in the top-scoring 10% of all decoys in each set. The all-atom score produces the greatest mean enrichment in this experiment (E = 9.10), followed by the reduced-atom score (E = 7.36), and the FTDock score (E = 5.34). Thus, on average, the all-atom score identifies 91% of decoys with ≥95% correct intermolecular contacts within the best 10% of decoy structures in their respective decoy sets. This is significantly better than the performance of either the reduced-atom ($p = 0.002$) or the FTDock score ($p < 0.001$) in this test.

### 3.3.4 Examples of Decoy-Discrimination Performance

Native Z-scores measure the ability of a score to discriminate native structures from incorrect conformations, and provide a one-dimensional representation of scoring function performance. To give a more complete picture of the performance of the all-atom score at decoy discrimination, we show three representative decoy sets in Figure 3.3, Figure 3.4 and Figure 3.5. In these, the 5/10/1 parameterization of the all-atom score is plotted versus native RMSD for the docking decoys, chosen as examples of good, average, and poor native structure discrimination, respectively.

A particularly successful example of discrimination by the all-atom scoring function is shown in Figure 3.3a, for the I-CreI homing endonuclease (pdb id: 1g9z). Here, three decoy structures have similar scores to the native complex, each with scores that are significantly separated from the larger pool of decoys. This separation produces a native Z-score of -14.3, the best observed for any complex in the test set. The similarity of these three decoys to the native structure is illustrated in Figure 3.3b, where the protein backbones are rendered for each of the decoys, and superimposed on the native DNA structure. These decoys were used as starting structures to create the near-native docking decoy set for this structure, shown in Figure 3.3c. Here, a graphical interpretation of the decoy enrichment metric is provided, with the most native-like decoys (those decoys with ≥95% correct contacts) plotted in green, and the tenth-percentile score cutoff shown as a dashed line. In this case, all of the native-like decoys score within the tenth percentile, a decoy enrichment of 10.0. A graphical representation of the score-RMSD correlation is also provided in Figure 3.3c, where the line of best-fit for all decoys better than 5Å RMSD to the native complex structure is shown in gray. In this example, the score-RMSD correlation is not strong ($r = 0.22$), despite the large score separation between the near-native and low-resolution decoy structures.

A more typical example of scoring function performance is shown in Figure 3.4a, for decoys of the avian myeloblastosis virus v-Myb structure in complex with its target promoter sequence (pdb id: 1h8a). The separation between the native structure and the decoy pool is less dramatic in this example (native Z-score: -6.24), and the scores plotted

in Figure 3.4a appear to correlate with RMSD at decoy RMSD values below approximately 10Å. The near-native decoy set for this complex (shown in Figure 3.4b), illustrates a common pathology of the scores we have tested, in that the native structure is frequently not the top-scoring structure of a given decoy set. In this example, the native structure score is higher than those of several of the near-native decoy structures, as well as those of a few decoys worse than 2Å RMSD from the native complex. The structures of the near-native decoys (illustrated in Figure 3.4c) are tightly clustered, suggesting that the differences between these decoys may be too subtle to be captured by a statistical score based on 1Å bins. The lowest-scoring, false-positive decoys (shown in Figure 3.4d), are also tightly clustered, but are uniformly nearer to the DNA structure than is the native protein. This may indicate that the wide initial bin of our potentials (5Å, in this example) does not result in a score that adequately penalizes atom-atom clashes, leading to higher scores for excessively-packed decoy structures.

The final decoy-discrimination example (Figure 3.5), shows decoy scores for a TATA-binding protein from *Arabidopsis thaliana* in complex with an engineered DNA-recognition site (pdb id: 1qna). This set is interesting because, although the native complex is separated from the rest of the docking decoys (native Z-score: -4.95), there is a spike of low-scoring decoys with RMSD values of more than 25Å from the native complex structure (the lowest-scoring example of which is illustrated in Figure 3.5d). All of these false-positive decoys are inversions of the native binding geometry, and are deceptively visually similar to the native complex structure (shown in Figure 3.5c). In this example, the native structure is distinguishable from the best-scoring, false-positive decoy only through the ordering of the strands in the β-sheet, and a slight bend in the C-terminal helix of the protein structure.

When considering the near-native decoys for this complex (Figure 3.5b), the performance of the score is mixed. The score-RMSD correlation is strong ($r = 0.73$), but the native structure scores worse than all of the nearest-native decoys (those decoys with ≥95% correct contacts). Moreover, a number of the nearest-native decoys are not within the tenth percentile of decoy scores, resulting in a near-native decoy enrichment of only

6.6, one of the lowest in our test set. Together, these metrics highlight the difficulty of successfully recognizing the correct binding conformation of a minor-groove binding protein.

### 3.3.5 Determining the Impact of Training Set Composition on Score Performance

Despite significant recent growth, there are still relatively few high-resolution structures of protein-DNA complexes in the public repositories. Thus, there is a legitimate concern that the size and composition of the training set biases the performance of the scores tested here. In order to investigate the impact of training-set size and composition on score performance, we constructed scores for randomly-chosen 95%, 90%, 85% and 80% subsets of the training set, and used these subsets to repeat the native discrimination experiments discussed above. Five replicate experiments were performed for each subset size, using a different randomly-chosen training data set for each replicate score (i.e., five replicates per set size, for a total of 20 random tests per score).

The results of this analysis are shown for the all-atom and the reduced-atom DFIRE scores in Figure 3.6. Here, each point represents the mean native Z-score produced by averaging the native Z-scores found for the 45 protein-DNA docking decoy sets. Thus, four columns are shown, each containing five points (one per replicate experiment) for the all-atom score, and five points for the reduced-atom DFIRE score. The means of these data points (i.e. the super-mean for the 45 decoy sets) are connected by dashed lines, to guide the eye. In order to compare the all-atom and reduced-atom DFIRE scores on a common axis, the maximum data point for each score was placed at the plot origin, and the remaining data points drawn relative to these values. Thus, the y-axis of Figure 3.6 shows the change in mean native Z-score observed over the test set, as a function of the size of the training set used to derive the score.

This figure demonstrates that the performance of the reduced-atom DFIRE score is virtually independent of the size of the training set. In contrast, the all-atom potential has a clear dependence on training set size, with an average change in mean native Z-score of

nearly -0.4 as the size of the data set is increased from 80% to 95%. In addition, the reduced-atom DFIRE mean Z-scores are more tightly clustered than the all-atom mean Z-scores, indicating that the performance of the all-atom score is sensitive to the size and the exact composition of the training set used to derive the score, whereas the reduced-atom DFIRE score is relatively insensitive to these factors.

### 3.3.6 Exploring the Relative Contributions of Nucleic Acid Base and Backbone Contacts

We investigated the relative contributions of protein contacts to the DNA backbone and bases by creating versions of the all-atom score using two subsets of the training data — one consisting exclusively of protein-DNA backbone atom pairs, and another of protein-DNA base atom pairs. We then used these functions to re-evaluate native Z-scores and score-RMSD correlations for the near-native docking decoys used to create . The results of this test are shown in Table 3.1.

We find no significant differences (for either the mean native Z-scores or mean score-RMSD correlation coefficients) between the backbone-only score, the base-only score and the full, all-atom scoring function. The all-atom score most frequently yields the greatest native Z-score (31 out of 45 structures), but, the backbone-only score most frequently yields the greatest score-RMSD correlation coefficients (28 out of 45 structures). There is a very poor correlation of the native Z-score and score-RMSD correlation data with mean $\Gamma$ angle (a quantitative measurement of DNA base-step deformation (164)), suggesting that for docking decoy discrimination, the performance of the all-atom score is independent of DNA bend. Because the all-atom score does not consider intramolecular interactions, however, it is not safe to assume that these results reflect indirect recognition phenomena within the tested protein-DNA complexes.

### 3.3.7 Examining the Relationship Between Score Form and Function

The all-atom scores for the interaction of the arginine guanidinium group with the guanine base (the most prevalent base-amino acid contact in our data set) provide insight as to the origin of the superior performance of the all-atom potential, compared to

reduced-atom representations. Figure 3.7a shows the scores for interactions of the arginine NH1 and NH2 nitrogen atoms with guanine O6 oxygen, while Figure 3.7b shows the scores for these arginine atoms with guanine N7, and Figure 3.7c, the scores for the interaction of arginine $C_\zeta$ with the same base atoms. All of the scores are represented as plots relative to interatomic distance, where a given score value is assumed to be centered on the upper distance cutoff of each distance bin.

In Figure 3.7a, the interaction between arginine NH2 and guanine O6 has a score minimum in the 3Å distance bin, while the corresponding minimum for the interaction between arginine NH1 and guanine O6 is in the 5Å bin. This is consistent with the scores observed in Figure 3.7b, which show score minima in the 4Å bin for the interactions of arginine NH1 and NH2 with guanine N7. Together, these scores suggest a conformation wherein arginine NH2 is placed nearest to guanine O6, while arginine NH1 and NH2 are roughly equidistant to guanine N7. This interpretation is consistent with the score data in Figure 3.7c, which shows that arginine $C_\zeta$ has a minimum score when placed at a distance of approximately 4Å from both guanine N7 and guanine O6 (for illustration, an interaction geometry that meets these distance constraints is provided as an inset in all three panels). While it is difficult to make a single structural argument using scores representing the geometries of a number of different arginine-guanine interactions, it is clear that the all-atom scores are able to capture atom-specific distance preferences that the equivalent reduced-atom scores (shown as dashed lines) cannot. The reduced-atom methods' treatment of structurally-distinct atom types (such as arginine NH1 and NH2, which have a chiral relationship with arginine $N_\varepsilon$) as identical atom types restricts their ability to represent atomic detail.

Further insight into the advantages of the all-atom representation is provided in Figure 3.8. Here, all-atom scores are shown for interactions between asparagine and adenine (Figure 3.8a,b) as well as the "chemically-equivalent" interactions between glutamine and adenine (Figure 3.8c,d). In a, scores for the interaction of asparagine $O_{\delta 1}$, $C_\gamma$, and $N_{\delta 2}$, with adenine N6 are shown. There is a clear minimum in the 5Å distance bin for asparagine $N_{\delta 2}$, as well two clear minima (3Å and 6Å) for asparagine $O_{\delta 1}$, and a broad

minimum between 4 and 5Å for asparagine $C_\gamma$. However, no equivalent minima are observed for glutamine $O_{\varepsilon 1}$, $C_\delta$ or $N_{\varepsilon 2}$ (Figure 3.8c), despite their obvious chemical similarity to the asparagine atom types used to create Figure 3.8a (which would lead to the equivalent treatment of these atoms in a reduced-atom representation). The interaction scores for asparagine $O_{\delta 1}$, $C_\gamma$ and $N_{\delta 2}$ with adenine N7 (Figure 3.8b) are more similar to those of glutamine $O_{\varepsilon 1}$, $C_\delta$ and $N_{\varepsilon 2}$ (Figure 3.8d), but are still notably different.

### 3.3.8 Discriminating Cognate DNA-Binding Sites from Non-Cognate Sequences

Although docking-decoy discrimination is a useful test of performance, one of the most relevant and difficult applications of any scoring function for protein-DNA structures is the identification of the DNA recognition sequence of a given protein structure. In Figure 3.9, we investigate the ability of the statistical potentials to discriminate cognate binding sites from a large pool of non-cognate DNA sequences. This test set was created by swapping nucleotide bases onto the complexes' native DNA backbones. These altered DNA structures were scored with the four statistical potentials, using the parameter combinations that produced the greatest native Z-scores in the docking tests, discussed previously. The Z-score of the cognate DNA recognition sequence was subsequently calculated relative to the 10,000 randomly-chosen, non-cognate sequences. Thus, each box plot in Figure 3.9 represents the native Z-scores for the 45 proteins in the structure test set, as determined using the scores listed at the bottom of the figure.

The all-atom score exhibits the greatest discriminatory power in this test, producing the largest mean separation between the scores of the cognate DNA binding sites for the structure set, and the pools of non-cognate decoy sequences for these structures. Every complex scored with the all-atom potential had a negative cognate sequence Z-score, indicating that the scores of the native DNA sequences for these structures were at least nominally below the mean scores of their respective decoy pools. In contrast, the reduced-atom, all-atom DFIRE, and reduced-atom DFIRE scores each had a significant percentage of complexes with positive cognate sequence Z-scores. This suggests that the

all-atom potential function is both more sensitive and more reliable than the other tested potential functions when discriminating native DNA-binding sequences. Nevertheless, we note that the magnitude of the Z-scores in this experiment are lower than those observed for the docking decoy discrimination tests shown in Figure 3.1, reflecting the greater difficulty of the test.

In order to address the possibility that non-native interactions in mutant complexes that were not energy minimized had biased the test results, we obtained the set of EcoRI sequence mutants used by Havranek *et al.* to test the ability of the Rosetta physical potential function to discriminate between cognate and non-cognate protein-DNA binding sites (28). These mutant protein-DNA complexes were prepared in a manner similar to that described above, but protein side-chain positions were optimized against the mutant DNA structures using the Monte Carlo simulated annealing method of Kuhlman *et al.* (63). We evaluated these repacked mutant (and native) structures with the all-atom potential function, and compared the all-atom scores with the corresponding Rosetta scores for each complex. The results of this analysis are shown in Table 3.2.

Havranek *et al.* computed the scores for ten repacked variants of the cognate EcoRI complex, as well as for ten repacked structures for every possible palindromic DNA sequence mutation. Thus, the EcoRI decoy set consists of 630 structures, ten of which are repacked variants of the native structure itself. Table 3.2 shows the ranks and Z-scores of these repacked native structures relative to the 620 repacked mutant structures in the decoy set. The all-atom score is a significantly better discriminator than the Rosetta score, with a mean Z-score of -2.22 relative to a mean of -1.58 for the Rosetta method ($p<0.01$, Welch's t-test). However, the mean rank of the native structures assigned by the all-atom score (17.7) is worse than the mean rank assigned by Rosetta (4.90). Thus, while the all-atom potential produces a greater separation between the scores of the cognate and non-cognate complexes than does the Rosetta score, it does a poorer job of ranking the decoy structures. Examining these scores directly, we find that the all-atom potential scores several two-base mutants (e.g. one mutation of the palindromic EcoRI recognition sequence) with slightly lower energies than the cognate recognition sequence. This

difference is sufficient to explain the discrepancy between the Z-Score and rank data observed here.

## 3.4 Discussion

The structure-based identification of transcription-factor binding sites in genome sequences, and the rational design of novel, nucleic-acid binding proteins require potential functions that can accurately predict protein-nucleic acid interactions. Despite their wide application to protein structure prediction, docking and other areas of structural biology (13-21, 24, 25, 56, 57, 166-168), statistical potentials have only recently been applied to this problem (1, 22, 26, 59, 60). Thus, there are many open questions concerning their application, including the most appropriate level of score resolution, the optimal definition of an atomic contact and choice of reference state. In order to address these questions, we have developed an all-atom, distance-dependent potential function for protein-DNA interactions, and have tested this method extensively using multiple decoy-discrimination experiments of increasing difficulty. Direct comparisons of this potential to several similar, but lower-resolution potentials demonstrate that the all-atom formalism significantly improves the potential's ability to discriminate native-like protein-DNA complexes from non-native structures.

### 3.4.1 The Increased Atomic Resolution of the All-Atom Score Improves its Decoy Discrimination Performance

Given the limited number of high-resolution protein-DNA complexes available, we were concerned that the performance of the all-atom statistical potential would be limited by increased statistical noise due to sparse training data. However, in tests of docking decoy discrimination (Figure 3.1), we find that the all-atom potential produces the greatest overall native Z-scores, followed by the reduced-atom potential, with the reduced-atom DFIRE and all-atom DFIRE scores performing worst in our tests. It is somewhat paradoxical that the all-atom formalism performs worst when the DFIRE reference state is used, and best when it is omitted, but we believe that this reflects the

greater sensitivity of the all-atom scores to the choice of reference state (due to the greater number of scores that must be trained with a given quantity of structural data).

In tests of near-native docking decoy discrimination (Figure 3.2), we find that the all-atom potential is best able to enrich decoy sets for native-like decoy structures (Figure 3.2c), and rank docking decoys according to their RMSD from the native structure (Figure 3.2b). Interestingly, the reduced-atom score produces a slightly lower mean native Z-score than does the all-atom function in this experiment (Figure 3.2a), but this result is inconsistent with all other decoy discrimination tests we have conducted.

These results strongly suggest that the performance advantages of the all-atom potential outweigh the disadvantage of statistical noise associated with the increased number of trained parameters. We find that the low-counts correction (13) is important for the performance of our scoring method, reflecting the small size of the training set. However, the all-atom potential's superior performance in our experiments, despite the limited size of the structure set used to train the potential, suggests that model detail is of greater importance than statistical quality when developing scores for protein-DNA systems.

To this end, we have explored the impact of training-set size on score performance in Figure 3.6. The all-atom potential performs best in these tests, but is sensitive to the composition and size of the training set. In contrast, the reduced-atom DFIRE score (chosen here for its use of the DFIRE reference state, which has been shown to be less sensitive to database size and composition (58)) is relatively insensitive to the quantity of data used for training, but has consistently worse native-discrimination performance. Thus, the choice of reference state appears to add robustness to the potential when the training set is small, but at the cost of reduced discrimination. The all-atom score also exhibits greater per-experiment variation than the reduced-atom DFIRE score, indicating that the former method is more sensitive to the exact composition of the structure training set.

It should be noted that we have purposely chosen a training set that is minimal in size (i.e. strictly non-redundant in sequence), in order to conservatively estimate score

performance. Including additional data in the training set by relaxing sequence-identity cutoffs increases score performance of the all-atom score in many cases (a finding consistent with the work of Velec *et al.* for protein-small molecule docking (23)). Because of this, as well as the demonstrated sensitivity of the all-atom score to training-set size and composition, we are optimistic that the performance of the method will increase as new structures are generated.

### 3.4.2 Short-Range Atomic Interactions Are Most Important for Recognition Specificity

As shown in Figure 3.1, we find that the decoy-discrimination performance of the all-atom and reduced-atom scores decreases as the distance cutoff used to define a contact is increased. It seems likely that extending the definition of atomic contacts to include those atoms separated by more than 10Å only increases the number of non-specific interactions in the training set. This finding is consistent with the results of Liu *et al.* (59) for protein-DNA systems, but is opposite of the conclusion of Samudrala and Moult, who found instead that an increased distance cutoff allowed for improved discrimination of protein structure decoys (14). This result provides evidence that protein-DNA recognition specificity is primarily a short-range molecular phenomenon, with long-range interactions contributing less significantly to sequence discrimination (although not necessarily to overall binding energy).

### 3.4.3 The All-Atom Score Performance Is Unaffected by DNA Deformation

In order to explore the degree to which the all-atom score is sensitive to DNA deformation (i.e. deviation from ideal, B-form DNA structure), we calculated native Z-scores and score-RMSD correlations for the near-native docking decoy set using backbone-only and base-only versions of the all-atom score (Table 3.1). We then compared these values to the mean net-bending angle ($\Gamma$ – a quantitative description of DNA base-step deformation (164)) for each complex in our test set.

We observe no significant correlation between the native Z-score or score-RMSD correlation data and mean $\Gamma$ for any version of the all-atom score, suggesting that the

decoy-discrimination performance of the score is insensitive to DNA deformation. However, we observe that the backbone-only and base-only versions of the score have approximately equivalent mean performance under either metric. This result suggests that for rigid-body docking decoy discrimination, it may be sufficient to recognize shape complementarity (which can be evaluated using DNA backbone or base contacts alone), and that the all-atom potential is capturing this information in a relatively high-resolution manner (as opposed to FTDock's grid-based score, which performed significantly worse than the all-atom score on this decoy set, as shown in ).

### 3.4.4 The All-Atom Score Formalism Can Capture Specific Intermolecular Interactions

Analysis of atomic interactions from the all-atom and reduced-atom scores indicates that the superior performance of the all-atom representation possibly derives from its ability to capture fine details of intermolecular interactions (Figure 3.7 and Figure 3.8). In the example shown in Figure 3.7, the all-atom potential appears to capture a chiral preference for the interaction of arginine with the major groove face of guanine, wherein arginine NH1 favors placement near to guanine N7, and arginine NH2 favors placement near to guanine O6. The argument for this geometry is compelling, since the guanidinium group is chiral, presenting a hydrogen-bond donor (arginine $N_{\varepsilon}$) *cis* to the NH2 atom. This chiral preference cannot be observed in the equivalent reduced-atom scores for the same interactions, because the reduced-atom mapping treats the arginine NH1 and NH2 atoms as a single chemical type (Np13).

Justification for the use of residue-specific atom types is provided in Figure 3.8. Here, all-atom scores for the interaction of two chemically-equivalent sets of protein side-chain atoms (asparagine $O_{\delta 1}$, $C_{\gamma}$, $N_{\delta 2}$; glutamine $O_{\varepsilon 1}$, $C_{\delta}$, $N_{\varepsilon 2}$), are shown with the nucleotide base atoms adenine N6 and N7. These atoms would be treated as identical in virtually any reduced-atom representation where atoms are mapped to "chemical" types, but have clearly different contact preferences in the all-atom scores shown here.

Examples such as these imply that the use of residue-specific atom types allows for the capture of structural preferences that would be missed by less detailed formalisms.

### 3.4.5 The All-Atom Score Can Predict Cognate Protein-Binding Sequences from Structure

Many applications of potential functions for protein-nucleic acid complexes (*e.g.* structure-based detection of protein binding-sites in DNA sequences) require the ability to recognize native-like protein-DNA interactions based on subtle differences in interface geometry. We investigated the ability of our all-atom method to discriminate these subtle chemical differences by evaluating the scores for a large set of sequence-mutant decoys, generated by computationally threading 10,000 random DNA sequences onto the DNA structures of each complex in the test set. The results for this experiment (Figure 3.9) show that the all-atom potential is the only potential we have tested that is capable of detecting the cognate binding sequences for every protein in the test set. All of the other potentials had at least one complex where the native DNA sequence scored worse than the mean score for the 10,000 decoy sequences.

The cognate Z-scores for this experiment are lower in magnitude than the native Z-scores observed for the near-native docking decoy discrimination test (), which are themselves lower in magnitude than the native Z-scores observed in the initial decoy-discrimination tests (Figure 3.1). Thus, the progressive decrease in discriminatory power reflects the progressive increase in difficulty of the tests. However, in sequence-scanning experiments, not all base pairs within a binding sequence are expected to contribute equally to specific recognition — many base changes are likely to be neutral, thereby lowering the Z-scores observed in this experiment. Nonetheless, if sequence scores are normally distributed, the mean native Z-score observed here (-1.3) represents an elimination of more than 90% of non-cognate DNA sequences, or better than a 10-fold enrichment.

### 3.4.6 In a Test of Binding-Sequence Identification, the All-Atom Score Performs Competitively with a Physical Potential Function

Discrimination of a native structure from a collection of minimized decoys is a more rigorous test than discrimination from structures that have not been minimized. In previous work, Havranek *et al.* produced a minimized sequence-variant decoy set for the EcoRI restriction endonuclease using the Rosetta software package (28). Remarkably, the all-atom method is significantly better at discriminating the cognate recognition sequence from these decoys than is Rosetta's physical potential (mean all-atom Z-score: -2.22, mean Rosetta Z-score: -1.58), although we find that the Rosetta method is better at ranking these structures (mean all-atom rank: 18.3, mean Rosetta rank: 5.5).

This paradoxical result is explained by the fact that the all-atom score produces a distribution of decoy scores with a longer tail than does the Rosetta method (which has a relatively narrow distribution of sequence scores, clustered near the score of the native sequence). In this case, the greater separation of the all-atom native sequence score from the mean of the all-atom score distribution makes up for the slight increase in variance of the all-atom score distribution, and results in a significantly better native Z-score. It is likely that the superior performance of the physical method in ranking structures arises from its very construction: the Rosetta potential was created by choosing weights for the different components of the potential so as to optimize its ability to identify native sequences at protein-DNA interfaces. No equivalent optimization has been made for the all-atom potential.

### 3.4.7 Areas for Improvement

The all-atom potential appears to achieve its greatest performance when discriminating between near-native and incorrect decoy conformations, but has more difficulty when discriminating between structures that are very close to the native conformation. Even in the most successful examples of decoy discrimination, we frequently observe that the very best decoys score better than the native structure, suggesting that the all-atom potential may be missing important details of sequence-

specific interactions. Typically, these high-scoring decoys are well-packed at the molecular interface, with disruptions in some of the contacts that define sequence-specific recognition.

The all-atom potential does not explicitly consider a number of features that may affect sequence-specific recognition (e.g. hydration, electrostatics, molecular flexibility or indirect readout). However, statistical potentials are presumed to implicitly account for some of these features, because the distribution of atomic contacts in the structural database reflects the contribution of these forces in protein-DNA structures. We believe that the greatest weakness of the current all-atom potential is its lack of consideration of local structural detail, such as the directionality of hydrogen bonds (the utility of which was demonstrated for protein-protein (54), and protein-RNA interactions (1)).

### 3.4.8 Conclusion

We have constructed an all-atom statistical potential function for evaluating protein-DNA interactions based on structure, shown that it outperforms previous statistical models, and that it is competitive with a physical potential function developed for protein-DNA interfaces. The performance of the all-atom score in tests of cognate sequence recognition suggests that it may soon have application to problems previously reserved to more complex approaches, such as the large-scale detection of protein binding sequences within a genome, or the computational design of nucleic-acid binding proteins.

A potentially important element missing from this approach is the directional information contained in hydrogen-bonding geometry. We plan to investigate the incorporation of an orientation-dependent description of hydrogen bonding, and we are confident that the ability of the all-atom potential to discriminate near-native structures and identify cognate DNA-binding sequences will be improved once this property is added to the model.

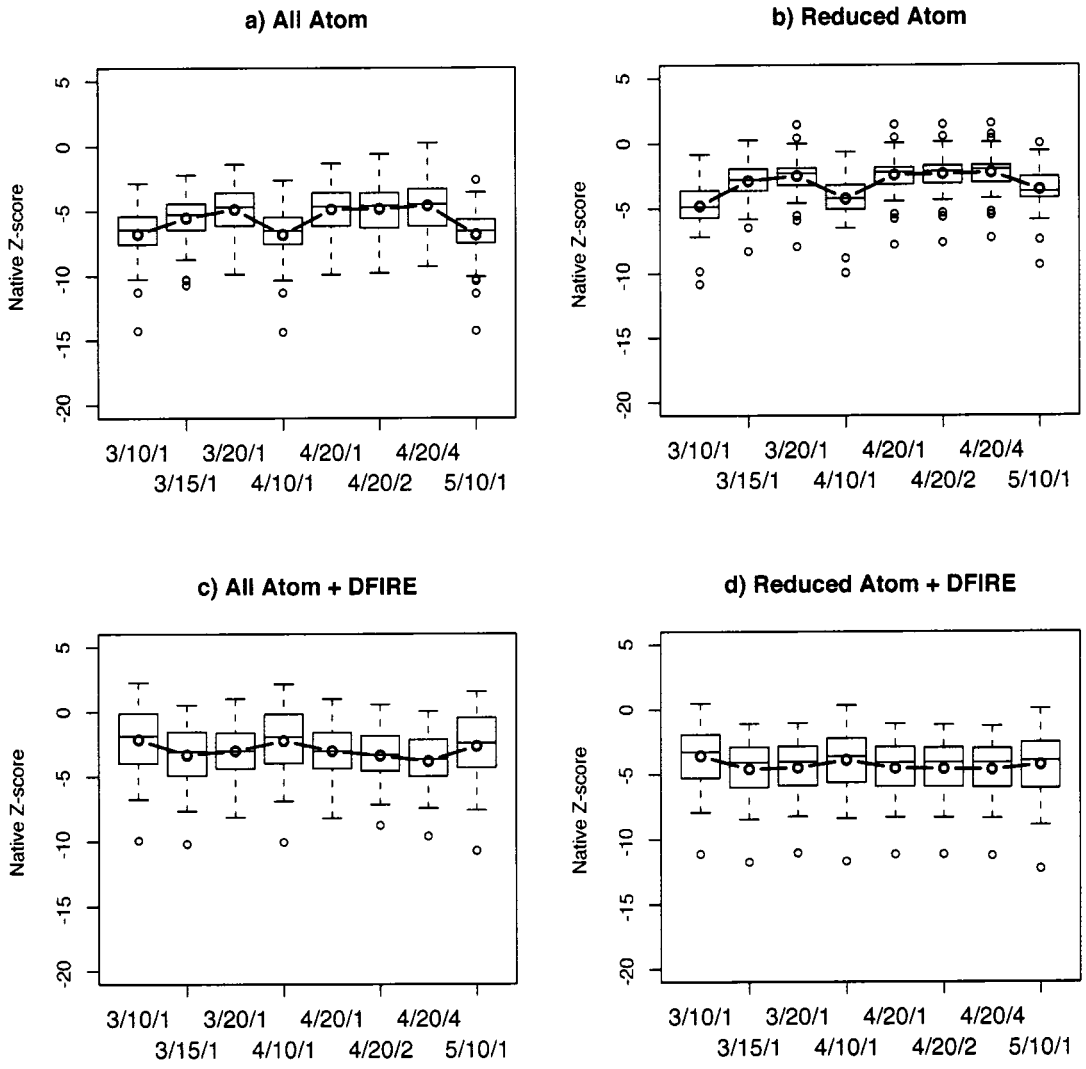



**Figure 3.1: Native structure Z-scores for different score/parameter combinations.** Each boxplot represents the native Z-scores for a set of 45 protein-DNA complexes, calculated relative to 2000 docking decoys per complex. The lower, middle and upper horizontal lines in each boxplot represent the 25th, 50th and 75th percentile native Z-scores for each set, respectively. Whiskers extend to 1.5 times the interquartile range, and outliers are represented as points. The mean Z-scores are connected by lines. One score type is shown per panel, with individual boxes representing a combination of score parameters, where the first parameter is the initial distance bin width, the second is the distance cutoff used to define a contact, and the third is the width of all remaining bins.

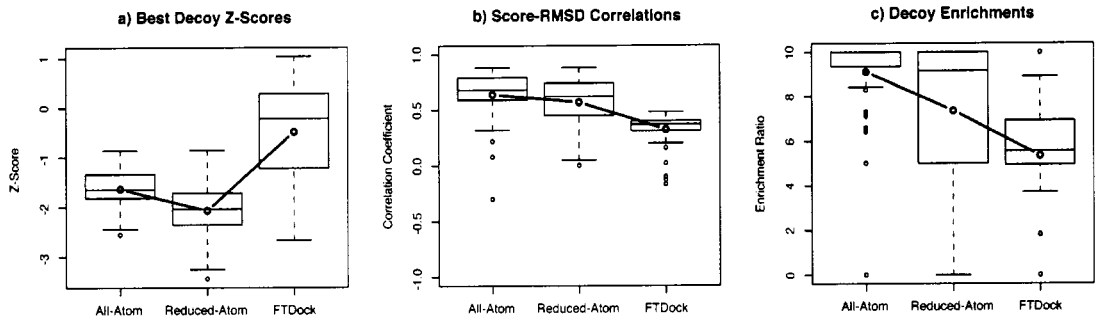



**Figure 3.2: Score discrimination performance for near-native decoys.**
Each boxplot represents the specified metric for a set of 45 protein-DNA complexes, calculated relative to the scores for a set of near-native, "refined" docking decoys created using a variation of the FTDock docking protocol. a) Z-score of the best (nearest-native) decoy structure in each decoy set, relative to all other members of the decoy set. b) Score-RMSD correlation coefficient for all decoys nearer than 5Å RMSD to the native structure. c) Native-like decoy "enrichment," defined as the percentage of decoys with ≥95% correct contacts in the 1% best-scoring decoy structures.

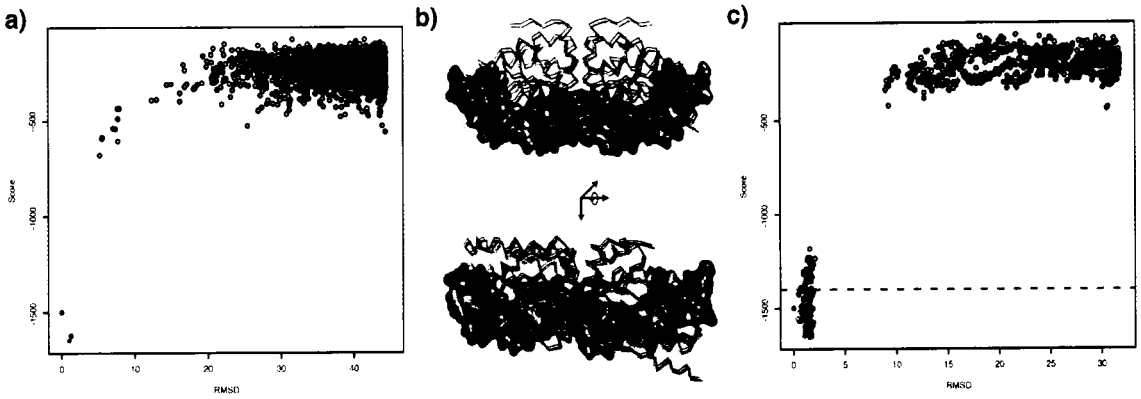



**Figure 3.3: An example of good decoy-discrimination performance.**

a) Scores for the rigid-body docking decoys of homing endonuclease structure (1g9z) are plotted relative to their root-mean-square-deviation (RMSD) from the native complex structure. The native structure is plotted in red, and the remaining colored points represent the three decoys whose structures are shown. b) The structures of the three highlighted decoys, as well as the native backbone conformation, are rendered relative to the native DNA molecule for this complex. c) The near-native docking decoy set for this complex. Green points represent those decoys with ≥ 95% correct contacts, the dashed line represents the tenth percentile score cutoff, and the solid gray line is the line of best fit for those decoys ≤ 5Å RMSD to the native complex structure.

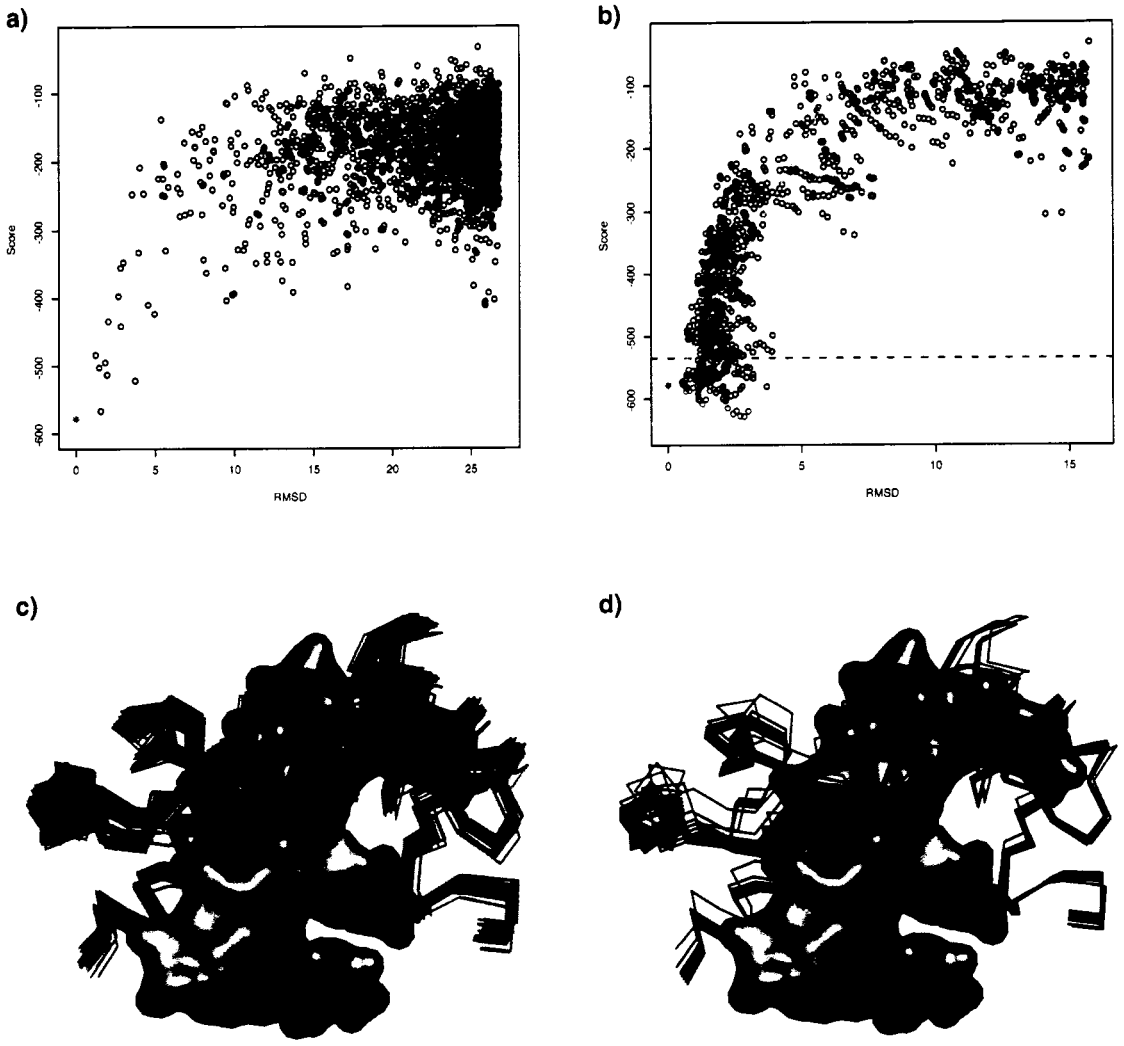



**Figure 3.4: An example of average decoy-discrimination performance.**

a) Scores for the rigid-body docking decoys of the avian myeloblastosis virus v-Myb structure (1h8a) are plotted relative to their native RMSD values, with the native structure score shown in red. b) The near-native decoy set for this complex. Green points represent those decoys with ≥ 95% correct contacts, the dashed line represents the tenth percentile score cutoff, and the blue points represent the six best-scoring decoys for the decoy set. The line of best fit for decoys ≤ 5Å RMSD to the native complex structure is shown in gray. c) The backbone structures of the native structure (red) and the near-native decoys with ≥ 95% correct contacts (green), rendered relative to the native DNA molecule for the complex. d) The backbone traces of the native structure (red) and the six best-scoring decoy structures in the near-native decoy set (blue).

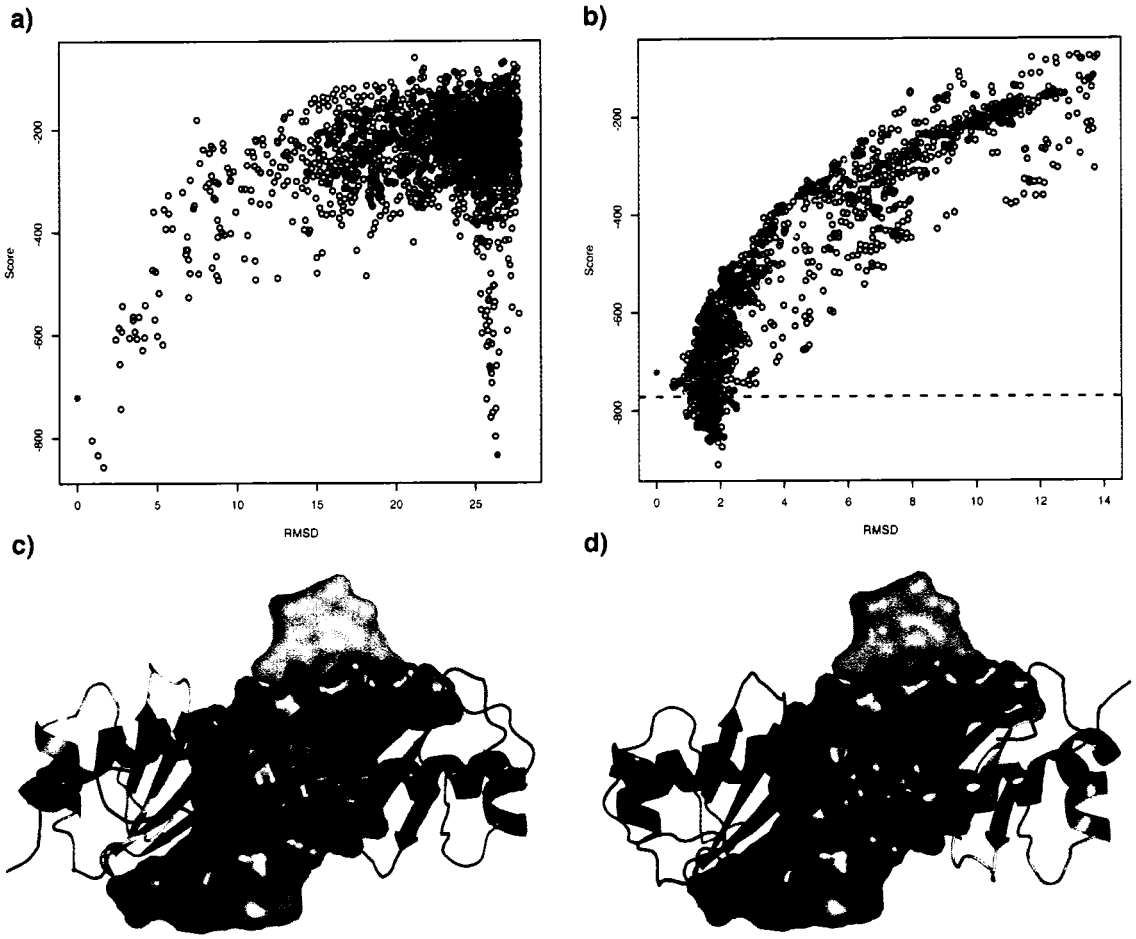



**Figure 3.5: An example of poor decoy-discrimination performance.**

a) Scores for the rigid-body docking decoys of a TATA-binding protein from Arabidopsis (1qna) are plotted relative to their native RMSD values. The red asterisk represents the native complex, and the blue point represents the "flipped" false-positive decoy shown in panel d. b) The near-native docking decoy set for this complex. Green points represent those decoys with ≥ 95% correct contacts, the dashed line represents the tenth percentile score cutoff, and the gray line shows the line of best fit for those decoys ≤ 5Å RMSD to the native complex structure. c) The structure of the native complex. d) The structure of the false-positive decoy plotted in light blue in panel a, as an example of the "flip" pathology.

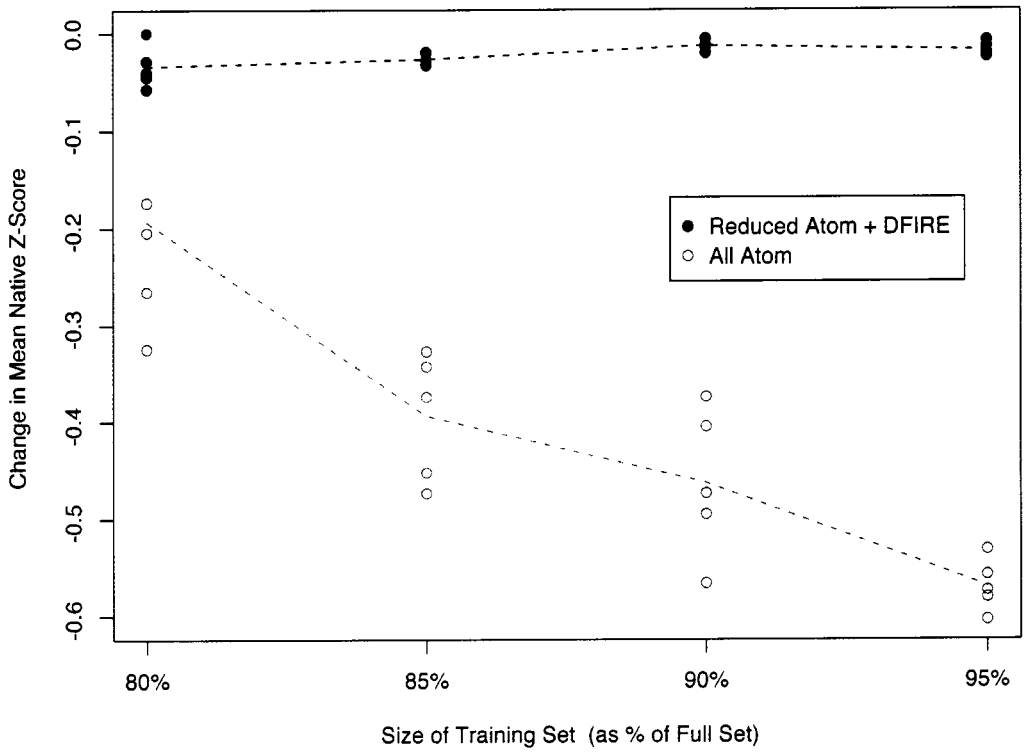



**Figure 3.6: The impact of training set size on native structure discrimination.** Two data sets are plotted on a common axis, one representing the native Z-score data for the reduced-atom DFIRE potential, and another representing the equivalent data for the all-atom statistical potential. Each data point represents the change in mean native Z-score for 45 protein-DNA docking decoy sets, calculated from a potential trained using a randomly-selected training set of the size listed. Five replicate experiments were performed per score for each training set, and the means of these replicate experiments are connected using dotted lines. In order to plot to a common axis, the data sets were independently translated such that the maximum mean native Z-score for each potential was placed at a y-axis value of 0.0.

**Table 3.1: The relative contribution of nucleic-acid base and backbone contacts to the native Z-score.**

All-atom scores were constructed using a subset of training data consisting exclusively of contacts to the nucleic-acid backbone ("BB") or contacts to the nucleic acid bases ("Base"), and these were compared to the full all-atom function ("All") for their ability to calculate native Z-scores and score-RMSD correlations for the near-native docking decoy sets described in . For each structure, the best Z-score and correlation coefficient are shown in bold. Structures are arranged in order of descending mean Γ angle, a measurement of overall DNA deformation (per base-step), where higher values indicate a greater overall deformation (164).

| | | Native Z-Score | | | RMSD Correlation | | |
|---|---|---|---|---|---|---|---|
| PDB ID | Γ (°) | BB | Base | All | BB | Base | All |
| 1qna | 35.70 | **-0.98** | -0.84 | -0.92 | **0.76** | 0.67 | 0.71 |
| 1d02 | 13.62 | -1.56 | -1.69 | **-1.72** | 0.76 | **0.89** | 0.83 |
| 1eon | 13.41 | -1.69 | -1.63 | **-1.71** | **0.66** | 0.25 | 0.62 |
| 1ckq | 12.29 | -1.04 | -1.02 | **-1.09** | **0.72** | 0.65 | 0.70 |
| 1dmu | 9.12 | -1.45 | -1.44 | **-1.47** | **0.40** | 0.32 | 0.38 |
| 1qpz | 8.53 | -2.21 | -2.16 | **-2.26** | **0.86** | 0.80 | 0.83 |
| 1au7 | 8.48 | -1.62 | -1.63 | **-1.67** | **0.49** | 0.42 | 0.46 |
| 1je8 | 8.15 | -1.66 | -1.62 | **-1.73** | **0.81** | 0.79 | **0.81** |
| 2cgp | 7.84 | -1.03 | **-1.31** | -1.27 | 0.35 | **0.70** | 0.60 |
| 1b3t | 7.74 | **-1.28** | -1.17 | **-1.28** | **0.59** | 0.45 | 0.54 |
| 1tc3 | 7.30 | -1.35 | -1.45 | **-1.53** | 0.60 | **0.68** | **0.68** |
| 1g9z | 7.17 | -2.48 | **-2.59** | -2.58 | 0.21 | 0.22 | **0.26** |
| 1zme | 6.84 | -2.09 | -2.02 | **-2.14** | 0.85 | **0.91** | 0.89 |
| 1a73 | 6.56 | -2.07 | -2.05 | **-2.14** | **0.74** | 0.70 | 0.72 |
| 1jko | 6.55 | **-1.84** | -1.44 | -1.69 | 0.84 | 0.83 | **0.86** |
| 1bdt | 6.41 | -1.77 | -1.92 | **-1.93** | 0.81 | **0.85** | 0.84 |
| 2bop | 6.28 | **-1.60** | -1.44 | -1.58 | **0.79** | 0.76 | 0.78 |
| 1a1i | 6.21 | -1.55 | -1.73 | **-1.75** | 0.69 | 0.67 | **0.71** |
| 1bc8 | 6.10 | **-1.53** | -1.22 | -1.41 | 0.77 | 0.77 | **0.78** |
| 1pdn | 6.04 | **-1.48** | -0.90 | -1.28 | **0.69** | 0.63 | 0.67 |
| 1skn | 5.96 | -1.65 | -1.55 | **-1.66** | **0.58** | 0.46 | 0.53 |
| 1mjo | 5.94 | -1.90 | **-2.12** | -2.07 | **0.90** | 0.88 | **0.90** |
| 1bl0 | 5.88 | -0.93 | **-1.01** | **-1.01** | **0.60** | 0.48 | 0.53 |
| 2dgc | 5.75 | -1.88 | -1.78 | **-1.89** | **0.74** | 0.54 | 0.69 |
| 3pvi | 5.71 | -1.67 | -1.62 | **-1.68** | **0.60** | 0.40 | 0.56 |
| 2hdd | 5.61 | -2.66 | -2.56 | **-2.76** | **0.69** | 0.45 | 0.67 |
| 1ign | 5.19 | -1.69 | -1.75 | **-1.79** | **0.70** | 0.65 | 0.67 |
| 1qpi | 5.09 | -2.33 | -2.25 | **-2.40** | **0.86** | 0.81 | 0.84 |
| 1a3q | 5.08 | -1.47 | -1.51 | **-1.56** | **0.21** | 0.12 | 0.16 |
| 1dfm | 5.05 | -1.11 | -1.14 | **-1.15** | -0.18 | -0.11 | -0.22 |
| 1lq1 | 5.04 | -1.97 | -2.16 | **-2.22** | 0.76 | **0.85** | 0.83 |
| 1tro | 5.02 | -1.51 | -1.52 | **-1.60** | 0.73 | 0.73 | **0.75** |
| 1fjl | 4.95 | -1.63 | -1.58 | **-1.72** | **0.82** | 0.74 | 0.79 |
| 1h8a (a) | 4.82 | **-1.97** | -1.81 | **-1.97** | **0.45** | 0.15 | 0.36 |

Table 3.1 continued

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1h8a (b) | 4.82 | **-1.70** | -1.44 | -1.68 | 0.58 | 0.59 | **0.60** |
| 1f4k | 4.80 | -0.98 | -1.05 | **-1.07** | 0.76 | 0.77 | **0.78** |
| 6pax | 4.73 | **-1.24** | -0.91 | -1.13 | **0.64** | 0.53 | 0.60 |
| 1hlv | 4.53 | -1.78 | -1.80 | **-1.86** | **0.82** | 0.78 | 0.81 |
| 1mnn | 4.46 | -1.62 | -1.53 | **-1.67** | **0.71** | 0.62 | 0.67 |
| 1dsz | 4.38 | **-1.22** | -0.81 | -1.09 | **0.71** | 0.67 | **0.71** |
| 1hwt | 4.13 | -1.69 | -1.82 | **-1.87** | 0.72 | **0.89** | 0.85 |
| 1per | 4.09 | **-1.56** | -1.18 | -1.40 | **0.62** | 0.57 | 0.59 |
| 1l3l | 4.02 | -1.80 | **-2.16** | -2.11 | 0.85 | **0.87** | 0.86 |
| 3hts | 3.87 | **-0.80** | -0.57 | -0.72 | **0.81** | 0.78 | 0.80 |
| 3bam | 3.77 | -1.63 | -1.63 | **-1.69** | 0.36 | 0.35 | **0.41** |
| | mean | -1.61 | -1.57 | -1.66 | 0.65 | 0.61 | 0.65 |
| | σ | 0.40 | 0.46 | 0.44 | 0.21 | 0.24 | 0.22 |

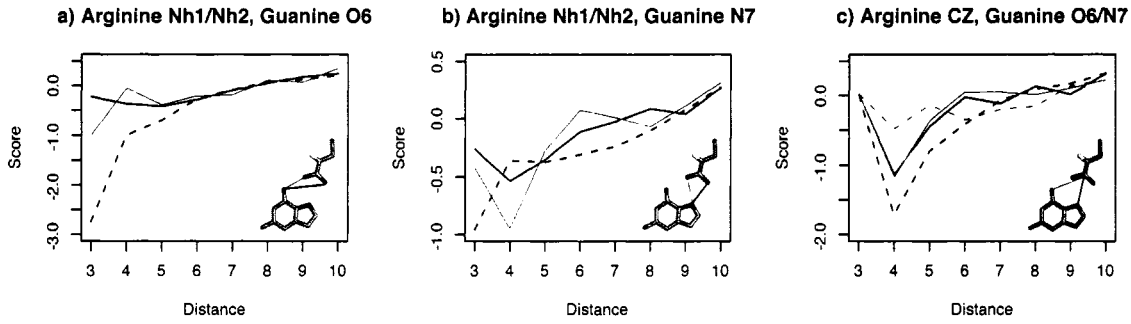



**Figure 3.7: All-atom scores for the arginine-guanine interaction.**

In each panel, scores for the interaction of specific arginine and guanine atoms are plotted relative to interatomic distance, with the equivalent reduced-atom scores shown as dashed lines. a) The interaction of arginine NH1 (black) and NH2 (red) with guanine O6. b) The interaction of arginine NH1 (black) and NH2 (red) with guanine N7. c) The interaction of arginine $C_\zeta$ with guanine O6 (red) and N7 (black). Here, the reduced-atom equivalent to the arginine $C_\zeta$-guanine O6 interaction is shown as a red dashed line, with the reduced-atom equivalent to arginine $C_\zeta$-guanine N7 shown as a black dashed line.

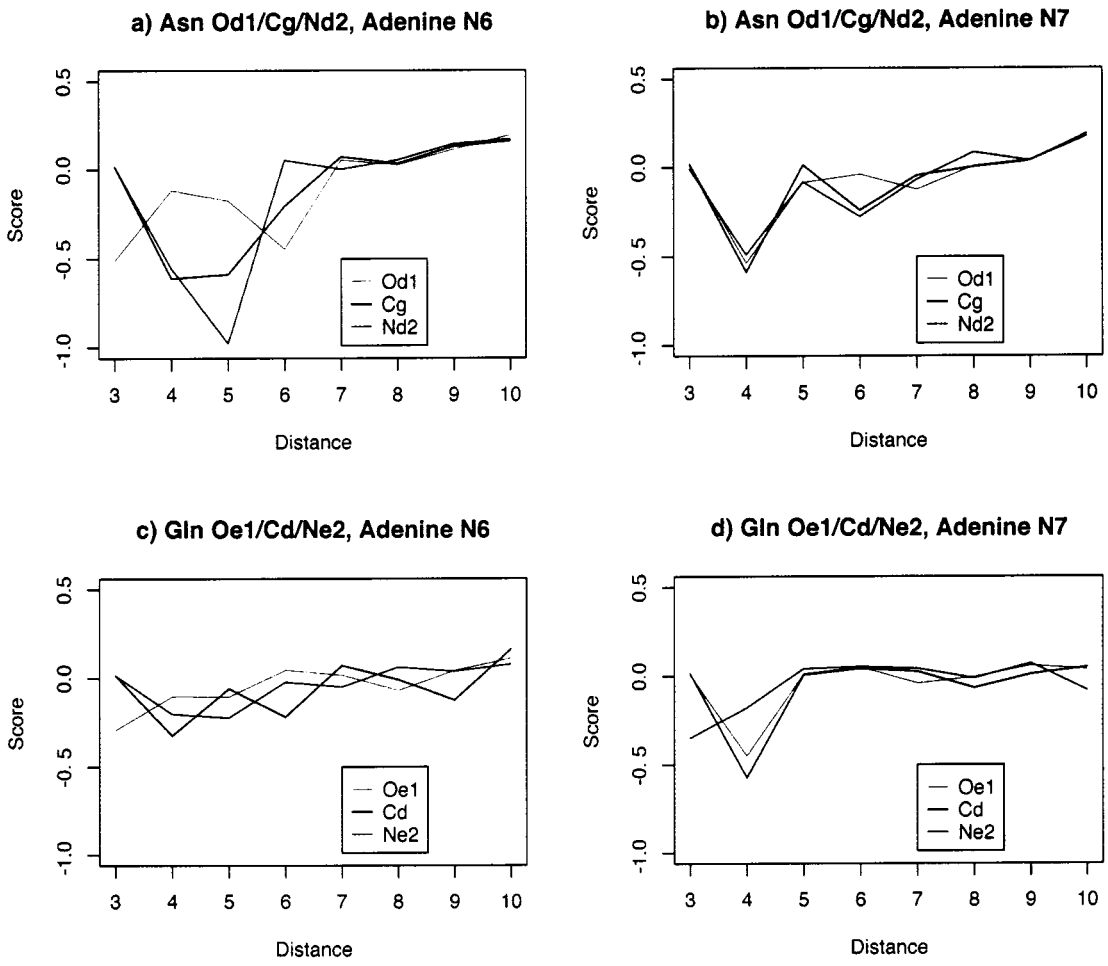



**Figure 3.8: All-atom scores for the interaction of asparagine and glutamine with adenine.**

Scores are plotted relative to interatomic distance. a) Asparagine $O_{\delta 1}$, $C_{\gamma}$, and $N_{\delta 2}$ interactions with adenine N6. b) Same asparagine atoms, interactions with adenine N7. c) Glutamine $O_{\varepsilon 1}$, $C_{\delta}$ and $N_{\varepsilon 2}$ interactions with adenine N6. d) Same glutamine atoms, interactions with adenine N7.

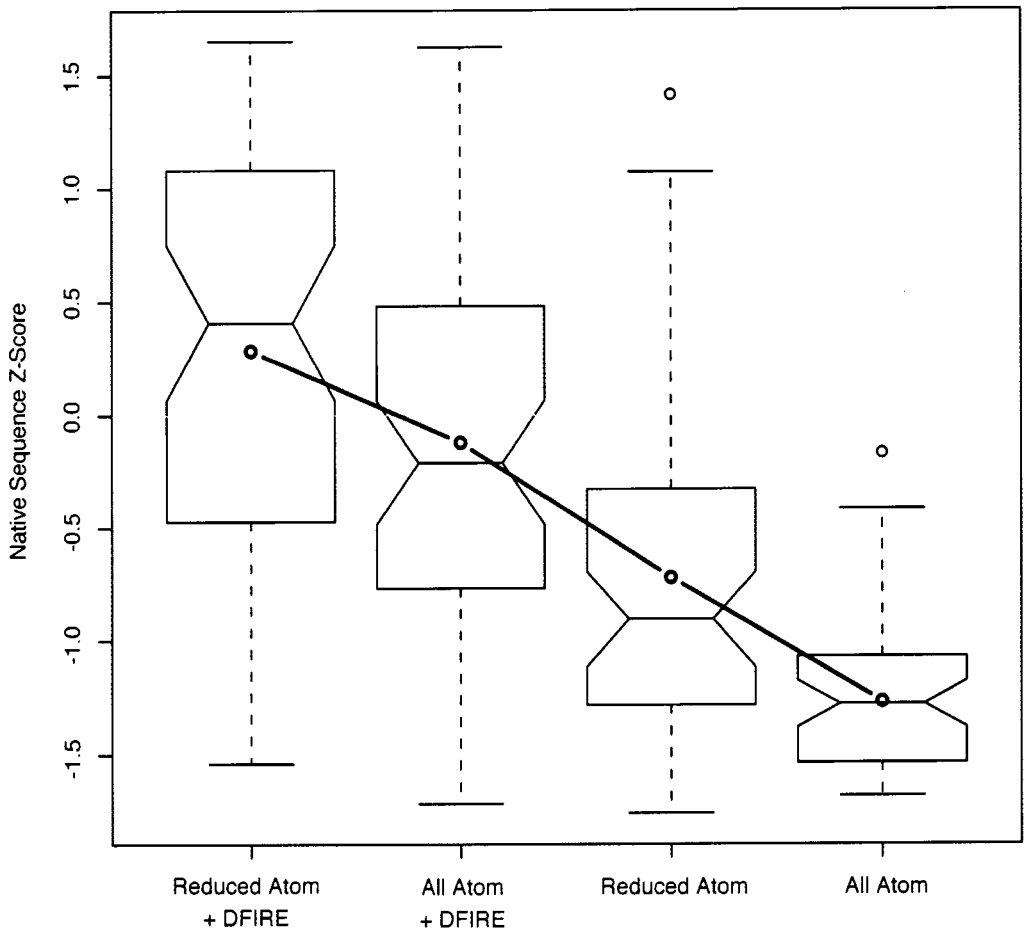



**Figure 3.9: DNA sequence discrimination test.**

Each box plot represents the distribution of Z-scores for the cognate DNA-binding sequences of 52 protein-DNA complexes, calculated relative to 10,000 random DNA sequences per complex. Box plots are interpreted as in Figure 3.1, with the addition of notches to denote significant differences in median Z-score (non-overlapping notches are considered strong evidence of significance). Mean Z-scores for each score type are shown in red, and connected by lines.

**Table 3.2: A comparison of the all-atom method with the Rosetta physical potential function.**

The all-atom Z-score and ranking of ten minimized native EcoRI complexes are shown relative to 620 minimized DNA sequence mutant decoys generated by Havranek *et al.* (28) using the Rosetta package. The all-atom Z-score, the Rosetta Z-score and their ranks are shown for each minimized native structure, with the mean and standard deviation for all structures shown in the final row. Rank is determined by ordering the minimized structures according to score, and counting the best-scoring structure as rank 1, the second-best as rank 2, etc.

| Repacked Structure | All-Atom | | Rosetta | |
|---|---|---|---|---|
| | Z-Score | Rank | Z-Score | Rank |
| 1 | -2.21 | 18 | -1.60 | 2 |
| 2 | -2.27 | 14 | -1.58 | 7 |
| 3 | -2.21 | 19 | -1.60 | 3 |
| 4 | -2.27 | 12 | -1.53 | 10 |
| 5 | -2.27 | 13 | -1.58 | 6 |
| 6 | -2.05 | 40 | -1.61 | 1 |
| 7 | -2.27 | 15 | -1.56 | 8 |
| 8 | -2.28 | 11 | -1.54 | 9 |
| 9 | -2.21 | 20 | -1.60 | 4 |
| 10 | -2.21 | 21 | -1.60 | 5 |
| Mean ($\sigma$) | -2.22 (0.07) | 18.3 (8.38) | -1.58 (0.03) | 5.5 (3.02) |

# 4 A Knowledge-Based Potential Function Predicts the Specificity and Binding Energy of RNA-Binding Proteins

## 4.1 Introduction

The sequence-specific recognition of RNA by proteins plays a fundamental role in gene expression by directing different cellular RNAs to specific processing pathways or sub-cellular locations. Experimental studies have explored the molecular basis for the sequence dependence of protein-RNA recognition (169-174), and a number of studies have explored this problem from a structural perspective (1, 8, 106-108, 153, 175-179). However, these previous efforts have mostly emphasized qualitative descriptions of the recognition process, and relatively few attempts have been made to quantify the characteristics of protein-RNA interactions using computational approaches (1). In order to better understand the biological function of RNA-binding proteins, a quantitative description of protein-RNA interactions is needed.

Among the simplest, most successful approaches to the quantitative analysis of biological structures are knowledge-based potential functions. These have been employed in protein structure prediction (13, 14, 16-19, 55, 168, 180), as well as in the prediction of protein-protein (19, 22, 24, 25) and protein-ligand interactions (20-23), yet despite their simplicity, they are useful predictors of structure and molecular interactions. A few works have explored the use of knowledge-based methods for the prediction of protein-DNA interactions from structure (22, 26, 59), and more recently, our group (2) and others (30) have independently demonstrated that simple, knowledge-based potentials can provide quantitative descriptions of protein-DNA interfaces that are comparable to much more complex approaches, such as force fields used for molecular mechanics (30).

The relative scarcity of high-resolution structures of protein-RNA complexes has represented an understandable barrier to the application of knowledge-based computational approaches to the problem of protein-RNA recognition. However, we have

previously demonstrated that a statistical hydrogen bonding (HB) potential can discriminate native structures of protein-RNA complexes from docking decoy sets (1). Nonetheless, hydrogen bonds represent only about 25% of contacts between protein and RNA (106), and there is reason to believe that a more comprehensive, knowledge-based approach could describe these interactions more effectively.

In this work, we report the application of an all-atom, distance-dependent statistical potential to the prediction of sequence-specific recognition between proteins and RNA. We demonstrate that this simple approach can discriminate native complex structures from near-native decoys, can recapitulate experimentally determined relative binding energies (ΔΔG's) for several protein-RNA complexes, and can be used to predict the RNA sequences recognized by a number of different RNA recognition motif (RRM) and K homology (KH) domains. These results represent the first large-scale, structure-based, quantitative analysis of protein-RNA recognition processes, and demonstrate that simple, statistical models can be as powerful as physical potential functions when applied to problems requiring the high-resolution modeling of protein-RNA interactions.

## 4.2 Methods

### 4.2.1 All-Atom Distance Potential

The potential function used here is identical to the method described by Robertson and Varani (2), with the exception of a modified low-count correction. In this work, the correction described by Sippl (13) is replaced with a weighted pseudocount method, where a constant number of pseudocounts ($P$) are added to the observed counts for each atom pair. These pseudocounts are allocated over distance bins in proportion to the background frequency ($f(d_{ij})$) values, leading to an updated expression for $f(d_{ij}, t_i, t_j)$:

$$f(d_{ij},t_i,t_j)_{adj} = \frac{N_{obs}(d_{ij},t_i,t_j) + P \cdot f(d_{ij})}{\sum_{d_{ij}} N_{obs}(d_{ij},t_i,t_j) + P} \qquad (4.1)$$

Where $d_{ij}$ is the cartesian distance observed between two atoms $(i,j)$, of types $t_i$ and $t_j$, and $N_{obs}(d_{ij},t_i,t_j)$ represents the number of atoms of types $t_i$ and $t_j$ observed in the structure training set, separated by a distance of at least $d_{ij}$.

In the present work, $P$ is set to 75 (a value found empirically to maximize the discrimination power of the score in tests using protein-DNA complexes); no attempt was made to optimize the value of this constant for protein-RNA interactions. As a control, we also tested a simple, contact-counting method, wherein every contact between protein and RNA (within a given distance cutoff) was assigned a score of -1.

### 4.2.2 Atom Type Selection

Atom score types were assigned using the method of Robertson and Varani (2). Briefly, the all-atom potential treats every atom, in every residue, as a unique type (*e.g.* Alanine Cβ and Arginine Cβ are considered unique atom types under this scheme), resulting in a total of 158 protein, and 81 RNA atom types.

### 4.2.3 Training Set Selection

The training set contains crystal structures of protein-RNA complexes downloaded from Protein Data Bank (PDB) (142)with resolution better than 2.5Å. Structures with more than 20% sequence identity were identified using the *ExPASy* sequence-redundancy tool (181); the higher-resolution structure of every homologous pair was retained. After filtering, the training set contained 72 protein-RNA complexes (the 50S ribosome structure comprises 28 individual peptide chains in complex with RNA, plus 44 independent protein-RNA crystal structures). Because of the limited number of protein-RNA structures, it was necessary to use a combined training/test set. Thus, in order to assess the performance of the potential without biasing the result, the native structure of scored complexes were excluded from the training set for their scores (*i.e.* leave-one-out cross-validation).

### 4.2.4 Construction of Test Sets

Five rigid-body docking decoy sets (originally generated by Chen *et al.* (1)) were used for initial decoy-discrimination tests in this work. These were constructed using the docking module of ROSETTA, which incorporates energy minimization through the use of a protein side-chain repacking algorithm (70, 182). Each of these decoy sets contains 2000 structures with deviations of between 1-35Å from the native complex structure.

Additionally, near-native decoys were generated for 21 different protein-RNA complexes by extracting time-step structures from molecular dynamics (MD) trajectories, created using AMBER 8 in a deformation-like process with *ff99* force field (183). The initial structure of each complex was first minimized in 500 steps (250 steps of steepest-descent and 250 steps of conjugate gradient minimization), then heated from 0K to 400 K (a temperature range empirically chosen to produce mean decoy set deviations in the range 1-5Å RMSD) in 20ps using a Langevin dynamics algorithm (184, 185). Snapshots were taken every 0.05ps, and a total of 400 structures were extracted from each MD simulation. The binding free energy was calculated using the MM_GBSA module of AMBER 8 as:

$$\Delta G_{bind} = G_{complex} - (G_{protein} + G_{RNA}) \qquad (4.2)$$

where $G_{complex}$, $G_{protein}$ and $G_{RNA}$ represent the MM_GBSA-calculated free energies of the protein-RNA complex, the free protein and the free RNA, respectively.

### 4.2.5 Prediction of Sequence Specificity for RNA-Binding Domains

Because many RNA-binding domains of the RRM superfamily interact in a conserved fashion with four nucleotides across the surface of the β-sheet of the structure (175, 186, 187), and recognition by KH domains appears to be conserved between different domains as well (175, 188-191), we adopted a four-nucleotide model for our sequence-specificity tests. Starting with each complex in our training set (containing one or more RRM or KH domains), we extracted the protein coordinates and the four nucleotides bound at the center of the domain (for structures containing more than one RNA-binding domain, the structure was divided into two independent domains). This

approach was chosen to allow for an unbiased evaluation of the potential function, despite the knowledge that this overly simple model of RRM recognition would fail for certain protein domains that bind anomalously (e.g. U1A protein), or in situations where two domains cooperatively define sequence specificity.

For the selected domains, every nucleotide was replaced by A, U, C and G, systematically, in all possible combinations, using Insight II 2000 (192). Thus, 256 different structures were generated for each binding domain, and minimized using AMBER 8 (183) in 20 steps to regularize the local structure. Some RRM and KH domains in complex with single strand DNA (PDB code: 2UP1, 1WTB, 1X0F, 1ZZI and 1ZZJ) were also included in the test set because recognition of single stranded RNA and DNA are mechanistically similar. Protein mutations were modeled using MOE (193), followed by energy minimization with AMBER; the conformation of the mutated residue with side chain conformation most similar to the native residue was retained.

## 4.3 Results

### 4.3.1 Docking Decoy Discrimination

An important property of any potential function used to evaluate the quality of protein-RNA complexes is its ability to discriminate cognate (native, crystallographically determined structures) from non-cognate (decoy) structures. Thus, as a preliminary test of our method, and a direct comparison to previous work, we used our distance-dependent potential to evaluate the docking decoys generated by Chen *et al.* (1) for their application of the ROSETTA physical potential function to protein-RNA interactions. These decoys were created using a combination of rigid-body docking and protein side-chain repacking, and range in RMSD (to the native complex structure) from less than 1Å, to over 20Å. Thus, although these decoys are not especially near-native structures, they represent a solid basis for comparison to a much more complicated scoring method (the multi-term, hybrid physical/statistical potential function used by ROSETTA).

When scored with the distance-dependent potential, the native complex can always be identified as the best-scoring structure in each of the five decoy sets (Figure 4.1

through Figure 4.5); the native structure Z-scores for these decoy sets are listed in Table 4.1. These values indicate a strong discriminatory ability, comparable to that reported by Chen *et al.* (1) using their significantly more complicated method. Specifically, the distance-dependent potential outperforms the Chen *et al.* method for the 1CVJ decoy set (Z-score: -7.02 vs. -5.11), is approximately equivalent to their results for the 1EC6 and 1FXL decoy sets (Z-score: -6.46 vs. -6.53 and -2.66 vs. -2.70), and is worse for the 1JID and 1URN decoys (Z-score: -6.29 vs. -9.12 and -4.80 vs. -8.39). Overall, the distance potential (using a 6Å cutoff) results in a mean native Z-score of -5.45, versus the value of -6.37 obtained by Chen *et al.* (Table 4.1); this a statistically insignificant difference ($p=0.53$, Welch's two-sided t-test). As a control, a simple, contact-counting "potential" was used to evaluate the decoys; this approach was relatively unsuccessful, obtaining a mean Z-score of -2.64, less than half of the mean native Z-score found using the distance potential.

This result is remarkable, because the potential used by Chen *et al.* contained multiple terms corresponding to packing and solvation, as well as others designed to reproduce the statistical frequency of protein side chains in protein-nucleic acid interfaces. The relative weight of these different terms was found by an optimization procedure that aimed to reproduce the average composition of the interfaces observed in nature. In contrast, the current statistical potential was generated 'as-is' from the observed frequency of intermolecular contacts in the database of protein-RNA structures.

Interestingly, we find that the magnitude of the observed Z-scores declines significantly as the contact cutoff is increased from 6Å to 12Å (data not shown), suggesting that short-range contacts provide the bulk of the discriminatory power in these tests. Together, these results suggest that protein-RNA recognition specificity is primarily determined by short-range intermolecular contacts. Long-range phenomena (such as non-local electrostatics) appear to play a more limited role, at least with regard to specificity.

### 4.3.2 Near-Native Decoy Discrimination

In order to better understand the capabilities of the distance-dependent potential function, we compared its ability to discriminate near-native protein-RNA structures, with that of force field implemented in the AMBER 8 molecular simulation package. We generated near-native protein-RNA decoys for 21 protein-RNA complexes by using AMBER 8 to conduct molecular dynamics simulations of the native protein-RNA complexes, and selecting time-steps from the resulting MD trajectories for each structure. We then scored these decoy structures using the distance-dependent potential function, and examined the correlations between distance scores and AMBER energies for each decoy set; the results of this test are summarized in Table 4.2. This is a difficult test of score performance, and neither the distance potential nor the Amber potential appears to reliably discriminate native structures from these very near-native, MD-generated decoys (average Z-score of -0.69 vs. -0.59, Table 4.2). The observed correlation of the scores with RMSD is also poor (*ibid.*, average $R^2$ of 0.01 and 0.05 respectively). Despite this, the distance-dependent statistical potential shows moderate to strong correlations (average $R^2$ of 0.41) with the energy values predicted by the AMBER force field. This is a remarkable result, in that the distance-dependent potential is able to achieve discrimination performance on par with the (significantly more complicated) method used to generate the decoys in question.

### 4.3.3 Identifying RNA-Binding Sequences from Structure

Having established the performance of the statistical potential function in tests of decoy discrimination, we investigated the ability of the potential to predict the cognate recognition sequences of RNA-binding proteins. This is a particularly important problem, because sequence specificity is known for only a fraction of all RNA-binding proteins. The ability to predict (or at least, narrow down) the cognate sequence for 'orphan' RNA-binding proteins would greatly facilitate the design of biological experiments. In order to study whether the distance potential can be used to identify the

nucleotide sequences recognized by RNA-binding proteins, the potential was used to score sequence-variant decoys for a number of RRM and KH domain complexes.

This test relied on a specific structural model of the RNA recognition process of RRM and KH domains, wherein four RNA nucleotides are recognized specifically by each domain – a mechanism strongly supported by previous research on the mechanism of protein-RNA recognition for RRM proteins (175, 186, 187) and by the structure of existing KH domains bound to RNA (194). Complexes containing two RNA-binding domains were divided into independent structures (e.g. 1CVJ_1 and 1CVJ_2 represent the first and second Poly-A binding protein domain of structure 1CVJ, respectively), and the two domains were considered as structurally and thermodynamically unrelated. Since the model assumes that each RRM and KH domain binds to 4 nucleotides, we generated a set of $4^4$ (256) different structures for each protein-RNA complex by computationally "threading" all possible four-nucleotide combinations onto the RNA bases nearest the center of the β-sheet structure of the RRM. We then scored these sequence-variant structures with the distance-dependent potential function.

Figure 4.6a shows the results of this analysis. If the potential and model of recognition were perfect – and if each structure was sequence-specific and corresponded to the most favorable sequence recognized by a given domain – the cognate sequences of the tested structures would be expected to rank as number 1. In reality, it is unlikely that the cognate recognition sequences for all domains will be consistently assigned the best score, and therefore, it is more reasonable to express sequence-discrimination performance in terms of percentiles (where perfect discrimination of the cognate recognition sequence would result in a percentile score of 100). Remarkably, we find that 18 of the 29 tested RRM and KH domain complexes had their cognate recognition sequence ranked above the 90th percentile (*i.e.* had better than 10-fold enrichment for the correct sequence). Furthermore, the distance-dependent potential ranks the cognate recognition sequences of the protein-RNA complexes in our test set above the 90th percentile, on average, compared to the 41st percentile for the control score (Figure 4.6b).

Among successful examples of cognate binding-sequence discrimination, the native sequences of the RRM1 of Sex-lethal protein (1B7F_1) and KH1 domain of Poly-C-binding protein-2 were both ranked first of 256 sequences, while KH domain 3 of hnRNP K (1ZZI), RRM of U2B" protein (1A9N) and RRM 4 of Polypyrimidine Tract Binding protein (2ADC_1) each had their cognate recognition sequences ranked in the top 3. However, prediction was less successful for other classical RRM domains. Again, we observe poor performance in tests with the U1A complex, (the cognate recognition sequence of U1A protein was ranked at 30), but perhaps this is not surprising, due to the non-canonical, 7-nucleotide recognition sequence (AUUGCAC) of the U1A protein (which is obviously not in accordance with the four-nucleotide model of RRM sequence recognition used here) (195). Similar difficulties were likely responsible for the poor results obtained for the Poly-A binding potein (1CVJ_1, rank 19), and RRM1 of the HuD protein (1FXL_1, rank 32). Both Pab and HuD utilize two domains to achieve sequence-specific recognition in a cooperative manner, which, again, is not supported by our simplified model of RRM sequence recognition (196). Notably, however, the non-sequence-specific RNA helicase protein (PDB code: 2DB3, included as a negative control) had an expectedly poor cognate sequence rank of 226/256.

### 4.3.4 Estimating Experimentally Determined RNA-Binding Affinities

An important characteristic of any potential function for protein-nucleic-acid interactions is the ability to recapitulate the sequence dependence of experimental binding energies; this is a pre-requisite if the potential is to be applied to problems of protein-RNA interface prediction or design. Fortunately, among the limited number of known protein-RNA complex structures, a few also have a relatively dense set of experimentally determined binding constants for interface mutations. We have used these experimentally characterized mutants to create a set of computationally "mutated" structures of the complexes (shown in Table 4.3), and have scored these structures using the distance-dependent statistical potential.

A first instructive example is provided by mutants of bacteriophage MS2 coat protein bound to a small RNA hairpin (197, 198). Starting with the crystal structure of the complex between MS2 coat protein and the cognate RNA hairpin (PDB code: 1ZDI), a series of structures were generated, representing the RNA and protein for which binding constants are mutations reported in the literature. Then the distance-dependent potential scores for these structures were compared to the known binding constants for each mutation. When all of the MS2 mutations were considered together, we observed a poor correlation between distance score and experimental binding affinities (data not shown). However, we obtained excellent correlations between these values when we divided the binding-affinity data into two subsets (Table 4.3, Figure 4.7, Figure 4.8). One set of protein mutations correspond to complexes where the bound RNA hairpin contained an adenine, guanine or uracil base at position -5, and a second set containing five protein mutants where the bound RNA contained a cytosine at this position. For both sets of mutants, the correlation between distance score and experimental binding affinity is strong ($R^2 = 0.65$, Figure 4.7; $R^2 = 0.97$, Figure 4.8), and statistically significant at the 95% confidence level. Figure 4.9 shows a likely explanation for this result – the characteristic intramolecular hydrogen bond formed by the cytosine at position -5 (197). When this nucleotide is mutated to any other base, the intramolecular hydrogen bond is lost, leading to a reorganization of the RNA structure. At present, the statistical potential does not consider intramolecular contacts; therefore, contributions to binding energy due to changes in RNA structure are not captured by our current approach.

A second example is the human U1A protein (PDB code: 1URN), which is an extensively studied model of protein-RNA recognition, for which considerable binding studies have been conducted. Interestingly, we observe only poor correlations between the distance-dependent score and the experimentally determined dissociation constants ($K_d$) (149) for this protein complex when we conducted a test using a training set of strictly non-homologous protein-RNA structures. However, when the U1A complex itself was included in the training set, we obtained moderate to strong correlations ($R^2$ values between 0.27 and 0.65, depending on the choice of distance cutoff). This suggests

an intriguing hypothesis – U1A may bind to RNA in a manner that is structurally unusual, relative to other known RNA-binding proteins. This hypothesis is further supported by our observation that the inclusion of a U1A homolog (the U2B" complex) in the training set can improve the results of this experiment ($R^2$ from 0.04 to 0.39, Table 4.2). Thus, it appears that the structure of the U1A or homology complex adds a set of protein-RNA atomic contacts (*i.e.* interatomic distances) to the training set that are substantially different than those observed in the 71 other protein-RNA complexes in our set. The U1A complex has long been considered a model for the RRM superfamily, largely because of the availability of NMR and crystallographic structures (199, 200), but it appears that this protein binds to RNA in a manner that is highly unusual. This may reflect the longer RNA sequence recognized by the U1A protein (discussed below), or possibly, the very large conformational changes which have been observed in the RNA upon protein binding (201).

A third example is provided by the Fox-1 protein, which regulates alternative splicing of tissue-specific exons by binding to the GCAUG sequence. The structure of the complex (PDB code: 2ERR) and the experimental binding constants for two sets of related mutations have been recently reported (202): one set for mutations on the Fox-1 protein, and a second set for mutations to its target RNA molecule. A moderately strong correlation was observed between the distance score and the protein mutation data ($R^2$ = 0.46, Figure 4.10), but an anti-correlation was observed for the set of RNA mutations ($R^2$ = -0.57, Table 4.3). The poor performance in this second test likely reflects the failure of the current form of the statistical potential to capture the energetic contribution associated with the disruption of RNA intramolecular interactions that are a characteristic of this complex (202).

Figure 4.11 shows a final example, a universally conserved component of the core of the signal recognition particle (SRP), a ribonucleoprotein that plays a role in targeting newly translated proteins to the endoplasmic reticulum. The structure of the complex (PDB code: 1HQ1) and the binding affinity of a series of RNA mutants have been determined (203). The distance potential results in scores that correlate significantly ($R^2$

= 0.52, p <= 0.05) with experimental binding affinities for mutations involving substitutions of deoxynucleotides for their corresponding ribonucleotides. However, as observed for Fox1, no significant correlation was found for mutations of nucleotides that disrupt critical RNA intramolecular interactions – in this case, formation of base pairs at or near the binding interface.

## 4.4 Discussion

The central role of protein-RNA interactions in the regulation of gene expression has led to considerable interest in the biochemical processes underlying these interactions (204-206). However, much of this research has been devoted to the study of the structure/function relationship for individual protein-RNA complexes, and little effort has been made to develop quantitative models that might describe these interactions. Thus, our understanding of the mechanisms driving protein-RNA recognition is currently largely descriptive (177). Recent work on protein-DNA interactions has shown that quantitative models of protein-nucleic-acid recognition can provide insight into the mechanisms of gene regulation (2, 26, 28-30, 82), and promise to allow the rational design of DNA-binding proteins with altered specificity (64, 207). The development of computational tools capable of predicting the specificity of RNA-binding proteins across entire families (such as the RRM superfamily), or of redesigning the specificity of these proteins, would be of equal importance in dissecting post-transcriptional regulatory mechanisms, and in providing new tools to interrogate gene expression pathways.

In previous work, our group demonstrated that a simple, statistical potential function could be surprisingly sensitive and accurate when used to predict protein-DNA interactions from structure (2); this result was corroborated by a similar study, published concurrently by another group (30). Given these results, we hypothesized that the same approach would be equally successful with protein-RNA interfaces. And while various statistical techniques have been used by a number of groups for the prediction of protein structures, protein-DNA and protein-ligand interactions (13, 14, 16, 18, 19, 21-26, 55, 59, 208), such an approach has never been applied to protein-RNA interactions.

In this work, we describe the successful application of the distance-dependent, all-atom statistical potential function developed by our group (for protein-DNA systems), to the prediction of the energetics and recognition specificity of protein-RNA interactions. We demonstrate that scores created by this simple, statistical method correlate well to the energies predicted by a complex physical potential used for molecular dynamics simulations (the AMBER force field), and that the statistical potential can recapitulate experimentally determined binding constants for a number of protein-RNA complexes (with the caveat that it cannot yet capture the effect of mutations on RNA-RNA interactions at or near an interface). Most significantly, however, we demonstrate that this simple technique is remarkably successful at predicting the cognate recognition sequences of a wide variety of sequence-specific, RNA-binding proteins – a first demonstration of this capability by any method. Together, these results suggest that this relatively simple approach can capture the fine structural details of protein-RNA interactions, and that its discriminatory power is on par with far more complicated methods – such as physics-based potentials developed and optimized for molecular dynamics and protein design.

### 4.4.1 The Impact of Contact Distance Cutoff on Discriminatory Power

The contact distance cutoffs used in this work were varied in order to find the value that maximizes decoy discrimination performance for protein-RNA complexes. Previously, Robertson *et al.* showed that shorter contact cutoffs result in optimal discrimination ability in protein-DNA complexes (2), whereas Samudrala *et al.* found that a longer cutoff (>10Å) was better able to discriminate correct structures during protein structure prediction experiments (14). However, Lu *et al.* has demonstrated that the first coordination shell (*i.e.* a cutoff between 3.5Å and 6.5Å) achieves the greatest selectivity for protein decoys created using gapless threading procedures (16), so some question remains as to the best choice of contact cutoff for any particular problem. In order to evaluate the influence of different cutoff values in our tests, replicate experiments were conducted using 6Å. 10Å and 12Å distance cutoffs. In nearly all of

our tests, the use of a shorter contact cutoff (6Å) results in greater selectivity for structural details of the interface (Table 4.1). For the prediction of mutation energies, however, a longer cutoff appears to outperform shorter cutoff values for some sets of mutation data (Table 4.3). Some of these mutations are not near the protein-RNA interface (*e.g.* in the U1A mutations, where the closest mutation to the protein-RNA interface is 8.89Å from the RNA molecule), and only the use of a longer cutoff value can capture these effects. In light of the differing conclusions of previous research into this question, these results imply that a one-size-fits-all approach to energy function design may be limiting. In other words, it may be possible to significantly improve potential functions by customizing their parameterization to particular problem domains.

### 4.4.2 Prediction of RNA Recognition Sequences from Protein-RNA Structures

An obvious application of any potential function for protein-RNA interactions is the prediction of the cognate binding sequences for RNA-binding proteins. In this work, we have presented the first large-scale, quantitative analysis of the sequence-specificities for K-Homology (KH) and RNA-recognition motif (RRM) domains. In a test of sequence recognition for 29 unique KH and RRM domains, we find that the potential is able to identify (within the 10th percentile) the cognate RNA recognition motifs of these domains approximately 70% of the time. Since not all RRM/KH domains (for example, U1A) obey the simple 4-nucleotide recognition model that we have used here (where each nucleotide makes independent interaction with the protein) (175), this is a remarkably strong result. Despite the simple form of the statistical potential, and the obvious over-simplifications of the 4-nucleotide recognition model, this method is surprisingly robust over the diverse set of RNA-binding domains that we have considered.

### 4.4.3 Prediction of Protein-RNA Binding Energetics

When we evaluated the relative free energy of a set of mutations for several protein-RNA complexes of known structure, the distance-dependent potential was successful within defined structural classes. Indeed, we observe strong, significant ($p <= 0.05$)

score-energy correlations for several sets of mutations that we tested; however, in order to achieve these results, it was necessary to subdivide several of the mutation data sets. For example, for the MS2 complex, the mutation data had to be divided into two classes, based on the presence or absence of a cytosine at position -5 in the RNA. A likely explanation for the importance of the -5 cytosine mutation is offered by the observation that mutants containing a cytosine at the -5 position have much lower $K_d$ than all the other mutations. Furthermore, in the complex structure, the amino group of the cytosine at position -5 makes an intramolecular hydrogen bond that increases the propensity of the free RNA to adopt the structure seen in the complex (198), as shown in Figure 4.7c. Such an intramolecular interaction may lead to a preferable contribution to the binding energy; because the distance potential currently measures only intermolecular interactions, it is unable to capture the thermodynamic effect of interactions within the RNA or protein, and of mutation-induced changes in RNA (or protein) structure. The good correlations of distance potential (when sequence mutations are grouped according to the base identity at position –5), suggests strongly that the potential is able to capture the energetic contributions of intermolecular interactions.

The same limitations observed in the MS2 mutation data led to the failures in prediction for RNA mutations in the Fox1 and SRP complexes. In the structure of the Fox-1 complex, nucleotide U1 interacts with C3 by forming an intra-molecular hydrogen bond, while and G2 and A4 form a non-Watson-Crick base pair (202) (Figure 4.12). Four out of the seven Fox-1 RNA mutations we tested directly affect these intramolecular interactions, which are not evaluated by the statistical potential used in this work. In the case of the RNA mutations to the SRP complex, the mutated RNA residues are located in a double-stranded region of RNA, and do not form specific interactions with the protein (203). However, these mutations are expected to alter the structure of the RNA itself, and are likely to introduce conformational changes in the structure of the complex, as well. Again, the effect of these changes in RNA conformation cannot be captured by the intermolecular potential function used here.

Given these observations, it is reasonable to conclude that the omission of protein intramolecular contacts might also limit the predictive power of the method. Nevertheless, we have been able to obtain good predictions for the energetics of a diverse set of protein mutations in our tests; this result may indicate that protein-protein interactions involving interface residues do not contribute significantly to the energetics of protein-RNA recognition. In order to address this question, we are exploring the impact of incorporating intramolecular terms (both protein-protein and RNA-RNA) into the potential function.

### 4.4.4 The Effect of Training Set Composition on Potential Function Performance

All knowledge-based potentials face the possibility of unintentional bias, because their training depends upon the selection of a representative sample of structural data. If great care is not exercised to ensure that this training set is unbiased (*i.e.* structurally heterogeneous), it is possible to create a statistical potential that unfairly scores certain structures more favorably than others simply because they are over-represented in the training set.

The challenge of over-fitting is particularly acute for protein-RNA interactions, as there are relatively few high-resolution structures of protein-RNA complexes. Because of this limitation, a combined training/test set was used in this work. In order to avoid bias, a leave-one-out cross-validation strategy was employed: the tested structure was always excluded from the training set. Thus, every test in this work was conducted with a different score, trained using only those structures that were not homologous to the tested protein-RNA complex.

Unfortunately, this strategy leads to situations where there is insufficient training data to capture particular structural phenomena. For example, we observe virtually no correlation between the distance-dependent score and the experimental binding affinity for mutations of U1A protein until the U1A complex structure is added to the training set (Table 4.3). Addition of the homologous U2B" complex structure (PDB code: 1A9N) to the training set in place of U1A can improve these results, indicating that the training set

is missing critical structural information that would help to discriminate native-like contacts unique to the U1A complex (an unusually high-affinity RRM, with a long, seven-nucleotide recognition sequence)(149, 195). Nevertheless, it is important to note that the standards we have established for protein non-homology in the current training set are quite strict, and we anticipate that this problem can be greatly mitigated with the use of less stringent sequence-similarity cutoffs, without unduly biasing the results obtained by the method. Moreover, the performance of the method is certain to improve with the size of the structural database, as more high-resolution protein-RNA structures are made available for analysis.

### 4.4.5 The challenge of near native decoy discrimination

The question of how to generate and discriminate near-native decoys is still an open challenge for many areas of computational structural biology, such as docking (23, 182) and protein structure prediction (168, 209). In this work, the docking decoy sets created by Chen *et al.* contain many near-native decoys (e.g. structures < 3Å RMSD) that can be successfully discriminated by the distance-dependent potential (Figure 4.1, insets). However, when testing against the exceptionally near-native decoys generated by extracting snapshots from MD simulations (Table 4.2), we find that none of the potentials we have tested are able to reliably distinguish such non-native decoys from native structures. Moreover, not even Amber (which was used to conduct the MD simulations), was able to discriminate these very near-native structures – on average, Amber does no better than the much simpler, distance-dependent potential at this difficult task. Thus, the question of how to create a potential that is sensitive to the extremely subtle structural variations present in very near-native decoys remains a challenging and important area of research. We are hopeful that the incorporation of terms describing the higher-order geometric preferences of protein-RNA interfaces (e.g. the incorporation of a directional hydrogen-bonding potential, analogous to that used by Chen *et al.* (1)) will help to enhance the discriminatory power of our method, as will the inevitable increase in high-resolution structural data available for training. Nevertheless, the power of the

distance-dependent potential function is remarkable, given its simplicity; it is on par with the (significantly more complicated) Amber force field in many of our experiments.

## 4.5 Conclusion

We have introduced a simple, statistical potential function with the ability to discriminate the structures of native protein-RNA complexes from near-native decoys, to reproduce experimentally determined binding affinities for a number of RNA-binding proteins, and to predict cognate binding sequences for a large set of protein-RNA complexes. Despite the simplicity of our approach, this statistical potential performs as well as a significantly more complicated, knowledge-based potential function in tests of docking decoy discrimination. Moreover, in tests with near-native decoy structures, the statistical potential is strongly correlated with the physics-based potential function used by the AMBER package.

This work represents the first large-scale, quantitative investigation of the structural phenomena responsible for sequence-specific protein-RNA recognition. Several results indicate that the performance of this method is on par with the significantly more complicated methods that are widely used for protein design and molecular simulation, yet this work represents only a preliminary study of the approach, and we have made a number of simplifying assumptions that almost certainly negatively affect our results. We anticipate that the performance of the approach will increase with the size of the structural database, and as terms are added to the model to account for protein and RNA intra-molecular interactions that are currently ignored. Nevertheless, even in its current, naïve implementation, this simple, statistical model achieves a high degree of sensitivity to subtle changes in protein-RNA interface structure. We are optimistic that this knowledge-based potential function will find broad application to problems requiring the high-resolution modeling of protein-RNA interfaces, such as structure-based genome annotation, or the rational design of novel RNA-binding proteins.

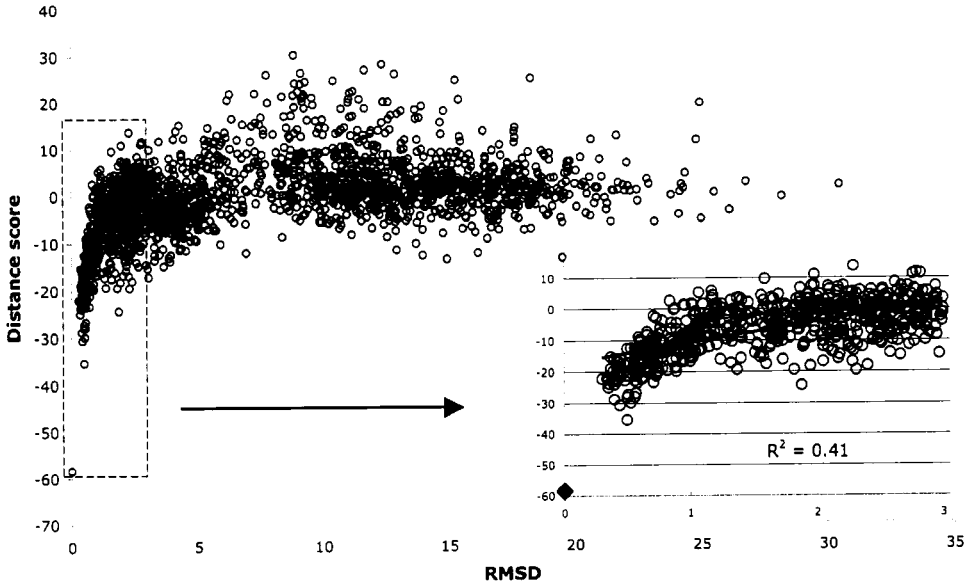



**Figure 4.1: Score-RMSD plot for the Poly-A binding protein (PDB: 1CVJ) docking decoy set.**

The score generated by the distance-dependent potential (in arbitrary units) is plotted vs the deviation from the native structure (open circle at RMSD=0). An enlarged view of the near-native decoys (0-3Å RMSD) is shown as an inset.

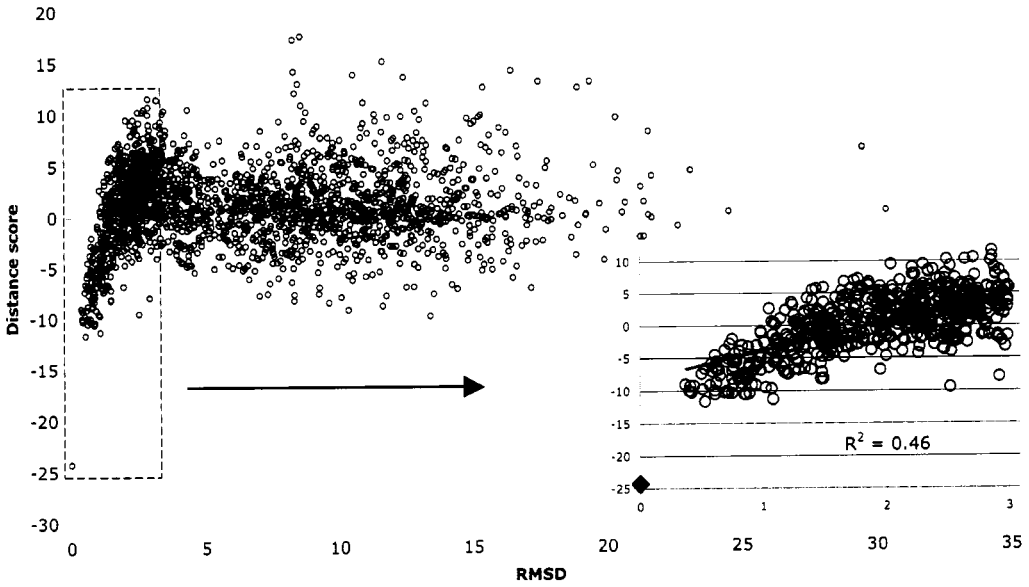



**Figure 4.2: Score-RMSD plot for the Nova-2 K-Homology domain 3 (PDB: 1EC6) docking decoy set.**

The score generated by the distance-dependent potential (in arbitrary units) is plotted vs the deviation from the native structure (open circle at RMSD=0). An enlarged view of the near-native decoys (0-3Å RMSD) is shown as an inset.

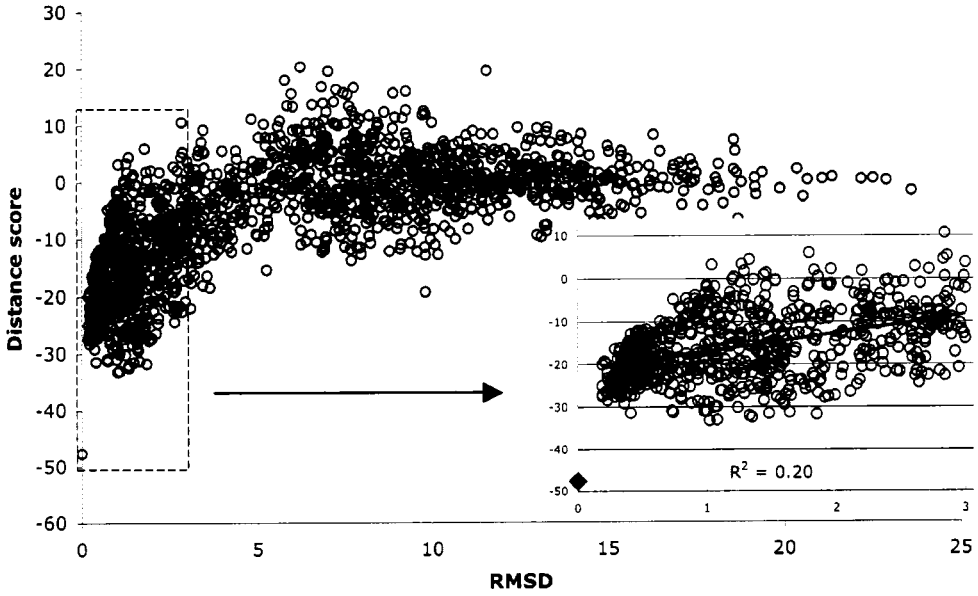



**Figure 4.3: Score-RMSD plot for the HuD protein (PDB: 1FXL) docking decoy set.** The score generated by the distance-dependent potential (in arbitrary units) is plotted vs the deviation from the native structure (open circle at RMSD=0). An enlarged view of the near-native decoys (0-3Å RMSD) is shown as an inset.

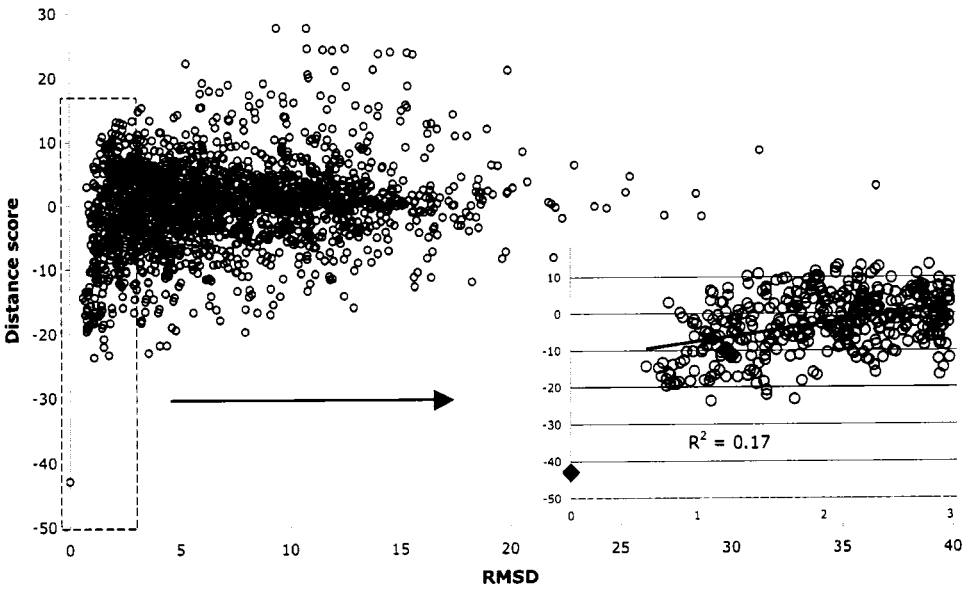



**Figure 4.4: Score-RMSD plot for the SRP19 protein (PDB: 1JID) docking decoy set.** The score generated by the distance-dependent potential (in arbitrary units) is plotted vs the deviation from the native structure (open circle at RMSD=0). An enlarged view of the near-native decoys (0-3Å RMSD) is shown as an inset.

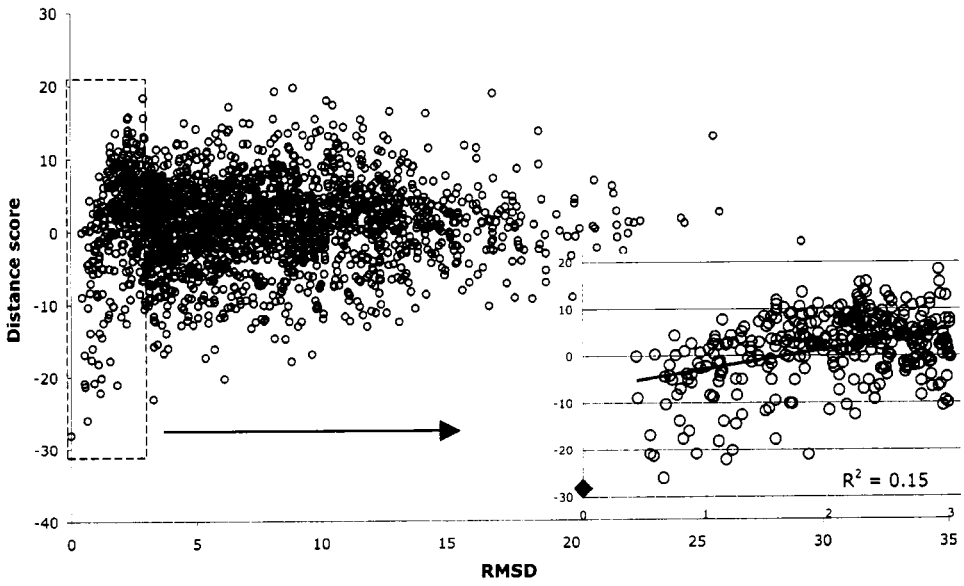
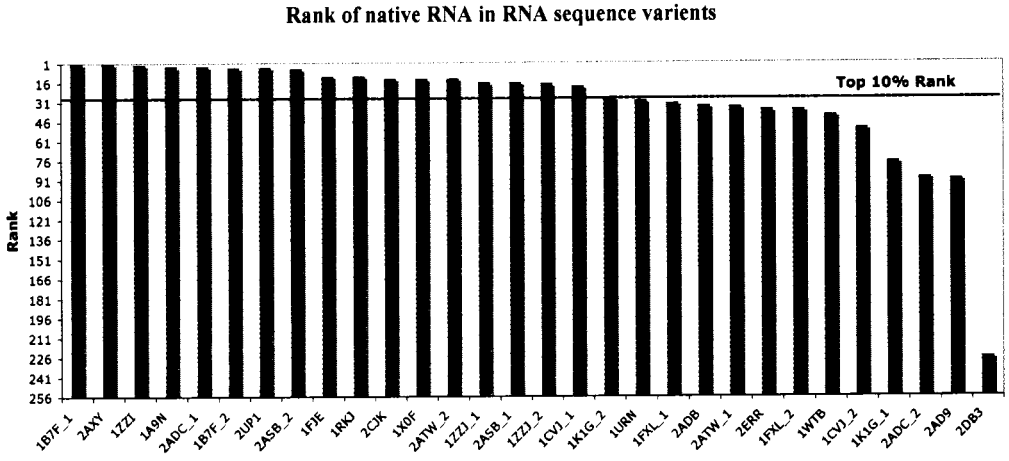
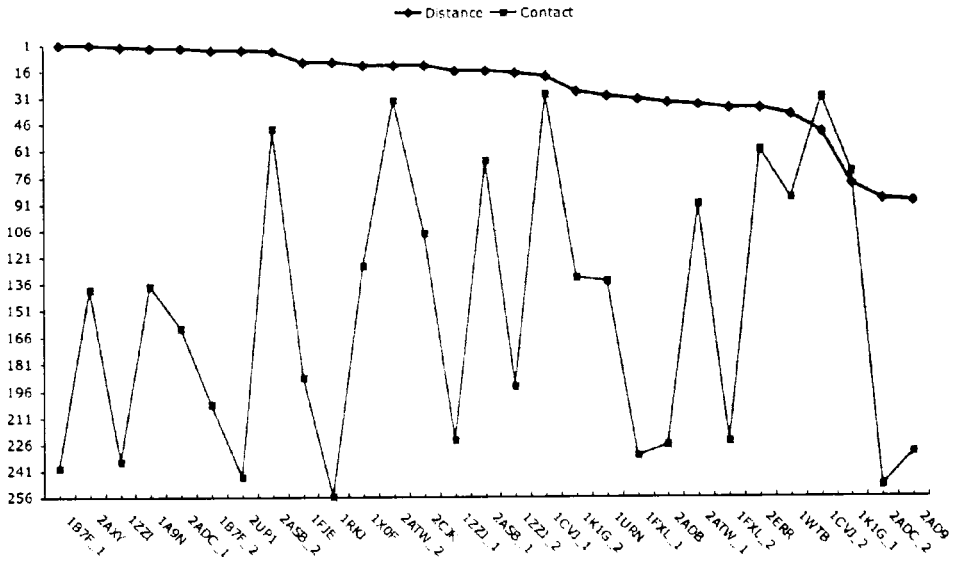



**Figure 4.5: Score-RMSD plot for the U1A protein (PDB: 1URN) docking decoy set.** The score generated by the distance-dependent potential (in arbitrary units) is plotted vs the deviation from the native structure (open circle at RMSD=0). An enlarged view of the near-native decoys (0-3Å RMSD) is shown as an inset.




(a)




(b)

**Figure 4.6: Structure-based identification of RRM recognition sequences.**
(a) The cognate sequence rank by distance potential (cutoff = 6Å) in RRM/KH domain sequence decoy sets. The dashed line represents the 10th percentile rank. (b) Comparison of the discrimination ability of the distance potential and the contact-counting method.

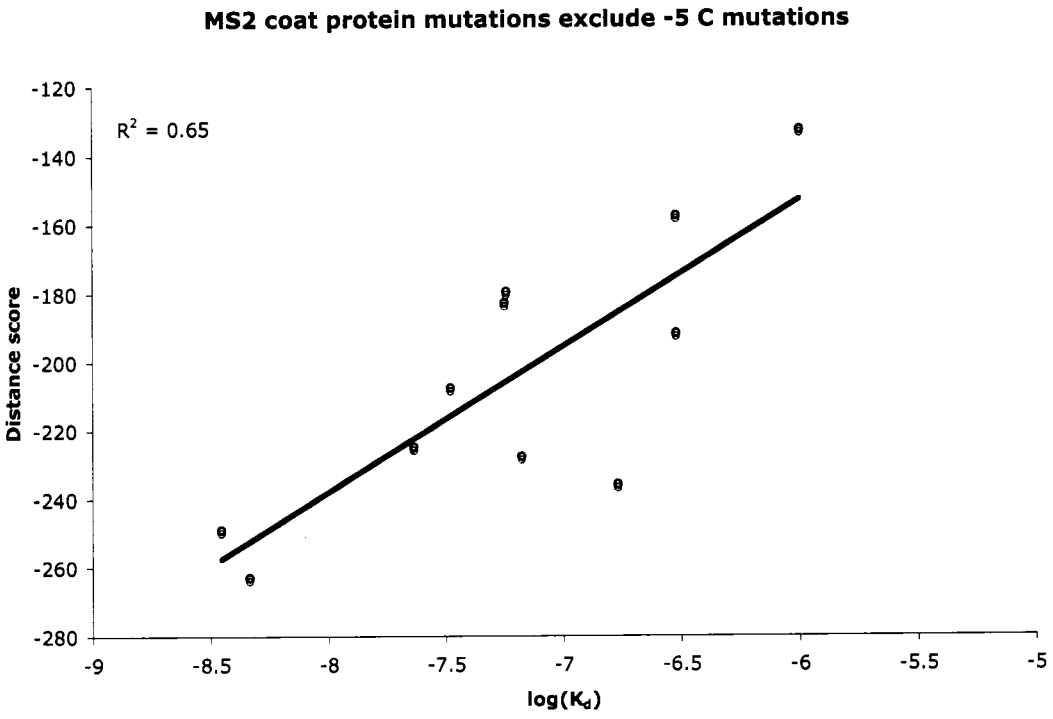



**Figure 4.7: Correlation between scores generated by the distance-dependent statistical potential and experimental binding free energies (logKd) for mutants of the MS2 coat protein containing nucleotides other than cytosine at position -5 of the RNA molecule.**

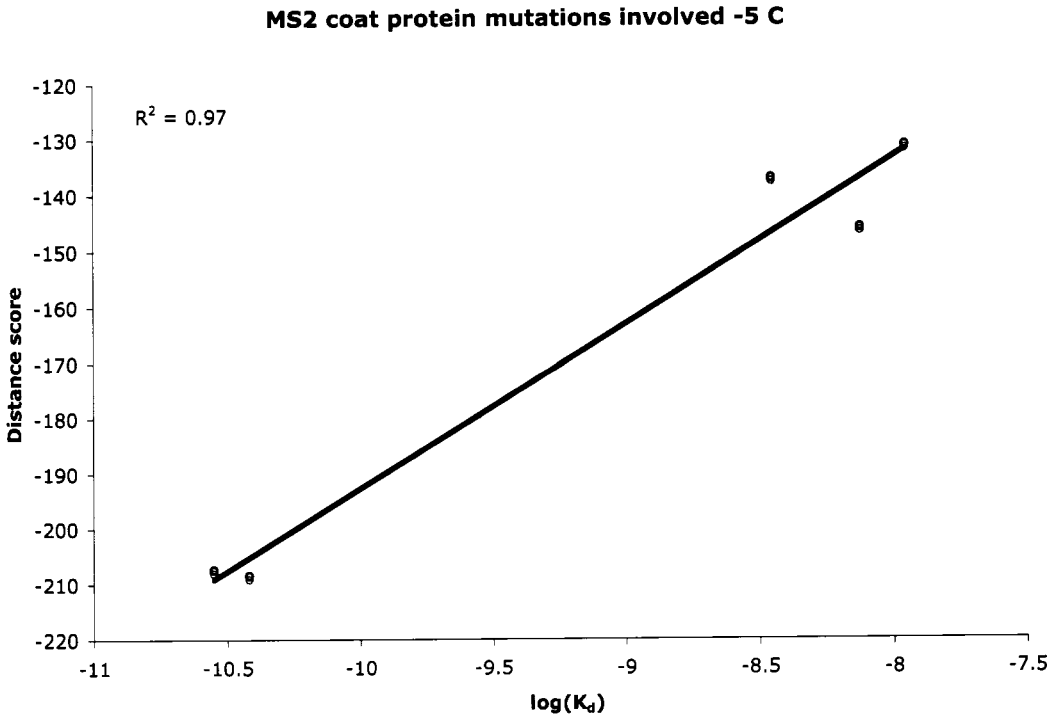



**Figure 4.8: Correlation between scores generated by the distance-dependent statistical potential, and experimental binding free energies for mutants of the MS2 coat protein complex containing cytosine at position -5 of the RNA molecule.**

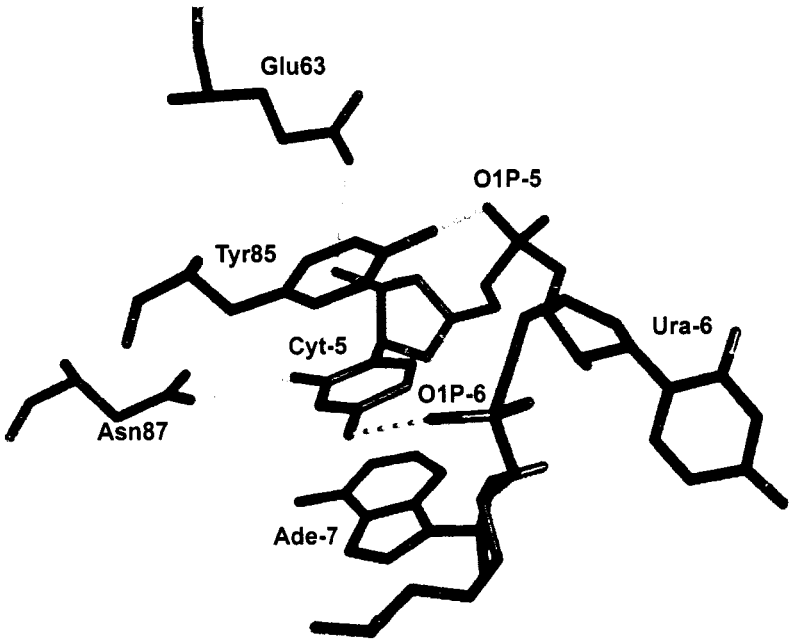



**Figure 4.9: The characteristic hydrogen bond between the amino group of cytosine -5 and the phosphate oxygen of uracil -6 observed in the structure of the MS2 coat protein complex with RNA.**

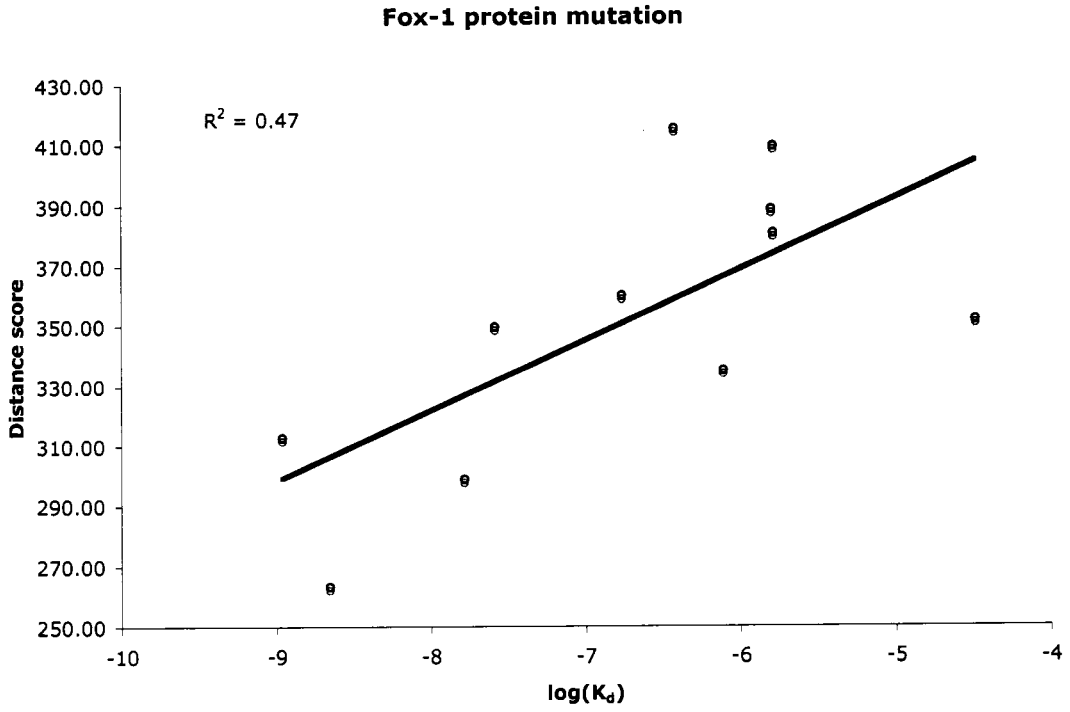



**Figure 4.10: Correlation between scores generated by the distance-dependent statistical potential and experimental binding free energies (logKd) for mutants of the Fox1 protein.**

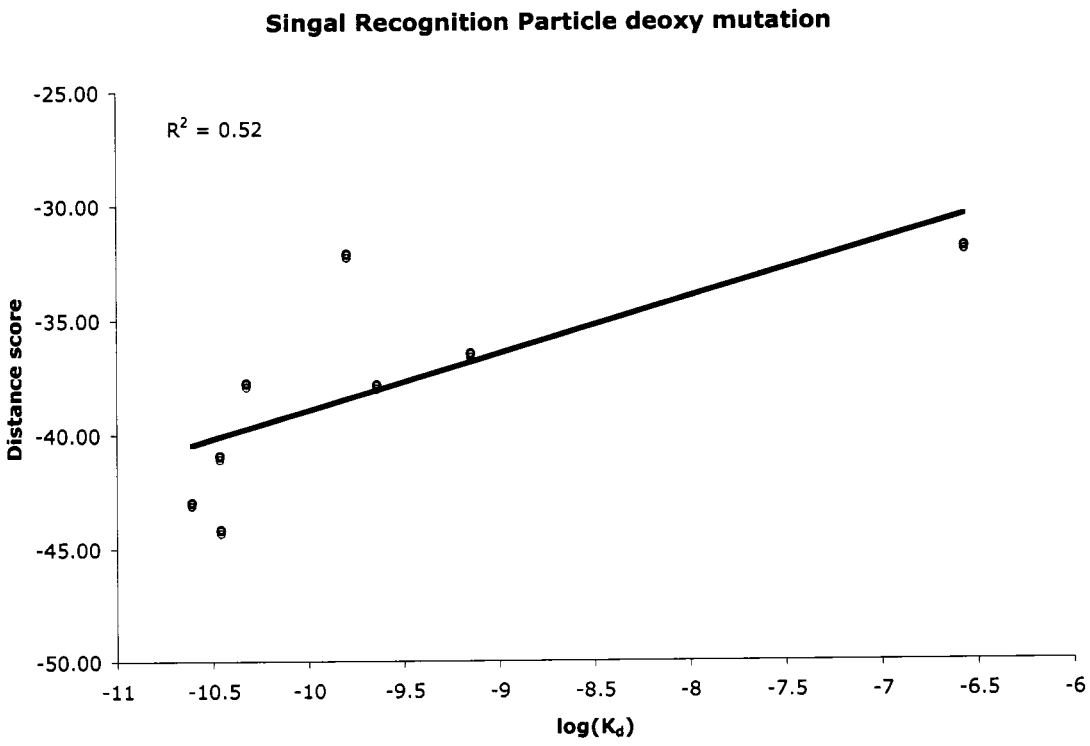



**Figure 4.11: Correlation between scores generated by the distance-dependent statistical potential and experimental binding free energies (logKd) for ribose to deoxyribose mutants of a universally conserved protein component of the Signal Recognition Particle.**

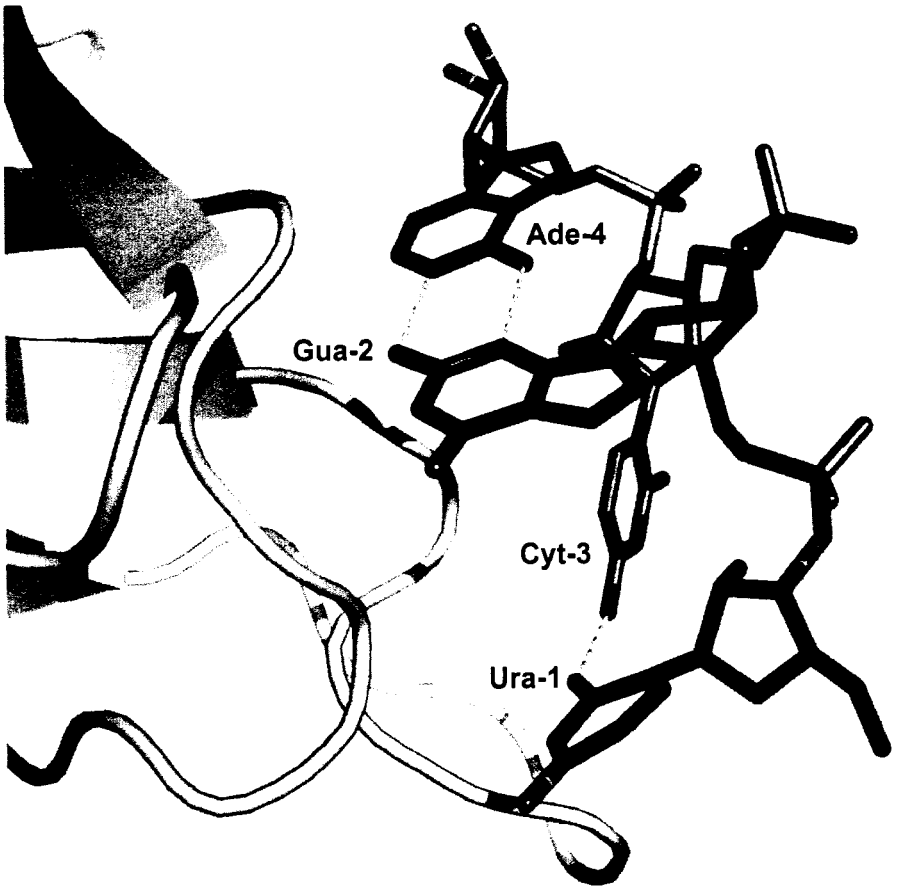



**Figure 4.12: The intramolecular hydrogen bond between uracil-1 and cytosine-3, and the non-Watson-Crick base pair between guanine-2 and adenine-4 for the RNA in complex with Fox-1 protein (PDB code: 2ERR).**

**Table 4.1: Native Z-scores and score-RMSD correlation coefficients for the protein-RNA docking decoy sets prepared by Chen *et al.***

| | Z-Score | | | |
|---|---|---|---|---|
| | Distance-Dependent[a] | Coulombic[b] | Pseudo-HB[b] | Contact Count[b] |
| 1CVJ | -7.02 | -1.19 | -5.11 | -2.44 |
| 1EC6 | -6.46 | -1.09 | -6.53 | -3.00 |
| 1FXL | -2.66 | -1.55 | -2.70 | -1.26 |
| 1JID | -6.29 | -1.36 | -9.12 | -3.09 |
| 1URN | -4.80 | -1.35 | -8.39 | -3.39 |
| Mean | -5.45 | -1.31 | -6.37 | -2.64 |
| σ | 1.76 | 0.18 | 2.58 | 0.84 |

[a] Using a 6Å contact cutoff.

[b] From Chen *et al.*

**Table 4.2: Z-scores and correlations for observed for near-native decoys generated by MD simulation.**

| | Largest RMSD (Å) | Z-Scores | | $RMSD\ R^2$ | | Distance-Dependent vs. Amber ($R^2$) |
|---|---|---|---|---|---|---|
| | | Distance-Dependent | Amber | Distance-Dependent | Amber | |
| 1B7F | 2.33 | -3.51 | -3.01 | 0.34 | 0.31 | 0.37 |
| 1CVJ | 2.63 | -1.13 | -0.44 | 0.01 | -0.02 | 0.51 |
| 1DFU | 2.45 | -1.82 | -1.93 | 0.07 | 0.08 | 0.39 |
| 1E7K | 2.58 | -0.38 | 0.83 | -0.03 | -0.13 | 0.34 |
| 1EC6 | 5.21 | -3.02 | -2.29 | 0.5 | 0.83 | 0.71 |
| 1FJE | 3.54 | 1.49 | 2.51 | -0.24 | -0.27 | 0.42 |
| 1FXL | 2.22 | -1.32 | -1.34 | 0.05 | 0.11 | 0.47 |
| 1JBS | 2.46 | -0.31 | 0.60 | 0.00 | -0.23 | 0.38 |
| 1JID | 2.92 | 0.26 | 1.05 | -0.19 | -0.24 | 0.54 |
| 1K1G | 3.48 | 0.52 | 1.95 | -0.1 | -0.26 | 0.52 |
| 1KNZ | 2.44 | -1.39 | -1.69 | 0.00 | 0.22 | 0.24 |
| 1M8W | 2.47 | -1.10 | -1.21 | 0.21 | 0.36 | 0.44 |
| 1R9F | 2.55 | -1.24 | -0.19 | 0.1 | -0.04 | 0.11 |
| 1RKJ | 3.85 | 0.92 | 3.00 | -0.08 | -0.56 | 0.19 |
| 1URN | 3.06 | -1.08 | -2.88 | 0.04 | 0.36 | 0.46 |
| 2AD9 | 3.94 | -0.68 | -2.09 | 0.00 | 0.02 | 0.42 |
| 2ADB | 2.66 | 0.37 | -2.15 | -0.24 | 0.01 | 0.16 |
| 2ADC | 3.01 | -0.21 | -0.65 | -0.23 | -0.23 | 0.63 |
| 2ASB | 2.12 | -0.94 | -2.07 | 0.23 | 0.64 | 0.48 |
| 2ATW | 2.19 | -1.38 | -2.94 | 0.19 | 0.76 | 0.33 |
| 2CJK | 2.75 | 1.42 | 2.47 | -0.36 | -0.65 | 0.42 |
| Mean | | -0.69 | -0.59 | 0.01 | 0.05 | 0.41 |
| $\sigma$ | | 1.28 | 1.94 | 0.21 | 0.39 | 0.15 |

**Table 4.3: Correlations observed between the distance-dependent score and the experimental free energy of binding for several mutant protein-RNA complexes.**

| | Distance-dependent | | | Contact Counting | | |
|---|---|---|---|---|---|---|
| | 6Å | 10Å | 12Å | 6Å | 10Å | 12Å |
| **Protein mutations** | | | | | | |
| MS2 mutations (w/o cytosine-5) | 0.43 | 0.50 | 0.65 | 0.19 | 0.10 | 0.08 |
| MS2 mutations (w/ cytosine-5) | 0.81 | 0.81 | 0.97 | 0.43 | 0.14 | 0.09 |
| U1A protein mutations [a] | 0.27 | 0.48 | 0.65 | 0.29 | -0.06 | -0.03 |
| U1A protein mutations [b] | 0.04 | 0.14 | 0.39 | 0.29 | -0.06 | -0.03 |
| Fox-1 complex | 0.40 | 0.45 | 0.47 | 0.47 | 0.43 | 0.42 |
| **RNA mutations** | | | | | | |
| Fox-1 complex | 0.20 | -0.39 | -0.57 | 0.30 | 0.33 | 0.35 |
| Signal Recognition Particle; DNA mutants | 0.87 | 0.56 | 0.52 | 0.36 | 0.30 | 0.29 |
| Signal Recognition Particle; RNA mutants | -0.07 | -0.03 | -0.07 | 0.01 | 0.07 | 0.05 |

[a] The native U1A complex was included in the training set for this experiment.

[b] The U2B" complex (U1A homolog) was included in the training set for this experiment.

# BIBLIOGRAPHY

1. Chen Y, Kortemme T, Robertson T, Baker D, Varani G. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. Nucleic Acids Res. 2004;32(17):5147-62.

2. Robertson TA, Varani G. An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. Proteins. 2007;66:359-74.

3. Seeman N, Rosenberg J, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. Proceedings of the National Academy of Science. 1976;73:804-8.

4. Pabo C, Sauer R. Protein-DNA recognition. Annu Rev Biochem. 1984;53:293-321.

5. Pabo C, Nekludova L. Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition. J Mol Biol. 2000;301(3):597-624.

6. Matthews B. Protein-DNA interaction. no code for recognition. Nature. 1988;335(6188):294-5.

7. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein-DNA complexes. Genome Biol. 2000;1(1).

8. Jones S, Daley D, Luscombe N, Berman H, Thornton J. Protein-RNA interactions: A structural analysis. Nucleic Acids Res. 2001;29(4):943-54.

9. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: A three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. 2001;29(13):2860-74.

10. Jones S, van Heyningen P, Berman H, Thornton J. Protein-DNA interactions: A structural analysis. J Mol Biol. 1999;287(5):877-96.

11. Luscombe NM, Thornton JM. Protein-DNA interactions: Amino acid conservation and the effects of mutations on binding specificity. J Mol Biol. 2002;320(5):991-1009.

12. Mirny L, Gelfand M. Structural analysis of conserved base pairs in protein-DNA complexes. Nucleic Acids Res. 2002;30(7):1704-11.

13. Sippl M. Calculation of conformational ensembles from potentials of mean force. J Mol Biol. 1990;213:859-83.

14. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol. 1998;275:895-916.

15. Xu Y, Xu D, Uberbacher E. An efficient computational method for globally optimal threading. Journal of Computational Biology. 1998;5:597-614.

16. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins. 2001;44:223-32.

17. Skolnick J. In quest of an empirical potential for protein structure prediction. Curr Opin Struct Biol. 2006;16(2):166-71.

18. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science. 2002;11:2714-26.

19. Zhang C, Liu S, Zhou H, Zhou Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Science. 2004;13:400-11.

20. DeWitte R, Shakhnovich E. SMoG: De novo design method based on simple, fast and accurate free energy estimates. J Am Chem Soc. 1996;118:11733-44.

21. Ischchenko A, Shakhnovich E. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions. J Med Chem. 2002;45:2770-80.

22. Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem. 2005;48(7):2325-35.

23. Velec H, Gohlke H, Klebe G. DrugScoreCSD--knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J Med Chem. 2005;48:6296-303.

24. Jiang L, Gao Y, Mao F, Liu Z, Lai L. Potential of mean force for protein-protein interactions studies. Proteins. 2002;46:190-6.

25. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys J. 2003;84:1895-901.

26. Kono H, Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. Proteins. 1999;35(1):114-31.

27. Endres RG, Schulthess TC, Wingreen NS. Toward and atomistic model for predicting transcription-factor binding sites. Proteins. 2004;57(2):262-8.

28. Havranek JJ, Duarte CM, Baker D. A simple physical model for the prediction and design of protein-DNA interactions. J Mol Biol. 2004;344:59-70.

29. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res. 2005;33(1):5781-98.

30. Donald J, Chen W, Shakhnovich E. Energetics of protein-DNA interactions. Nucleic Acids Res. 2007;35:1039-47.

31. Schneider T, Stephens R. Sequence logos: A new way to display consensus sequences. Nucleic Acids Res. 1990;18:6097-100.

32. Matys V, Fricke E., Geffers R, Gossling E, Haubrock M, Hehl R, et al. TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003;31:374-8.

33. Weiner S, Kollman P, Case D, Singh U, Ghio C, Alagona G, et al. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc. 1984;106:765-84.

34. Weiner S, Kollman P, Nguyen D, Case D. An all-atom force field for simulations of proteins and nucleic acids. Journal of Computational Chemistry. 1986;7:230-52.

35. Cornell W, Cieplak P., Bayly C, Gould I, Merz Jr. K, Ferguson D, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc. 1995;117:5179-97.

36. Langley D. Molecular dynamic simulations of environment and sequence dependent DNA conformations: The development of the BMS nucleic acid force field and comparison with experimental results. Journal of Biomolecular Structure and Dynamics. 1998;16:487-509.

37. Cheatham III T, Cieplak P, Kollman P. A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat. Journal of Biomolecular Structure and Dynamics. 1999;16:845-862.

38. Foloppe N, MacKerell A. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. Journal of Computational Chemistry. 2000;21(2):86-104.

39. MacKerell A, Banavali N. All-atom empirical force field for nucleic acids: II. application to molecular dynamics simulations of DNA and RNA in solution. Journal of Computational Chemistry. 2000;21(2):105-20.

40. Wang J, Cieplak P, Kollman P. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? Journal of Computational Chemistry. 2000;21:1049-74.

41. MacKerell A, Banavali N, Foloppe N. Development and current status of the CHARMM force field for nucleic acids. Biopolymers. 2001;56(4):257-65.

42. Lebrun A, Lavery R. Modeling DNA deformations induced by minor groove binding proteins. Biopolymers. 1999;49:341-53.

43. Pastor N, MacKerell AD, Weinstein H. TIT for TAT: The properties of inosine and adenosine in TATA box DNA. J Biomol Struct Dyn. 1999;16:787-810.

44. Lebrun A, Lavery R, Weinstein H. Modeling multi-component protein-DNA complexes: The role of bending and dimerization in the complex of p53 dimers with DNA. Protein Eng. 2001;14:233-43.

45. Marco E, Garcia-Nieto R, Gago F. Assessment by molecular dynamics simulations of the structural determinants of DNA-binding specificity for transcription factor Sp1. J Mol Biol. 2003;328:9-32.

46. Gorfe A, Caflisch A, Jelesarov I. The role of flexibility and hydration on the sequence-specific DNA recognition by the Tn916 integrase protein: A molecular dynamics analysis. J Mol Recognit. 2004;17:120-31.

47. Gutmanas A, Billeter M. Specific DNA recognition by the antp homeodomain: MD simulations of specific and nonspecific complexes. Proteins. 2004;57:772-82.

48. Pichierri F, Aida M, Gromiha MM, Sarai A. Free-energy maps of base-amino acid interactions for DNA-protein recognition. J Am Chem Soc. 1999;121(26):6152-7.

49. Sayano K, Kono H, Gromiha M, Sarai A. Multicanonical monte carlo calculation of the free-energy map of the base-amino acid interaction. Journal of Computational Chemistry. 2000;21(11):954-62.

50. Yoshida T, Nishimura T, Aida M, Pichierri F, Gromiha MM, Sarai A. Evaluation of free energy landscape for base-amino acid interactions using ab initio force field and extensive sampling. Biopolymers. 2002;61:84-95.

51. Thayer KM, Beveridge DL. Hidden markov models from molecular dynamics simulations on DNA. Proc Natl Acad Sci U S A. 2002;99(1):8642-7.

52. Paillard G, Lavery R. Analyzing protein-DNA recognition mechanisms. Structure. 2004;12(1):113-22.

53. Huang N, MacKerell A. Specificity in protein-DNA interactions: Energetic recognition by the (cytosine-C5)-methyltransferase from HhaI. J Mol Biol. 2005;345(2):265-74.

54. Kortemme T, Morozov A, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol. 2003;326(4):1239-59.

55. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol. 1995;5(2):229-35.

56. Godzik A. Knowledge-based potentials for protein folding: What can we learn from known protein structures? Structure. 1996;4(4):363-6.

57. Rojnuckarin A, Subramaniam S. Knowledge-based interaction potentials for proteins. Proteins. 1999;36(1):54-67.

58. Zhang C, Liu S, Zhou H, Zhou Y. The dependence of all-atom statistical potentials on structural training database. Biophys J. 2004;86:3349-58.

59. Liu Z, Mao F, Guo J, Yan B, Wang P, Qu Y, et al. Quantitative evaluation of protein-DNA interactions using and optimized knowledge-based potential. Nucleic Acids Res. 2005;33(2):546-58.

60. Ge W, Schneider B, Olson W. Knowledge-based elastic potentials for docking drugs or proteins with nucleic acids. Biophys J. 2005;88:1166-90.

61. Kaplan T, Friedman N, Margalit H. Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput Biol. 2005;1(1).

62. Dickerson RE. DNA bending: The prevalence of kinkiness and the virtues of normality. Nucleic Acids Res. 1998;26(8):1906-26.

63. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Science. 2000;97(19):10383-8.

64. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Stoddard BL, et al. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature. 2006;441(7093):656-9.

65. Dunbrack RL. Rotamer libraries in the 21st century. Curr Opin Struct Biol. 2002;12(4):431-40.

66. Knegtel RM, Antoon J, Rullmann C, Boelens R, Kaptein R. MONTY: A monte carlo approach to protein-DNA recognition. J Mol Biol. 1994;235(1):318-24.

67. Knegtel RM, Boelens R, Kaptein R. Monte carlo docking of protein-DNA complexes: Incorporation of DNA flexibility and experimental data. Protein Eng. 1994;7(6):761-7.

68. Aloy P, Moont G, Gabb H, Querol E, Aviles F, Sternberg M. Modelling repressor proteins docking to DNA. Proteins. 1998;33(4):535-49.

69. Sternberg M, Aloy P, Gabb H, Jackson R, Moont G, Querol E, et al. A computational system for modelling flexible protein-protein and protein-DNA docking. Proceedings of the International Conference on Intelligent Systems in Molecular Biology. 1998;6:183-92.

70. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol. 2003;331(1):281-99.

71. van Dijk M, van Dijk ADJ, Hsu V, Boelens R, Bonvin AMJJ. Information-driven protein-DNA docking using HADDOCK: It is a matter of flexibility. Nucl Acids Res. 2006;34(11):3317-25.

72. Gromiha MM, Siebers JG, Selvaraj S, Kono H, Sarai A. Role of inter and intramolecular interactions in protein-DNA recognition. Gene. 2005;364:108-13.

73. Steffen N, Murphy S, Tolleri L, Hatfield G, Lathrop R. DNA sequence and structure: Direct and indirect recognition in protein-DNA binding. Bioinformatics. 2002;18 Suppl 1:22-.

74. Steffen NR, Murphy SD, Lathrop RH, Opel ML, Tolleri L, Hatfield GW. The role of DNA deformation energy at individual base steps for the identification of DNA-protein binding sites. Genome Inform. 2002;13:153-62.

75. Dixit SB, Andrews DQ, Beveridge DL. Induced fit and the entropy of structural adaptation in the complexation of CAP and lambda-repressor with cognate DNA sequences. Biophys J. 2005;88:3147-57.

76. Arauzo-Bravo MJ, Fujii S, Kono H, Ahmad S, Sarai A. Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: Toward understanding the indirect readout mechanism in protein-DNA recognition. J Am Chem Soc. 2005;127(4):16074-89.

77. Napoli AA, Lawson CL, Ebright RH, Berman HM. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: Recognition of pyrimidine-purine and purine-purine steps. J Mol Biol. 2006;357(1):173-83.

78. Aeling KA, Opel ML, Steffen NR, Tretyachenko-Ladokhina V, Hatfield GW, Lathrop RH, et al. Indirect recognition in sequence-specific DNA binding by escherichia coli integration host factor: The role of DNA deformation energy. J Biol Chem. 2006;281(51):39236-48.

79. Aeling KA, Steffen NR, Johnson M, Hatfield GW, Lathrop RH, Senear DF. DNA deformation energy as an indirect recognition mechanism in protein-DNA interactions. IEEE/ACM Trans Comput Biol Bioinform. 2007;4(1):117-25.

80. Kuszewski J, Schwieters C, Clore GM. Improving the accuracy of NMR structures of DNA by means of a database potential of mean force describing base-base positional interactions. J Am Chem Soc. 2001;123(17):3903-18.

81. Clore GM, Kuszewski J. Improving the accuracy of NMR structures of RNA by means of conformational database potentials of mean force as assessed by complete dipolar coupling cross-validation. J Am Chem Soc. 2003;125(6):1518-25.

82. Morozov AV, Siggia ED. Connecting protein structure with predictions of regulatory sites. PNAS. 2007;104(17):7068-73.

83. Pomerantz JL, Sharp PA, Pabo CO. Structure-based design of transcription factors. Science. 1995;267(5194):93-6.

84. Pomerantz JL, Pabo CO, Sharp PA. Analysis of homeodomain function by structure-based design of a transcription factor. Proc Natl Acad Sci U S A. 1995;92(21):9752-6.

85. Isalan M, Choo Y, Klug A. Synergy between adjacent zinc fingers in sequence-specific DNA recognition. Proc Natl Acad Sci U S A. 1997;94(11):5617-21.

86. Kim JS, Kim J, Cepek KL, Sharp PA, Pabo CO. Design of TATA box-binding protein/zinc finger fusions for targeted regulation of gene expression. Proc Natl Acad Sci U S A. 1997;94(8):3616-20.

87. Liu Q, Segal DJ, Ghiara JB, Barbas CF. Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. Proc Natl Acad Sci U S A. 1997;94(11):5525-30.

88. Kim JS, Pabo CO. Getting a handhold on DNA: Design of poly-zinc finger proteins with femtomolar dissociation constants. Proc Natl Acad Sci U S A. 1998;95(6):2812-7.

89. Pomerantz JL, Wolfe SA, Pabo CO. Structure-based design of a dimeric zinc finger protein. Biochemistry (N Y ). 1998;37(4):965-70.

90. Wolfe SA, Ramm EI, Pabo CO. Combining structure-based design with phage display to create new cys(2)his(2) zinc finger dimers. Structure. 2000;8(7):739-50.

91. Peisach E, Pabo CO. Constraints for zinc finger linker design as inferred from X-ray crystal structure of tandem Zif268-DNA complexes. J Mol Biol. 2003;330(1):1-7.

92. Wolfe SA, Grant RA, Pabo CO. Structure of a designed dimeric zinc finger protein bound to DNA. Biochemistry (N Y ). 2003;42(46):13401-9.

93. Mandell JG, Barbas CF. Zinc finger tools: Custom DNA-binding domains for transcription factors and nucleases. Nucleic Acids Res. 2006;34:516-23.

94. Papworth M, Kolasinska P, Minczuk M. Designer zinc-finger proteins and their applications. Gene. 2006;366(1):27-38.

95. Nomura W, Sugiura Y. Design and synthesis of artificial zinc finger proteins. Methods Mol Biol. 2007;352:83-93.

96. Klug A. Towards therapeutic applications of engineered zinc finger proteins. FEBS Lett. 2005;579(4):892-4.

97. Dobson N, Dantas G, Baker D, Varani G. High-resolution structural validation of the computational redesign of human U1A protein. Structure. 2006;14(5):847-56.

98. Lichtarge O, Yamamoto KR, Cohen FE. Identification of functional surfaces of the zinc binding domains of intracellular receptors. J Mol Biol. 1997;274(3):325-37.

99. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. J Mol Biol. 2002;321(1):7-20.

100. Raviscioni M, Gu P, Sattar M, Cooney AJ, Lichtarge O. Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity. J Mol Biol. 2005;350(3):402-15.

101. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 1996;257(2):342-58.

102. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct. 2000;29:183-212.

103. Jamieson AC, Miller JC, Pabo CO. Drug discovery with engineered zinc-finger proteins. Nat Rev Drug Discov. 2003;2(5):361-8.

104. Laird-Offringa IA, Belasco JG. Analysis of RNA-binding proteins by in vitro genetic selection: Identification of an amino acid residue important for locking U1A onto its RNA target. Proc Natl Acad Sci U S A. 1995;92(25):11859-63.

105. Laird-Offringa IA, Belasco JG. RNA-binding proteins tamed. Nat Struct Biol. 1998;5(8):665-8.

106. MichËle Treger EW. Statistical analysis of atomic contacts at RNA-protein interfaces. Journal of Molecular Recognition. 2001;14(4):199-214.

107. Allers J, Shamoo Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. Journal of Molecular Biology. 2001;311(1):75-86.

108. Nadassy K, Wodak SJ, Janin J. Structural features of protein-nucleic acid recognition sites. Biochemistry. 1999;38(7):1999-2017.

109. Cheng AC, Chen WW, Fuhrmann CN, Frankel AD. Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. Journal of Molecular Biology. 2003;327(4):781-96.

110. Choo Y, Klug A. Physical basis of a protein-DNA recognition code. Curr Opin Struct Biol. 1997;7(1):117-25.

111. Rooman M, Lievin J, Buisine E, Wintjens R. Cation-pi/H-bond stair motifs at protein-DNA interfaces. J Mol Biol. 2002;319(1):67-76.

112. Wintjens R, Lievin J, Rooman M, Buisine E. Contribution of cation-pi interactions to the stability of protein-DNA complexes. J Mol Biol. 2000;302(2):395-410.

113. Nadassy K, Tomas-Oliveira I, Alberts I, Janin J, Wodak SJ. Standard atomic volumes in double-stranded DNA and packing in protein-DNA interfaces. Nucl. Acids Res. 2001;29(16):3362-76.

114. Walberer BJ, Cheng AC, Frankel AD. Structural diversity and isomorphism of hydrogen-bonded base interactions in nucleic acids. J Mol Biol. 2003;327(4):767-80.

115. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucl. Acids Res. 2003;31(24):7189-98.

116. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol. 1999;285(5):2177-98.

117. Schreiber G. Kinetic studies of protein-protein interactions. Curr Opin Struct Biol. 2002;12(1):41-7.

118. Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, Schwab MS, et al. Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Biol. 2002;9(8):621-7.

119. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science. 1998;282(5393):1462-7.

120. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science. 2003;302(5649):1364-8.

121. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature. 2003;423(6936):185-90.

122. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. J Mol Biol. 2003;332(2):449-60.

123. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins. 1999;35(2):133-52.

124. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol. 1996;258(2):367-92.

125. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A. 2003;100(10):5772-7.

126. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature. 1986;319(6050):199-203.

127. Lumb KJ, Kim PS. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. Biochemistry (N Y ). 1995;34(27):8642-8.

128. Petrey D, Honig B. Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci. 2000;9(11):2181-91.

129. McGuire RF, Momany FA, Scheraga HA. Energy parameters in polypeptides. V. an empirical hydrogen bond potential function based on molecular orbital calculations. J Phys Chem. 1972;76(3):375-93.

130. Wiberg KB, Marquez M, Castejon H. Lone pairs in carbonyl compounds and ethers. J Org Chem. 1994;59:6817-22.

131. Neria E, Fischer S, Karplus M. Simulation of activation free energies in molecular systems. J Chem Phys. 1996;105:1902-21.

132. Buck M, Karplus M. Hydrogen bond energetics: A simulation and statistical analysis of N-methyl acetamide (NMA), water and human lysozyme. J Phys Chem Ser B. 2001;105:11000-15.

133. Grzybowski B, Ischchenko A, DeWitte R, Whitesides G, Shakhnovich E. Development of a knowledge-based potential for crystals of small organic molecules: Calculation of energy surfaces for C=O...H-N hyrogen bonds. J Phys Chem Ser B. 2000;104:7293-8.

134. Sippl MJ. Helmholtz free energy of peptide hydrogen bonds in proteins. J Mol Biol. 1996;260(5):644-8.

135. Mills JE, Dean PM. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. J Comput Aided Mol Des. 1996;10(6):607-22.

136. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. J Comput Aided Mol Des. 1993;7(4):473-501.

137. Grzybowski BA, Ishchenko AV, Shimada J, Shakhnovich EI. From knowledge-based potentials to combinatorial lead design in silico. Acc Chem Res. 2002;35(5):261-9.

138. Mitchell J, Laskowski R, Alex A, Thornton J. BLEEP - potential of mean force describing protein-ligand interactions: I. generating potential. J Comput Chem. 1999;20:1165-76.

139. Ben-Naim A. Statistical potentials extracted from protein structures: Are these meaningful potentials? J Chem Phys. 1997;107:3698-706.

140. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci U S A. 2002;99(22):14116-21.

141. Morozov AV, Kortemme T, Tsemekhman K, Baker D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. Proc Natl Acad Sci U S A. 2004;101(18):6946-51.

142. Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. Nat Struct Mol Biol. 2003;10(12):980-.

143. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. Science. 2000;289(5481):920-30.

144. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol. 1994;238(5):777-93.

145. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 1997;6(8):1661-81.

146. Auffinger P, Westhof E. Rules governing the orientation of the 2'-hydroxyl group in RNA. J Mol Biol. 1997;274(1):54-63.

147. Jeffrey J. Gray,Stewart E.Moughon Tanja Kortemme Ora Schueler-Furman Kira M.S.Misura Alexandre V.Morozov David Baker. Protein-protein docking predictions for the CAPRI experiment. Proteins: Structure, Function, and Genetics. 2003;52(1):118-22.

148. Scherly D, Boelens W, Dathan NA, van Venrooij WJ, Mattaj IW. Major determinants of the specificity of interaction between small nuclear ribonucleoproteins U1A and U2B" and their cognate RNAs. Nature. 1990;345(6275):502-6.

149. Timm-H.Jessen, Oubridge C, Teo CH, Pritchard C, Nagai K. Identification of molecular contacts between the U1 A small nuclear ribonucleoprotein and U1 RNA. The EMBO Journal. 1991;10:3447-56.

150. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181(96):223-30.

151. Lee LP, Tidor B. Barstar is electrostatically optimized for tight binding to barnase. Nat Struct Biol. 2001;8(1):73-6.

152. Morozov A, Kortemme T, Baker D. Evaluation of models of electrostatic interactions in proteins. J Phys Chem Ser B. 2003;107:2075-90.

153. Perez-Canadillas J, Varani G. Recent advances in RNA-protein recognition. Current Opinion in Structural Biology. 2001;11(1):53-8.

154. Hall KB. RNA-protein interactions. Curr Opin Struct Biol. 2002;12(3):283-8.

155. Muller C, Wolberger C. Protein-nucleic acid interactions. Curr Opin Struct Biol. 2002;12:69-71.

156. Varani G, Nagai K. RNA recognition by RNP proteins during RNA processing. Annu Rev Biophys Biomol Struct. 1998;27:407-45.

157. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D. Computational redesign of protein-protein interaction specificity. Nat Struct Mol Biol. 2004;11(4):371-9.

158. Havranek JJ, Harbury PB. Automated design of specificity in molecular recognition. Nat Struct Biol. 2003;10(1):45-52.

159. Shifman JM, Mayo SL. Exploring the origins of binding specificity through the computational redesign of calmodulin. Proc Natl Acad Sci U S A. 2003;100(23):13274-9.

160. Pabo CO, Peisach E, Grant RA. Design and selection of novel Cys2His2 zinc finger proteins. Annu Rev Biochem. 2001;70:313-40.

161. Friesen WJ, Darby MK. Specific RNA binding by a single C2H2 zinc finger. J Biol Chem. 2001;276(3):1968-73.

162. Berman H, Olson W, Beveridge D, Westbrook J, Gelbin A, Demeny T, et al. The nucleic acid database: A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J. 1993;63:751-9.

163. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402,

164. Lu X, Olson W. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 2003;31(17):5108-21.

165. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learning. 1997;29(2):131-63.

166. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol. 1996;6(2):195-209.

167. Subramaniam S, Tcheng DK, Fenton JM. A knowledge-based method for protein structure refinement and prediction. Proc Int Conf Intell Syst Mol Biol. 1996;4:218-29.

168. Lu H, Skolnick J. Application of statistical potentials to protein structure refinement from low resolution ab initio models. Biopolymers. 2003;70(4):575-84.

169. Draper DE. PROTEIN-RNA RECOGNITION. Annu. Rev. Biochem. 1995;64:593-620.

170. Guzman RND, Turner RB, Summers MF. Protein–RNA recognition. Biopolymers (Nucleic Acid Sciences). 1998;48:181-95.

171. Amosova O, Broitman SL, Fresco JR. Alanine-scanning mutagenesis of the predicted rRNA-binding domain of ErmC? redefines the substrate-binding site and suggests a model for protein–RNA interactions. Nucl. Acids Res. 2003;31(16):4941-9.

172. Law MJ, Rice AJ, Lin P, Laird-Offringa IA. The role of RNA structure in the interaction of U1A protein with U1 hairpin II RNA. RNA. 2006;12(7):1168-78.

173. Xia T, Wan C, Roberts RW, Zewail AH. RNA-protein recognition: Single-residue ultrafast dynamical control of structural specificity and function. PNAS. 2005;102(37):13013-8.

174. White SA, Hoeger M, Schweppe JJ, Shillingford A, V S, Zarutskie J. Internal loop mutations in the ribosomal protein L30 binding site of the yeast L30 RNA transcript. RNA. 2004;10(3):369-77.

175. Auweter SD, Oberstrass FC, Allain FHT. Sequence-specific binding of single-stranded RNA: Is there a code for recognition? Nucl. Acids Res. 2006:gkl620.

176. Chen Y, Varani G. Protein families and RNA recognition. FEBS J. 2005;272(9):2088-97.

177. Messias AC, Sattler M. Structural basis of single-stranded RNA recognition. Acc. Chem. Res. 2004;37(5):279-87.

178. Stefl R, Skrisovska L, Allain FH-. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO reports. 2005;6(1):33-8.

179. Frankel AD. Fitting peptides into the RNA world. Current Opinion in Structural Biology. 2000;10(3):332-40.

180. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins: Structure, Function, and Genetics. 2000;38(1):3-16.

181. Notredame C. Http://ca.expasy.org/cgi-bin/reduce_redundancy.cgi.

182. Gray JJ. High-resolution protein-protein docking. Current Opinion in Structural Biology. 2006;16(2):183-93.

183. Case DA, Darden TA, III TEC, Simmerling CL, Wang J, Duke RE, et al. AMBER8. University of California, San Francisco. 2004.

184. Pastor R, Brooks B, Szabo A. An analysis of the accuracy of langevin and molecular dynamics algorithms. Molecular Physics. 1988;65(6):1409-19.

185. Izaguirre JA, Catarello DP, Wozniak JM, Skeel RD. Langevin stabilization of molecular dynamics. J. Chem. Phys. 2001;114:2090-8.

186. Wang X, Tanaka Hall TM. Structural basis for recognition of AU-rich element RNA by the HuD protein. Nat Struct Mol Biol. 2001;8(2):141-5.

187. Maris C, Dominguez C, Allain FHT. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. 2005;272(9):2118-31.

188. Siomi H, Matunis MJ, Michael WM, Dreyfuss G. The pre-mRNA binding K protein contains a novel evolutionary conserved motif. Nucl. Acids Res. 1993;21(5):1193-8.

189. Lewis HA, Musunuru K, Jensen KB, Edo C, Chen H, Darnell RB, et al. Sequence-specific RNA binding by a nova KH domain: Implications for paraneoplastic disease and the fragile X syndrome. Cell. 2000;100(3):323-32.

190. Grishin NV. KH domain: One motif, two folds. Nucl. Acids Res. 2001;29(3):638-43.

191. Beuth B, Pennell S, Arnvig KB, Martin SR, Taylor IA. Structure of a mycobacterium tuberculosis NusA–RNA complex. The EMBO Journal. 2005;24:3576-87.

192. Insight II 2000, Accelrys, Inc.

193. Molecular Operationg Environment, Chemical Computing Group, Inc.

194. Chang K, Ramos A. The double-stranded RNA-binding motif, a versatile macromolecular docking platform. FEBS J. 2005;272(9):2109-17.

195. Tsai DE, Harper DS, Keene JD. U1-snRNP-A protein selects a ten nucleotide consensus sequence from a degenerate RNA pool presented in various structural contexts. Nucl. Acids Res. 1991;19(18):4931-6.

196. Lunde BM, Moore C, Varani G. RNA-binding proteins: Modular design for efficient function. Nat Rev Mol Cell Biol. 2007;8(6):479-90.

197. Valegard K, Murray JB, Stonehouse NJ, van den Worm S, Stockley PG, Liljas L. The three-dimensional structures of two complexes between recombinant MS2 capsids

and RNA operator fragments reveal sequence-specific protein-RNA interactions. Journal of Molecular Biology. 1997;270(5):724-38.

198. Johansson HE, Dertinger D, LeCuyer KA, Behlen LS, Greef CH, Uhlenbeck OC. A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein. PNAS. 1998;95(16):9244-9.

199. Oubridge C, Ito N, Evans PR, Teo CH, Nagai K. Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature. 1994;372(6505):432-8.

200. Allain FHT, Gubser CC, Howe PWA, Nagai K, Neuhaus D, Varani G. Specificity of ribonucleoprotein interaction determined by RNA folding during complex formation. Nature. 1996;380(6575):646-50.

201. Gubser CC, Varani G. Structure of the polyadenylation regulatory element of the human U1A pre-mRNA 3'-untranslated region and interaction with the U1A protein. Biochemistry. 1996;35(7):2253-67.

202. Auweter SD, Fasan R, Reymond L, Underwood JG, Black DL, Pitsch S, et al. Molecular basis of RNA recognition by the human alternative splicing factor fox-1. The EMBO Journal. 2006;25:163-73.

203. Batey RT, Sagar MB, Doudna JA. Structural and energetic analysis of RNA recognition by a universally conserved protein from the signal recognition particle. Journal of Molecular Biology. 2001;307(1):229-46.

204. Siomi H, Dreyfuss G. RNA-binding proteins as regulators of gene expression. Current Opinion in Genetics & Development. 1997;7(3):345-53.

205. Onesto C, Berra E, Grepin R, Pages G. Poly(A)-binding protein-interacting protein 2, a strong regulator of vascular endothelial growth factor mRNA. J. Biol. Chem. 2004;279(33):34217-26.

206. Kinnaird JH, Maitland K, Walker GA, Wheatley I, Thompson FJ, Devaney E. HRP-2, a heterogeneous nuclear ribonucleoprotein, is essential for embryogenesis and oogenesis in caenorhabditis elegans. Experimental Cell Research. 2004;298(2):418-30.

207. Morii T, Sato S, Hagihara M, Mori Y, Imoto K, Makino K. Structure-based design of a leucine zipper protein with new DNA contacting region. Biochemistry. 2002;41(7):2177-83.

208. Debasisa Mohanty,Brian N.Dominy Andrzej Kolinski Charles L.Brooks I.I.I.Jeffrey Skolnick. Correlation between knowledge-based and detailed atomic

potentials: Application to the unfolding of the GCN4 leucine zipper. Proteins: Structure, Function, and Genetics. 1999;35(4):447-52.

209. Wang K, Fain B, Levitt M, Samudrala R. Improved protein structure selection using decoy-dependent discriminatory functions. BMC Struct Biol. 2004;4:8.

# VITA

Timothy Allen Robertson was born in Clarksburg, West Virginia in 1976, and raised in Columbus, Ohio. In 1999, he graduated magna cum laude from the University of Denver with a B.Sc. in Biology and Computer Science. In 2007, he was awarded a Ph.D. in Biochemistry from the University of Washington.