Electrostatics and Packing in Biomolecules: Accounting for Conformational Change in Protein Folding and Binding

by

Justin Andrew Caravella

Bachelor of Science in Chemistry, Duke University, 1996

Submitted to the Department of Chemistry
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Chemistry

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2002

Author .....
Department of Chemistry
May 9, 2002

Certified by .....
Bruce Tidor
Associate Professor of Bioengineering and Computer Science
Thesis Supervisor

Accepted by .....
Robert W. Field
Chairman, Departmental Committee on Graduate Students

This thesis has been examined by a committee of the Department of Chemistry as follows:

Lawrence J. Stern .....
Thesis Committee Chair

Bruce Tidor .....
Thesis Supervisor

Robert Sauer .....

Electrostatics and Packing in Biomolecules: Accounting for Conformational Change in Protein Folding and Binding

by

Justin Andrew Caravella

Submitted to the Department of Chemistry
on May 9, 2002, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Chemistry

Abstract

The role of electrostatics and packing in protein folding and molecular association was assessed in different biomolecular systems. A continuum electrostatic model was applied to long-range electrostatic effects in the binding of human carbonic anhydrase II to a sulfonamide inhibitor. The effect of chemically modifying lysine ϵ-amino groups was computed, and the average calculated value showed good agreement with experimental results determined by capillary electrophoresis. In a second study, the continuum model was used to analyze all the electrostatic interactions in the Zif268 protein–DNA complex. The net electrostatic effect was unfavorable to binding, although many individual groups or group pairs had a favorable effect, and the residues most unfavorable to binding correspond to those thought to be important for specificity. Also, a measure of electrostatic complementarity was developed and applied to myoglobin—both to known sequences and to hypothetical chimeric myoglobin sequences. The complementarity measure rated the correct myoglobins higher than chimeric myoglobins when crystal structures were used, and performed better than other readily available measures of complementarity when myoglobin homology models were evaluated. In the second part of the thesis, methods for repacking proteins were presented and applied to Arc repressor. Sequence variants that are predicted to fold as heterodimers preferentially and variants that favor a switch-Arc structure over wild-type were found. In a final set of calculations, the search algorithms for repacking were combined with electrostatic effects predicted from an approximate continuum model. The structure of Zif268 zinc finger 1 complexed to DNA was predicted when limited docking and side chain flexibility were allowed. The predicted structure shows good agreement with the x-ray crystal structure. A second repacked structure provides insight into how sequence changes affect structure and hence binding specificity in the zinc finger protein.

Thesis Supervisor: Bruce Tidor
Title: Associate Professor of Bioengineering and Computer Science

Acknowledgments

I would like to thank my advisor, Bruce Tidor, for his help and guidance. His insight and experience have been valuable in all of my research.

I am also grateful for the help and advice of the people in Bruce Tidor's lab. They have made the lab a good environment in which to work throughout my graduate career. I would especially like to thank Zachary Hendsch, Karl Hanf, Amy Keating, Erik Kangas, David Green, and Michael Altman for discussions and technical advice.

I would like to thank Jeffrey Carbeck, David Duffy, and George Whitesides, who performed the experiments and aided in the writing of Chapter 2. I would also like to thank Mike Nohaile and Robert Sauer, whose experiments provided the impetus for the work described in Chapter 6.

Finally, I would like to thank my fiancée, Tania, and my family for their continuing support and encouragement.

Contents

Chapter 1

General Introduction

The process of molecular recognition is central to a variety of biological systems. A better understanding of molecular recognition would aid our ability to predict the structure and function of proteins and potentially enable the design of proteins with novel structure or function. Protein folding and binding properties arise from the physical chemical properties of both the solute and the solvent. In order to make predictions about these properties, we must address two basic issues. First, the energy of macromolecules must be evaluated in a folded protein or bound complex, and when the molecules are fully solvated. Second, most proteins have some degree of conformational flexibility, so there must be a way of sampling different configurations of the solute and solvent. These issues are interrelated—for instance, the choice of energy function often affects what method of conformational sampling is feasible. The methods that can be used are also limited in that they must be completed by a computer processor in a reasonable amount of time for the large molecules being considered.

One key component of the energy of macromolecules is electrostatic. Electrostatic interactions are significant both within the solute and between the solute and solvent. Protein folding or binding typically involves a tradeoff between solute–solvent interactions and solute–solute interactions. The balance between these interactions determines the extent to which folding or binding is favorable.

A continuum electrostatic model [61, 62] was used to compute electrostatic effects

in folding and binding of proteins. In this model, the solute is treated as a region of low dielectric ($\epsilon = 4$), with atom centers represented as point charges. The solvent is represented as a region of high dielectric constant ($\epsilon = 80$) with a Debye-Hückel treatment of salt. The electrostatic energy of the system is then determined by numerical solution of the Poisson–Boltzmann equation. This method helps address both of the issues raised above. It allows reasonably fast computation of the electrostatic energy of a given conformation of the solute. In addition, the continuum model greatly simplifies conformational searches of the system. The dielectric continuum represents electrostatic properties averaged over many configurations of the solvent and thus obviates the need for sampling all the solvent configurations.

The continuum model will be used to examine the role of electrostatics in stability and complexation in several biomolecular systems. In Chapter 2, the continuum model is applied to understanding electrostatic interactions at long-range in a complex of human carbonic anhydrase II with a small molecule inhibitor. Recent experiments [24, 43] have provided an experimental probe of the role of charged residues far from the active site in binding affinity. The experimental results give an average effect of the lysines in the protein on binding affinity. The calculated effects of these charged residues will be compared to the experimental average, and examined further to understand the effect that each individual lysine has on binding affinity.

In Chapter 3, the continuum electrostatic model is applied to study the electrostatic interactions in the Zif268 protein–DNA complex. The total binding free energy is divided into contributions of individual chemical groups in an effort to better understand the role of particular amino acids and bases in binding affinity and specificity. The effect of several amino acid and base mutations will be modeled and compared to experiment. The mutation experiments provide a test for how well the continuum electrostatic model can account for the role of different chemical groups in the protein and DNA. The continuum model is tested further in a different type of problem in Chapter 4. We make use of properties that a molecule is expected to have [69] when it is electrostatically optimized for binding to another molecule [79] to develop a measure of electrostatic complementarity. The complementarity is used to

predict the stabilities of different sequences of myoglobin based on their electrostatic properties. The results with sequences that are known to fold are compared to results from hypothetical chimeric myoglobins. Other possible measures of complementarity are then compared to the electrostatic measure.

The continuum electrostatic model addresses part of the problem we are interested in solving. The dielectric continuum implicitly accounts for arrangements of the solvent, but only accounts for arrangements of the solute in a limited way. We can allow rearrangement of a protein by making use of search algorithms that efficiently search a large number of discrete conformations. Algorithms such as dead-end elimination [32, 48] and A* [78] have been used in protein repacking [55, 70] and design [29, 128] calculations. Chapter 5 presents a review of these search algorithms and additions to the approach that will permit a search of a finer sampling of side chain conformations.

In the past, these search algorithms have been applied mainly toward stabilizing a particular desired structure. In Chapter 6, they are applied to two repacking problems in Arc repressor which involve a search for a sequence that will stabilize one structure over a different closely related structure. The search for this type of structural specificity will provide a more stringent test of the search methodology, since we must find a sequence that stabilizes a desired fold and destabilizes an undesired fold. The basic search approach must be modified to generate multiple sequences and conformations that are expected to be stable. These possibilities are then screened for the desired specificity.

Search algorithms such as DEE and A* have primarily been applied to problems involving the repacking of hydrophobic cores. When solvation or electrostatic effects were treated at all, a simple surface area term was used [84, 137]. Hence, the approach essentially is a way of optimizing the shape of a molecule while keeping its electrostatic properties constant. The continuum electrostatic model, by contrast, has provided a framework for a different type of optimization [69, 79]. In this optimization, the shape of a molecule is held constant, while the electrostatic properties are optimized. Ideally, we would like to have a way of treating biological molecules that accounts for

both shape and electrostatic effects.

An approach for repacking a molecule while accounting for both shape and electrostatic effects is presented in Chapter 7. The basic strategy is to use DEE/A* to identify a large set of possible conformations, then to use an approximate continuum electrostatic model [119] to evaluate each conformation with a better treatment of electrostatic effects. The approximate model is required because the full numerical solution of the Poisson–Boltzmann equation is too computationally expensive. This methodology is applied to repacking the Zif268 zinc finger–DNA complex. The protein–DNA complex has an interface with a substantial number of charged and polar chemical groups and is therefore expected to require appropriate modeling of electrostatic and shape effects in order to predict the correct structure of the complex.

Chapter 2

Long-Range Electrostatic Contributions to Protein–Ligand Binding Estimated Using Protein Charge Ladders, Affinity Capillary Electrophoresis, and Continuum Electrostatic Theory[a]

Abstract

Affinity capillary electrophoresis and charge ladders of proteins — populations of derivatives of proteins that differ in their net charge — together measure the contributions of long-range electrostatic interactions to the energetics of binding of ligands. Continuum electrostatic calculations allow for the detailed analysis of electrostatic interactions among protein, ligand, and solvent. These tools are combined to measure and analyze the role of long range electrostatic interactions

[a]appears in *J. Am. Chem. Soc.*: Justin A. Caravella, Jeffrey D. Carbeck, David C. Duffy, George M. Whitesides, and Bruce Tidor. *J. Am. Chem. Soc.* **121**:4340-7 (1999).

to the free energy of binding, ΔG, of substituted benzene sulfonamide inhibitors to derivatives of human carbonic anhydrase II that differ in net charge through chemical modification of Lys ϵ-amino groups. The two approaches are complimentary. The experimental results are essentially exact: they are averages over populations of proteins that have the same number of acetylated Lys ϵ-amino groups and net charge but differ in their pattern of acetylation. The calculations are approximate but afford a detailed analysis of interactions of individual members of a population of derivatives of the protein with an inhibitor. A Monte Carlo simulation of the experimental data using calculated contributions of individual Lys ϵ-amino groups to ΔG of inhibitors showed that a large number of different distributions of patterns of acetylation are consistent with the experimental results. The calculations predict significant differences in the contributions of some Lys ϵ-amino groups to ΔG and suggest that the resolution of current capillary electrophoresis techniques is not sufficient to measure these differences.

2.1 Introduction

Electrostatic interactions affect the affinity, specificity, and catalytic properties of biomolecules. Analysis of electrostatic effects is complicated by their long-range nature and by the observation that their net effect often results from opposing contributions, such as unfavorable desolvation penalties that are at least partially offset by favorable intermolecular interactions on binding. One particularly striking long-range electrostatic effect involves the Met repressor binding to its DNA operator. In this system, the binding affinity increases 1000-fold upon binding a positively charged cofactor; this increase has been attributed almost entirely to a long-range electrostatic attraction between the charged cofactor and the DNA [108, 133]. In another study, theoretical methods were used by Allewell and co-workers to suggest that linked networks of long-range electrostatic interactions could be important in the conformational changes of aspartate transcarbamylase [102].

Experiments have been devised to evaluate long-range electrostatic interactions

by measuring pK_a shifts of ionizable groups due to modification of other charged groups [83]. One complication in analyzing and applying such information is that a protein usually contains many titratable groups, and it can be difficult to predict how the mutation of one residue will affect the titration state of each of these groups. Mutations involving a charged group may affect electrostatic interactions directly or indirectly (e.g., by shifting the pK_a's of other charges). Theoretical methods have been used to predict pK_a shifts while taking such changes into account, but these methods have had only modest success [6, 9, 45, 150]. Interestingly, these models of pK_a shifts appear to produce better agreement with experiment when the interior dielectric constant of the protein is assumed to be 20, rather than the values of 2–4 generally used in electrostatic calculations [4]. Although it seems unlikely that such a high value is realistic to describe the protein core [129, 130], there is some suggestion that it might approximate the effect of conformational changes accompanying titration that are not included in static models [8].

Another approach that can be usefully applied to measuring long-range interactions is double mutant cycles [21, 63, 120, 122]. The principle is relatively straightforward and in common use. To measure the interaction between the side chains of residues A and B, the effect of mutating both simultaneously to a reference amino acid is subtracted from the sum of the two individual mutations. The difference gives an effective cooperativity between the sites that can be viewed as an interaction free energy, though the analysis is not straightforward because structural relaxation and dielectric effects may operate in non-additive ways among the mutants.

The combination of protein charge ladders and affinity capillary electrophoresis (ACE) is a useful tool for estimating the role of electrostatic interactions in the binding affinity of proteins for ligands [24, 43]. In the present study, human carbonic anhydrase II (HCA II) was modified to produce a charge ladder — a mixture of protein derivatives in which various numbers of Lys ϵ-amino groups are acetylated. The interactions between the components of this mixture of proteins and three differently charged benzene sulfonamide inhibitors of HCA II were measured experimentally using ACE. The use of differently modified proteins in combination

with measurements on differently modified inhibitors effectively provides a double-mutant cycle, which we use to measure interaction energies involving Lys residues in HCA II and their effects on binding benzene sulfonamide inhibitors. The advantage of this method over making individual mutant proteins is that it gives an interaction free energy that represents an average over many possible "mutants" (proteins with sets of acetylated lysines) without the enormous effort that would be required to make individual mutations of different combinations of lysines. The precision of this measurement of binding energy may be improved because it represents an average over many interactions. One potential disadvantage is that modified proteins with the same net charge are not separated from each other and their individual binding properties can not be measured.

The experiments are complemented by continuum electrostatic calculations, which allow us to predict individual contributions of each lysine to the binding affinity. The calculations are based on the crystal structure of HCA II bound to a benzene sulfonamide inhibitor [14]. We average the calculated contributions of individual lysines and compare them to the experimental results. The detailed experimental results depend on how the individual contributions are averaged, which is determined by the composition of the complex mixture present in the charge ladder. We demonstrate that the computed results are consistent with a variety of different acetylation patterns and also use the calculated results to simulate an ACE experiment. This analysis shows that the predicted energetics of acetylation are also consistent with the pattern of peaks seen in these experiments.

The structure of the paper is as follows. We first present our experimental approach, in which we describe the formation of charge ladders and the use of ACE to measure binding affinities of various ligands. We then present our theoretical approach, where we discuss the application of continuum electrostatic theory to this system. This is followed by our results and a discussion of how the experimental results may be compared to the calculated results.

2.2 Experimental Approach

The formation and analysis of charge ladders. The treatment of HCA II with acetic anhydride results in the formation of a set of derivatives that differ in their number and distribution of acetylated Lys ϵ-amino groups and, therefore, their net charge; the coefficient of friction is approximately unaffected by the modification. The change in charge upon acetylation of Lys ϵ-amino groups depends on the values of pH of the electrophoresis buffer (8.4 in these studies) and the pK_a of Lys ($\sim$10.2) and is therefore assumed to be -1.

This mixture of derivatives of proteins is analyzed by capillary electrophoresis (CE), which separates the components of the mixture based on their electrophoretic mobility (directly related to the net charge of the protein, and inversely related to the coefficient of friction of the protein with the solvent). CE groups the distribution of modified proteins according to their net charge, or equivalently, their number of acetylated Lys ϵ-amino groups, into individual peaks or "rungs" of a charge ladder: the rung with the least negative charge is the native HCA II, and is referred to as rung 0; rung 1 is a peak consisting of those proteins that have a single acetylated Lys ϵ-amino group, and is a mixture of up to 24 regio-isomeric derivatives of the protein. In general, rung n will be a mixture of up to $\binom{24}{n}$ derivatives, which may be as many as 2.7 million for the center rungs of the ladder. The measured mobilities for each rung of the ladder are averages over the regio-isomeric derivatives of HCA II that make up that rung. The set of rungs of a charge ladder appears in CE as a set of peaks whose spacing varies in a regular way with the number of acetylated Lys ϵ-amino groups (bottom plot in Figure 2.1).

The composition of the distribution of derivatives of the protein is determined by the amount of acetic anhydride added to the solution of protein and the details of the reaction conditions. If relatively few equivalents of acetic anhydride are added, then we observe only the first several rungs of the ladder; if more equivalents of acetic anhydride are added the later rungs appear. The final charge ladders used in our experiments, which contain measurable amounts of protein in each rung, are

mixtures of the products of several acetylations using from 10 to 50 equivalents of acetic anhydride.

Affinity Capillary Electrophoresis. We measure the binding affinity of the derivatives that make up the charge ladder of the protein for various ligands using affinity capillary electrophoresis (ACE). In this technique the mobility of the rungs of the charge ladder are measured at various concentrations of ligand added to the electrophoresis buffer: the change in mobility with concentration of ligand is assumed to be directly proportional to the fraction of protein that is bound with ligand. Scatchard analysis of the mobility as a function concentration of ligand yields a dissociation constant, K_d, and a standard free energy of binding, ΔG. Figure 2.1 shows stacked electropherograms of the charge ladder of HCA II in the presence of increasing concentrations of inhibitor 1 in the electrophoresis buffer and illustrates how the mobility of the rungs of the charge ladder are affected by the presence of the ligand. The ligand is positively charged and the effective charge on each derivative of the protein is changed by an amount that depends on the amount of ligand bound to the protein (i.e., the fraction of time spent in the bound state). At sufficiently high concentrations of ligand, the proteins are saturated and their apparent charge differs from that of the protein in the unbound state (i.e., in the absence of ligand) by an amount equal to the charge on the ligand. At all concentrations of ligand, the individual rungs of the charge ladder of HCA II are resolved and the amount of ligand bound to the proteins that make up each rung of the charge ladder can be measured in a single set of experiments.

In this study, three substituted benzene sulfonamides are used as ligands for HCA II. The structure of these inhibitors is shown in Scheme 2.1. The pendant group substituted in the *para* position determines the charge of these inhibitors in solution: inhibitor 1 is positively charged, inhibitor 2 is neutral, and inhibitor 3 is negatively charged. The partial atomic charges used to model the inhibitors bound to the active site of HCA II using continuum electrostatics calculations (described below) are also shown in Scheme 2.1.

2.3 Theoretical Approach

The structure of HCA II bound to a benzene sulfonamide inhibitor is shown in Figure 2.2 [14]. There are 23 lysines in the crystallographic model, each of which may be acetylated in the experiment. The 24th lysine (Lys-261) is disordered in the crystal structure and is omitted from the analysis. The N-terminus of the unmodified protein is pre-acetylated and is therefore not relevant here. Figure 2.2 also shows the Zn^{2+} ion in the binding site, which is coordinated by three histidine residues and the sulfonamide group of the inhibitor.

In the unbound state of HCA II, the Zn^{2+} ion is also coordinated by a water whose pK_a is 7 [82]. Thus at pH 8.4 (the pH at which the experiments were done), there is a hydroxide bound in the active site. The hydroxide ion is displaced upon binding of the ligand, which is believed to be deprotonated in the bound state. Therefore, there is an additional formal negative charge present in the binding site in both the unbound (hydroxide bound) and bound states. The net change in charge at the binding site is simply the charge on the pendant group of the inhibitor, as illustrated in Figure 2.3. Models for each of the three inhibitor complexes were constructed and used in the analysis of long-range electrostatic effects.

Figure 2.4 is a histogram of the distances of the lysines from the formal positive charge on inhibitor 1. Lys-170 of HCA II is only 10 Å from this nitrogen, while the other 22 lysines are 17–38 Å from it. The distribution of distances is approximately uniform. The calculations therefore give an indication of how charge–charge interactions modulated by dielectric effects are expected to vary with distance.

The experiment involves acetylation of the various lysines in HCA II. We wish to determine the difference in binding free energy, $\Delta\Delta G$, caused by acetylating a lysine. Figure 2.3 schematically shows two binding reactions. The first is between a ligand and HCA II with a fully charged lysine, and its binding free energy is ΔG_{b1}. The second is the same process with an acetylated, uncharged lysine, and its binding free energy is ΔG_{b2}. As seen in the thermodynamic cycle, the difference in binding free energy upon acetylation of a lysine ($\Delta G_{b2} - \Delta G_{b1}$) is related to the free energy of

charging a lysine in the unbound state (ΔG_{c1}) and that of charging a lysine in the bound state (ΔG_{c2}) as follows:

$$\Delta\Delta G = \Delta G_{b2} - \Delta G_{b1} = \Delta G_{c1} - \Delta G_{c2} \tag{2.1}$$

In order to calculate $\Delta\Delta G$ for each lysine, we determine the free energies of charging ΔG_{c1} and ΔG_{c2} using continuum electrostatic theory. The contribution of each lysine to the overall binding affinity should be entirely electrostatic, since all of the lysines are far (> 10 Å) from the binding site. Calculations on acetylated lysines show that the contribution of the polar N-acetyl group to $\Delta\Delta G$ are negligible (less than 5×10^{-3} kcal/mol; results not shown). The analysis is therefore simplified to understanding the effect of adding a charge to the lysine, illustrated in Figure 2.3.

The electrostatic free energy of ligand binding can be divided into three parts: 1) the solvent-screened direct coulombic interactions between charged and polar groups in the protein and charged and polar groups in the ligand/hydroxide, 2) the change in interactions of the protein and ligand with solvent, and 3) the *intra*molecular interactions of the protein and the ligand, whose magnitude is changed upon binding because of the change in screening by the solvent and ion environment. These three contributions are illustrated in Scheme 2.2, where they are labeled as direct, solvation, and indirect, respectively.

An example of this third type, which we call an indirect effect, is provided by DNA bending. Studies of protein transcription factors have shown that some bend DNA toward the bound protein, while others bend DNA away from the bound protein [71]. The former type may be induced by basic residues contacting the phosphates of the DNA, which mitigates the repulsion between phosphates on one side of the DNA, bending it toward the protein [105, 135, 136, 143]. The latter type of bending may be caused by the indirect effect. Binding of the protein to DNA results in desolvation of the phosphates, which causes the phosphates on the side where protein is bound to repel each other more strongly than on the side where the phosphate interactions are screened by solvent [2, 96]. Thus there is a precedent that the indirect effects predicted

by continuum dielectric theory are important in the energetics of biomolecules.

We are particularly concerned with how the three types of contributions listed above are affected by the acetylation of lysines. Since we assume that a lysine is the only group whose charge is significantly altered by acetylation, we need only consider how each lysine contributes to the interactions enumerated above. A more detailed analysis would compute changes in the titration behavior of all residues (especially His, Cys, and Tyr, of which there are 12, 1, and 8, respectively) due to lysine acetylation, though these effects are expected to be small due to solvent screening.

The first contribution (direct), $\Delta\Delta G_{dir}$, is the difference in the screened coulombic interactions of the lysine amino group with the ligand in the bound state and the hydroxide in the unbound state. The second is the change in Lys–solvent interactions upon binding, $\Delta\Delta G_{solv}$. This term is never favorable because ligand binding displaces solvent, which always interacts favorably with a charged group. $\Delta\Delta G_{indir}$, the third contribution to $\Delta\Delta G$, is the change in interactions between the Lys and other groups in the protein (e.g., the Zn^{2+} ion in the active site) that results from binding of the ligand. We now have

$$\Delta\Delta G = \Delta\Delta G_{dir} + \Delta\Delta G_{solv} + \Delta\Delta G_{indir} \tag{2.2}$$

Continuum electrostatic theory was used to calculate these components of $\Delta\Delta G$ for each of the 23 lysines' interactions with each of the three inhibitors.

It is conceivable that the $\Delta\Delta G$ of a particular lysine depends on whether the other lysines in the protein are acetylated. In other words, if the $\Delta\Delta G_{indir}$ term of lysine A has a significant contribution from lysine B, then the overall $\Delta\Delta G$ will depend on whether lysine B is acetylated. Calculations show that in the case of HCA II, there is only one lysine with Lys–Lys $\Delta\Delta G_{indir}$ contributions larger than 5×10^{-4} kcal/mol. This residue, Lys-170, has two contributions of energy less than 5×10^{-3} kcal/mol. Therefore, it is reasonable to ignore the Lys–Lys interactions and to assume that the $\Delta\Delta G$ for the acetylation of a Lys does not depend on which other Lys are acetylated,

which simplifies the analysis greatly. The data presented below were calculated with all lysines charged.

2.4 Results

Dependence of ΔG on the number of acetylations of the protein. ACE was used to measure ΔG of binding for each rung of the charge ladder to each of the three ligands. The plot of ΔG *versus* rung number (n) is approximately linear for the three inhibitors, as observed in previous, similar experiments using bovine carbonic anhydrase II [43]. The slopes of this plot give the average interaction energy between a lysine and the sulfonamide inhibitor. The slopes of the least-squares lines are shown in Figure 2.5 (solid lines). The slope for inhibitor 1 is -0.07 kcal/mol per acetylation, indicating that binding to a positively charged ligand is more favorable when the net negative charge on the protein increases due to lysine acetylation. Conversely, the slope for inhibitor 3 is $+0.05$ kcal/mol per acetylation, meaning that binding to a negatively charged ligand is less favorable when the net negative charge is increased. The slope for inhibitor 2 (the neutral inhibitor) is -0.01 kcal/mol per acetylation. One method of analyzing these results is through the standard double-mutant cycle argument. This is described here and can be compared with the more detailed electrostatic analysis presented below. We may use the slopes of the lines to determine an average interaction energy between the lysines and the charged pendant group on the inhibitor. The double-mutant cycle may be represented as follows. Let X represent the lysine, which is changed to an acetylated lysine, A. The positive (or negative) pendant group on the inhibitor, represented by Y, is changed to a neutral group, B. The interaction energy between X and Y is given by

$$\Delta\Delta G_{\mathrm{int,X-Y}} = (\Delta G_{\mathrm{X,Y}} - \Delta G_{\mathrm{A,Y}}) - (\Delta G_{\mathrm{X,B}} - \Delta G_{\mathrm{A,B}}) \tag{2.3}$$

In this experiment, the average value of $\Delta G_{\mathrm{X,Y}} - \Delta G_{\mathrm{A,Y}}$ (the change in binding energy of acetylating a lysine) is given by the slope ΔG *versus* n for the positive inhibitor.

The difference in the second terms $\Delta G_{X,B} - \Delta G_{A,B}$ is given by the slope of the ΔG *versus* n line for the neutral inhibitor. The difference of these slopes thus gives the average interaction energy between a lysine and a charged pendant group, which is -0.06 kcal/mol in the case of the positively charged inhibitor and $+0.06$ kcal/mol for the negatively charged inhibitor.

Calculated interactions of individual lysines. The contribution to $\Delta\Delta G$ for each lysine was calculated, as well as the $\Delta\Delta G_{solv}$, $\Delta\Delta G_{dir}$, and the $\Delta\Delta G_{indir}$ terms. These contributions to $\Delta\Delta G$ for each lysine depend only on its interactions with solvent, with the ligand at the active site, and with other groups in HCA II. Both the ligands and the groups whose dielectric environment change are a long distance from most of the lysines (at least 17 Å, see Figure 2.4.) As a result, the calculated values of $\Delta\Delta G$ are relatively small — on the order of hundredths of a kcal/mol. However, the long distances over which the interactions occur also means that the calculations are more precise. For example, conformational flexibility of lysine side chains should not substantially alter such long-range interactions. Interactions with groups that are near the lysines generally do not change upon ligand binding, and therefore they also should make no contribution to $\Delta\Delta G$.

We calculate $\Delta\Delta G_{solv}$ to be less than 4×10^{-4} kcal/mol for all lysines, except Lys-170, for which it is 0.02 kcal/mol in the presence of inhibitors 1 and 2. The small values of $\Delta\Delta G_{solv}$ are expected because the region of the protein that is desolvated on ligand binding (i.e., the binding site) is far from all of the lysines other than Lys-170.

The average values of the total calculated $\Delta\Delta G$ are -0.07, 0.00, and $+0.07$, for inhibitors 1, 2, and 3, respectively. The dashed lines in Figure 2.5 are plotted with these slopes. The three lines originate from the points for the zeroth rung of the charge ladder.

Figure 2.6 shows the distributions of calculated values for $\Delta\Delta G_{dir}$, $\Delta\Delta G_{indir}$, and the total $\Delta\Delta G$ for each of the three inhibitors. $\Delta\Delta G_{solv}$ is not represented in the histograms, but its contribution is included in the histograms for the total $\Delta\Delta G$. Lys-170 has an especially large value of $\Delta\Delta G$ in the presence of inhibitors 1 and 3. Larger values of $\Delta\Delta G$ are expected for Lys-170 because it is rather close to the formal

charge on these inhibitors. Its calculated contribution to $\Delta\Delta G$ is 0.68 kcal/mol for inhibitor 1 and -0.54 kcal/mol for inhibitor 3.

We note from the histograms that $\Delta\Delta G_{indir}$ makes a significant contribution to the total $\Delta\Delta G$ for all three inhibitors. The main contribution to this term for all 23 lysines comes from the Zn^{2+} ion and the coordinated histidines. In the unbound state the Zn^{2+} ion is only coordinated by a hydroxide ion, but in the bound state it is coordinated by a sulfonamide, whose larger low-dielectric region reduces solvent screening of Zn^{2+} interactions. It is an interesting question whether solvent in the roughly conical-shaped binding pocket is sufficiently oriented or restricted in motion that a dielectric constant substantially different from the bulk solvent value is required to simulate its effects. The values of $\Delta\Delta G_{indir}$ are similar among all three inhibitors, indicating that they are somewhat independent of the conformation of the R group on the benzene sulfonamide. (All three sulfonamide R groups have significantly different minimized conformations, although trial calculations suggest that the results are not particularly sensitive to the choice of conformation.)

In the case of the neutral inhibitor (inhibitor 2), the average value of the total $\Delta\Delta G$ is essentially zero, as one might expect for a neutral group interacting with a charged group. However, a closer look at the histograms reveals that $\Delta\Delta G$ is made up of two non-zero terms that nearly cancel. $\Delta\Delta G_{indir}$ averages to $+0.02$ kcal/mol because the desolvation of the Zn^{2+} ion results in greater Zn^{2+}–Lys repulsion in the bound state than in the unbound state, but $\Delta\Delta G_{dir}$ averages to -0.02 kcal/mol. The Lys–sulfonamide interaction is more favorable than the Lys–hydroxide interaction, even though both hydroxide and sulfonamide have a total charge of -1, primarily because the sulfonamide's negative charge is less screened by solvent.

2.5 Discussion

Predicting a charge ladder experiment from calculated $\Delta\Delta G$ values. In order to determine more thoroughly how well the calculated results agree with the experiment, we must examine how a predicted set of $\Delta\Delta G$ values would be expected

to give rise to the charge ladders seen in the experiments. Since we predict that the acetylation of each lysine has a different value of $\Delta\Delta G$, the $\Delta\Delta G$ predicted for any rung depends upon the relative amounts of each of these derivatives within a rung. Since we have no information about the reactivity of each lysine, we have little basis for knowing the distribution of derivatives in each rung.

We wish to find out whether our computed $\Delta\Delta G$ values are consistent with the experimental results using any distribution of protein derivatives. When we use different distributions of acetylated lysines, we find that there are many distributions that result in an agreement of ΔG_{expt} and ΔG_{calc} for a single inhibitor (results not shown). More importantly, one can ask whether there is one pattern of acetylation in the rungs of the charge ladder that results in agreement between ΔG_{expt} and ΔG_{calc} for all three inhibitors.

A full treatment of the possible patterns of acetylation in each rung allows a wide range of distributions. The first rung is comprised of those proteins that have been acetylated once. Each of the lysines may react at rates that may differ by several orders of magnitude, giving vastly different distributions of derivatives. The situation is further complicated by the fact that each of the mono derivatives may undergo acetylation at any of the remaining positions, giving a large number of possible di-acetylated products, which then may go on to form the millions of other possible derivatives that could comprise the charge ladder. An analysis of such a system would be enormously complicated, so we made the following simplifying assumption to represent the vastly different possible patterns of acetylation.

Let us define p_{ij} as the probability that lysine j is acetylated in going from the $(i-1)$th to the ith rung of the charge ladder. The set of p_{ij} values was optimized using a Monte Carlo algorithm so that the best possible agreement between ΔG_{expt} and ΔG_{calc} was obtained. The average deviation between ΔG_{expt} and ΔG_{calc} for only the first seven rungs of the charge ladders was used, since these ΔG values are expected to be the most accurate. Later rungs of the charge ladder have more uncertainty associated with them due to broadening of the peaks (particularly inhibitor 3). The results of the optimization are shown in Figure 2.7. We see that the overall trends in

ΔG *versus* number of acetylations are reproduced, although the fluctuations in ΔG from one rung of the charge ladder to the next do not match exactly.

The results shown in Figure 2.7 can be reproduced using different sets of p_{ij}. In order to illustrate this fact, we have plotted the fraction of each lysine that is acetylated in each rung of the charge ladder. In each part of Figure 2.8, the bottom line represents the fraction of each lysine that is acetylated in the second rung of the ladder. The second line in each part of Figure 2.8 represents the fraction of each lysine acetylated in the fourth rung, etc. (Every other rung is omitted from the Figure for clarity.) We assume that this is the pattern of acetylation in the presence of all three inhibitors, since acetylation occurs prior to addition of any inhibitor. Although the patterns of acetylation are significantly different among the six parts of the Figure, they all give identical predicted values of ΔG for each rung. There are thus multiple patterns of acetylation for which the calculated values of ΔG agree with the experimental values.

We may compare these results to those of the "null model," in which we assume that all lysines are acetylated at the same rate, and thus the observed $\Delta\Delta$G would simply be an average of the calculated $\Delta\Delta$G values for the lysines. These results are plotted in Figure 2.5. Here the slopes of the dashed lines are simply the average $\Delta\Delta$G values for each of the three inhibitors, and the lines are drawn so that their intercepts are the ΔG values for rung 0 in each ladder. The difference in the experimental and calculated slopes is 0.00 kcal/mol for inhibitor 1, 0.01 kcal/mol for inhibitor 2, and 0.02 kcal/mol for inhibitor 3. These errors are on the order of experimental error, and therefore by this measure the agreement between theory and experiment is good, as seen in Figure 2.5.

Insensitivity to outlying values of $\Delta\Delta G_{calc}$. The most striking feature of the distributions of $\Delta\Delta G_{calc}$ values is that, in the cases of inhibitors 1 and 3, all the lysines have $\Delta\Delta$G's in the range 0.01–0.15 kcal/mol in magnitude, with the exception of one outlier (Lys-170), whose $\Delta\Delta$G is much larger in magnitude. Here we ask whether, if such a large $\Delta\Delta$G is present in the experiment, those derivatives with Lys-170 acetylated would be separated from those without Lys-170 acetylated by ACE. We

addressed this question by simulating a charge ladder experiment. Each protein derivative is assumed to make a contribution to the peak on the charge ladder as a gaussian that is slightly narrower than the narrowest experimental linewidth. The derivatives are assumed to be present in the distribution predicted by the Monte Carlo method described above. The ligand concentration was chosen so that it would maximize separation of the derivatives on the calculated charge ladder. The results are shown in Figure 2.9. Those derivatives acetylated at position 170 do not appear to separate from those which are not acetylated at 170. There is still only one peak for each rung of the charge ladder, regardless of which distribution of protein derivatives is used. We conclude that the charge ladder experiments provide an accurate measure of the average value for $\Delta\Delta$G of acetylation over a large number of derivatives. However, they can not rule out the existence of one or two outliers from the measured average $\Delta\Delta$G in the large pool of derivatives unless higher resolution experiments are employed. Using our calculated values of $\Delta\Delta$G, splitting of peaks is expected when the experimental resolution is increased by a factor of roughly five.

Direct Interaction Strength as a Function of Distance. We determined the energy of electrostatic interactions between the charges on lysines and (1) the positive charge on inhibitor 1, (2) the negative charge on inhibitor 3, and (3) the negatively charged hydroxide. In order to quantify the extent to which the interactions are screened, we define the effective dielectric constant for an interacting pair of charges as the dielectric constant of a homogeneous medium that would give an identical interaction energy. This definition gives $\epsilon_{\text{eff}} = q_1 q_2 / Er$ where ϵ_{eff} is the effective dielectric constant, q_1 and q_2 are the charges on the interacting groups, E is the screened coulombic energy of interaction, and r is the distance between the groups. In Figure 2.10 we plot the effective dielectric constant for these direct interactions as a function of distance between charged groups. The trend is clearly that the effective dielectric constant increases with distance. Other similar calculations have found that long-range interactions at distances of 10–15 Å usually have an effective dielectric constant of approximately 40 [102]. These results agree with this value, however the computed effective dielectric constant increases rapidly with distance at

distances greater than 20 Å. Note that all of the interactions plotted in Figure 2.10 are between solvent-exposed charged groups. We might expect the effective dielectric constant to be substantially lower in cases where charges are buried.

Effects of parameters on the calculation. It is useful to know how the results of the calculation depend on the various parameters used in order to estimate errors. When the ionic strength was decreased to 15 mM (from 25 mM, which was used in the experiments), the calculated total $\Delta\Delta G$ values increased almost uniformly by 13%. An increase in interaction energies is expected due to the decrease in screening by salt. When the ion exclusion radius was increased to 4.0 Å (from 2.0 Å), the magnitudes of the total $\Delta\Delta G$ increased almost uniformly by 8%. Changing the interior dielectric constant from 4 to 3 resulted in an average change in $\Delta\Delta G$ of 3×10^{-3} kcal/mol, and changing the interior dielectric to 6 resulted in an average change in $\Delta\Delta G$ of 4×10^{-3}. The maximum change in the $\Delta\Delta G$ of any lysine caused by these changes in dielectric constant was less than 0.01 kcal/mol. Calculations were also performed with charges and radii from the CHARMM19 parameter set [18]. These calculations resulted in an average deviation from those made with the PARSE parameters of 7×10^{-3} kcal/mol. The results of the calculation are therefore not strongly dependent upon the choice of parameters.

2.6 Conclusion

The surface lysines of HCA II contribute to its binding of the family of benzene sulfonamide inhibitors studied here, primarily by long-range electrostatic interactions with partially buried active site groups. These interactions are of two types: direct interactions of lysine with ligand (sulfonamide in the bound state *versus* hydroxide in the unbound state) and indirect interactions of lysine with polar and charged protein groups in the active site that are differentially screened in the bound and unbound states. While most analyses of ligand binding focus on the former, both interactions are of similar magnitude here.

Experiments using affinity capillary electrophoresis provide aggregate information

that is complementary to the detailed, though approximate, analysis with continuum electrostatic theory. A visual examination of the experimental electropherograms suggests a straightforward, uniform binding free energy contribution from each lysine of the protein because the rungs do not split apart at intermediate ligand concentration (partial saturation). The nearly constant increment in binding free energy per protein charge adds further support. One insight from the calculations is that such seemingly uniform behavior may be masking more complex underlying energetics that are homogenized through population averaging and instrument resolution. The use of higher-resolution capillary experiments will be helpful in examining the underlying distributions in more detail.

2.7 Experimental and Theoretical Details

Materials. Carbonic anhydrase II (human; pI 7.6; E.C. 4.2.1.1) was purchased from Sigma (St. Louis, MO). The protein, as purified from human erythrocytes, is acetylated at the N-terminal α-amino group. Inhibitor 1 was synthesized as described in ref. [49]; inhibitors 2 and 3 were synthesized as described in ref. [5]. Uncoated fused silica capillaries with an internal diameter of 50 μm were purchased from Polymicro Technologies (Phoenix, AZ).

Acetylation of Amino Groups of Proteins. Proteins were dissolved in water at a concentration of 0.1 mM, and 10 vol % of 0.1 N NaOH was added to each solution to bring the pH to 12. Five to 20 equiv of acetic anhydride (100 mM in dioxane) were added to the protein solution and the reactants were quickly mixed by vortexing. Reactions were usually complete within 1 minute. Each rung in the final charge ladders was controlled to have similar intensity by mixing the products from several acetylations done with different equivalents of acetic anhydride. The sample was diluted in electrophoresis buffer (25 mM Tris, 192 mM Gly, pH 8.4) prior to analysis.

Capillary Electrophoresis (CE). CE experiments were conducted on a Beckman P/ACE 5500. Charge ladders of HCA II were analyzed at 25° C on an

uncoated capillary of fused silica (total length 47 cm, with a length to the detector of 40 cm) using 25 mM Tris-192 mM Gly buffer (pH 8.4) and an applied voltage of 15 kV.

Affinity Capillary Electrophoresis. The binding affinities of the proteins that make up the charge ladder of HCA II for the benzene sulfonamide inhibitors were measured using affinity capillary electrophoresis (ACE). In this technique the value of the electrophoretic mobility of a receptor, R, is measured as a function of increasing concentration of a ligand, L, in the electrophoresis buffer. The effective value of the electrophoretic mobility of the receptor, μ_{eff}, is the concentration-weighted average of the mobilities of the free, μ_{R}, and bound, $\mu_{\text{R}\bullet\text{L}}$ forms of the receptor. This effective mobility is expressed in eq 2.4, where θ is the mole fraction of the receptor–ligand complex.

$$\mu_{\text{eff}} = \theta\mu_{\text{R}\bullet\text{L}} + (1-\theta)\mu_{\text{R}} \tag{2.4}$$

The value of θ is estimated from

$$\theta = \frac{\mu_{\text{eff}} - \mu_{\text{R}}}{\mu_{\text{R}\bullet\text{L}} - \mu_{\text{R}}} = \frac{\Delta\mu_{\text{eff}}}{\Delta\mu_{\text{max}}} \tag{2.5}$$

where $\Delta\mu_{\text{eff}}$ is the difference between μ_{eff} and μ_{R}, and $\Delta\mu_{\text{max}}$ is the difference between μ_{R} and $\mu_{\text{R}\bullet\text{L}}$ (i.e., the value of μ_{eff} when the receptor is saturated with ligand). Scatchard analysis of the values of θ as a function of the concentration of ligand, L, yields the values of the binding constant, K_b

$$\frac{\theta}{[\text{L}]} = \text{K}_b(1-\theta) \tag{2.6}$$

The values of ΔG are then given by

$$\Delta\text{G} = -\text{RT}\ln\text{K}_b \tag{2.7}$$

In estimating the binding affinity of the proteins that make up a charge ladder for a ligand, we measure the changes in the mobility of each of the rungs of the

charge ladder as a function of the concentration of ligand in the electrophoresis buffer; that is, we measure simultaneously the binding affinities of this collection of protein derivatives for a ligand under a single set of experimental conditions.

The binding affinity of a receptor for a ligand can be measured by ACE only if the electrophoretic mobility of the receptor changes upon the association of the ligand. For inhibitors 1 and 3, the association of these ligands results in a measurable change in the mobility of HCA II; the receptor–ligand complex differs by approximately one unit of charge from the receptor in its free form. The change in mobility of HCA II due to the association of the neutral inhibitor 2 is negligible. To measure the binding affinity of a receptor for a neutral ligand, we used a competitive binding assay where we measured the change in mobility of the complex of the receptor and charged ligand, $L_{\pm}$, as a function of increasing concentrations of the neutral ligand, L_0 [49]. Applying eqs 2.4–2.6 to the observed changes in mobility, we obtain an apparent binding constant of the neutral ligand, $K_b^{0,\mathrm{app}}$. The true binding constant of the neutral ligand, K_b^0, is then obtained from $K_b^{0,\mathrm{app}}$ and the known value of the binding constant, $K_b^{\pm}$, and concentration of $L_{\pm}$:

$$K_b^0 = K_b^{0,\mathrm{app}}(1 + K_b^{\pm}[L_{\pm}]) \tag{2.8}$$

Structure of HCA II and complex. We used a crystal structure of HCA II complexed to a benzene sulfonamide inhibitor (PDB identifier 1cnw) [14]. Polar hydrogens were added to the structure using the HBUILD facility in the program CHARMM [18, 19]. The protein was kept at the same conformation in the unbound state in order to cancel grid energy terms from the solution of the linearized Poisson–Boltzmann equation. The RMS deviation in C_α positions between the ligand-bound structure which was used and an unbound structure of HCA II [39] is 0.17 Å, confirming that there is very little structural change in HCA II upon binding a benzene sulfonamide. The sulfonamide structure was modified so that its R groups were those of ligands 1, 2, and 3. Those parts of the ligand structures that were not present the x-ray crystal structure were minimized using the CHARMM22 all-atom parameter

set [85]. The protein and sulfonamide group were fixed at their crystal structure coordinates during the minimization.

The water bound in the active site of HCA II has a pK_a of 7 [82], and would therefore be hydroxide ion at the pH of the charge ladder experiments (8.4). The unbound state was therefore modeled with a hydroxide ion bound. The geometry of the hydroxide ion as bound to the zinc was taken from the geometry optimizations described below.

Partial Atomic Charges. The Poisson–Boltzmann calculations require partial atomic charges for all atoms in the unbound and bound state. Charges for HCA II (except the three histidine side chains coordinating Zn^{2+}, as described below) were taken from the PARSE parameter set [131]. The functional groups in the inhibitors that have parameters in the PARSE parameter set (e.g., amines, amides, and carboxylates) were assigned charges accordingly. However, many of the charges in the inhibitors (i.e., those of the benzene sulfonamide groups) were not given by the PARSE parameters and were therefore assigned by ab initio methods. The molecules were each divided into fragments separated by methylene groups. The electrostatic potential of the fragments that do not have PARSE charges was computed, and point charges were fit to the atom centers using the RESP method [7]. Geometry optimizations and electrostatic potential calculations were carried out with the program Gaussian 94 [40] as described below.

First, the geometry of the benzene sulfonamide fragment of inhibitor 1 was determined at the 6–31G** level. The charged nitrogen and the carbon atoms bonded to it were omitted from the optimization, since their charges can be taken from the PARSE parameter set, and their orbitals would not be expected to significantly alter the charges in the rest of the molecule.

The geometry of zinc complexed with three imidazole rings (to model His 94, 96, and 119) and with a hydrogen-substituted sulfonamide was optimized at the 3–21G level with the Wachters-Hay all-electron basis set for zinc as implemented in the program Gaussian 94 [40]. This smaller basis set was required in order to allow the calculation to finish in a reasonable time. Full optimization of this system resulted in

a significant conformational change of the imidazole rings with respect to the crystal structure, so the complex was optimized while keeping the dihedral angles fixed to their x-ray crystal structure values. The difference in partial charges assigned with dihedrals free and fixed was less than 0.05 for all atoms except for zinc and the histidine nitrogens liganded directly to it. These atoms were more highly charged (each nitrogen by ≈ -0.1 and zinc by $+0.3$) when the dihedrals were fixed.

The full benzene sulfonamide structure was appended to the zinc complex so that it had the proper orientation with respect to the zinc and the histidines. The electrostatic potential of the full complex was calculated at the 6–311G** level. Point charges were fit to the atomic centers using the RESP program, and the charges were constrained so that the total charge on the Zn^{2+} plus histidines was +2. The complex of hydroxide with the zinc and histidines was also assembled and optimized at the 3–21G level, again while keeping the dihedral angles along the Zn^{2+}–N coordination fixed. The charges on the Zn^{2+} and histidines were similar in both complexes (within 0.05 charge units), and therefore their charges were averaged to give the charges used in the electrostatic calculations. The final charges used on the sulfonamides, the Zn^{2+}, and the histidines are shown in Scheme 2.1.

Continuum Electrostatic Calculations. Calculations of the electrostatic free energy were carried out using a locally modified version of the program DELPHI [46, 47, 126]. The linearized Poisson–Boltzmann equation was solved iteratively at 2.0 grids/Å and at an ionic strength of 0.025 M. Each calculation of the potential in the bound and unbound states was repeated with ten different translations of the grid with respect to the molecule (i.e., the molecular geometry was unchanged). The standard deviation of the $\Delta\Delta G$ values over these translations was never more than 5×10^{-4} kcal/mol, except in the case of Lys-170, where it was not more than 3×10^{-3} kcal/mol for any of the inhibitors. Therefore, the numerical error associated with the finite difference method is small compared to the values being computed.

Modified Analysis of Binding Free Energies. Since the calculations suggest that each rung may contain a distribution of receptors with a distribution of binding affinities, a conventional Scatchard analysis may not describe the binding curve

correctly. Let us assume that the observed value of the fraction of bound receptors, θ_{obs}, represents an average over the receptors,

$$\theta_{\text{obs}} = \left\langle \frac{[L]}{K_{d,i} + [L]} \right\rangle \tag{2.9}$$

where each receptor i may have a different dissociation constant $\text{K}_{d,i}$. Given some distribution of receptors with an average ΔG of $\overline{\Delta\text{G}}$ and a standard deviation about the mean $\overline{\Delta\text{G}}$ of σ, we may express the observed theta as:

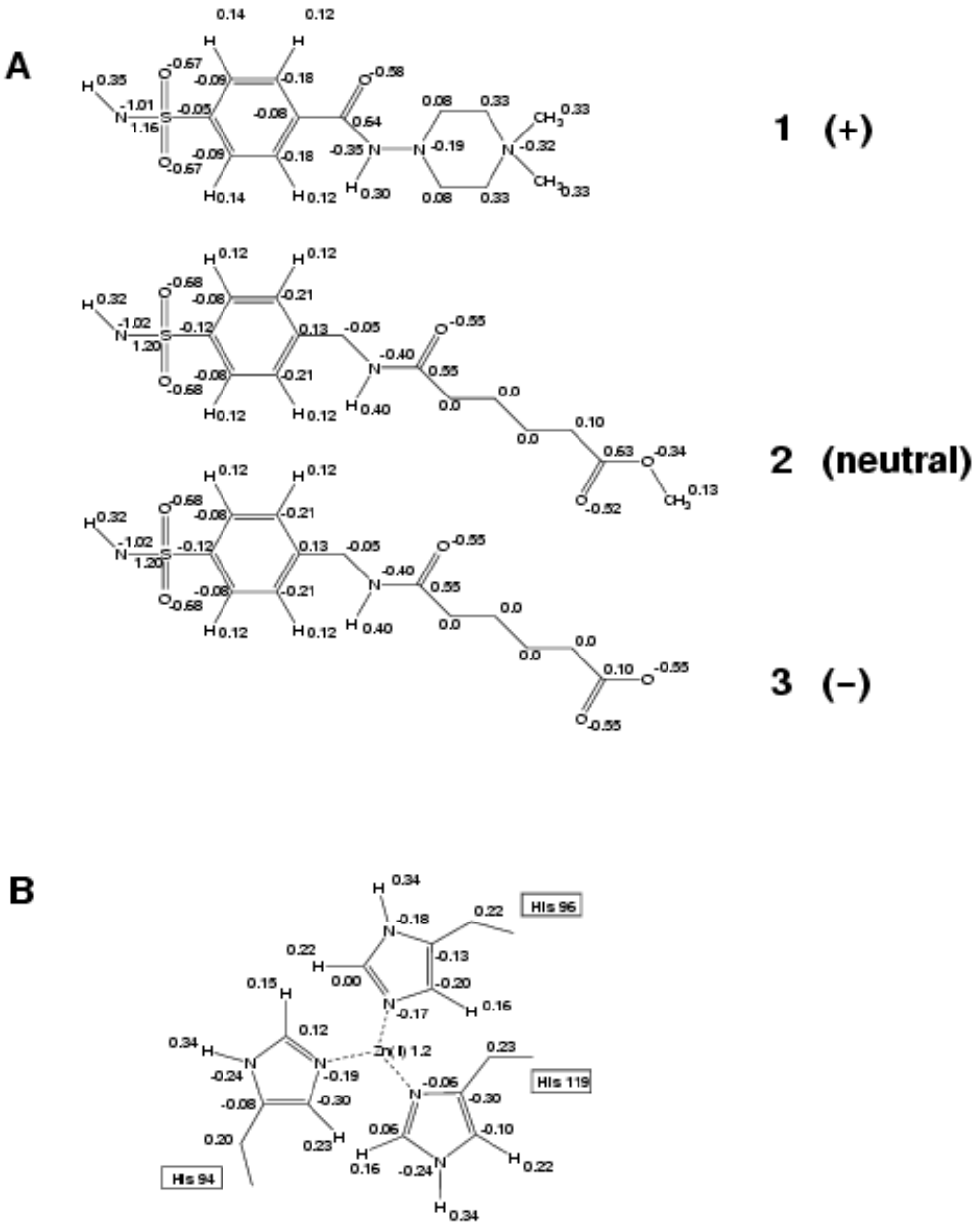
$$\theta_{\text{obs}} = \hat{\theta} + \left(\frac{\sigma}{RT}\right)^2 \left(\frac{\hat{K}}{[L]}\right)\left(\frac{\hat{K}}{[L]} - 1\right)\hat{\theta}^3 \tag{2.10}$$

Here, $\hat{K}$ is the dissociation constant corresponding to a receptor with the average binding energy. The theta *versus* ligand concentration data for each rung of the charge ladder can be fit to this equation (using a non-linear least-squares fit [113]) to give $\overline{\Delta\text{G}}$ and σ.

The results are not shown here because when computed the statistical significance of fitting to an additional parameter, σ, was dubious. The analysis, however, showed that the standard deviation of rung 0 was zero, since this rung represents only one species. The standard deviation gradually increased with increasing rung number, suggestive of a broader range of binding energies.

Acknowledgements

We thank Barry Honig for making the DELPHI computer program available and Martin Karplus for making the CHARMM computer program available. Figure 2.2 was made with QUANTA from Molecular Simulations, Inc. (San Diego, CA). This work was supported by the National Institutes of Health (GM51559 and GM30367 to G.M.W.; GM55758 and GM56552 to B.T.). J.A.C. is a National Science Foundation Predoctoral Fellow.



Scheme 2.1: **A** The structures of the benzene sulfonamide inhibitors substituted in the para position with neutral and charged pendant groups. The partial atomic charges used are shown next to the atoms. **B** The partial atomic charges used on the Zn^{2+} and the histidine side chains in the active site of HCA II.

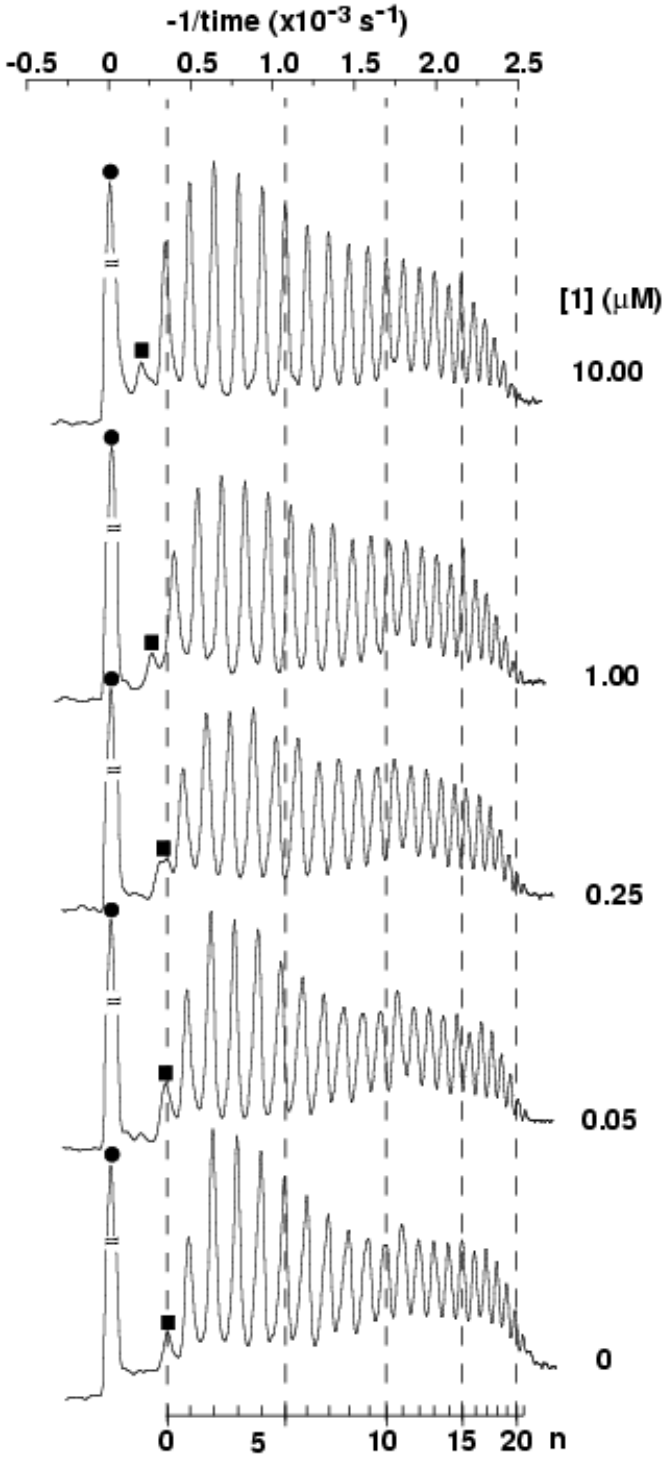


Figure 2.1: Electropherograms illustrating the change in the electrophoretic mobility of the rungs of the charge ladder of HCA II with increasing concentrations of inhibitor 1.

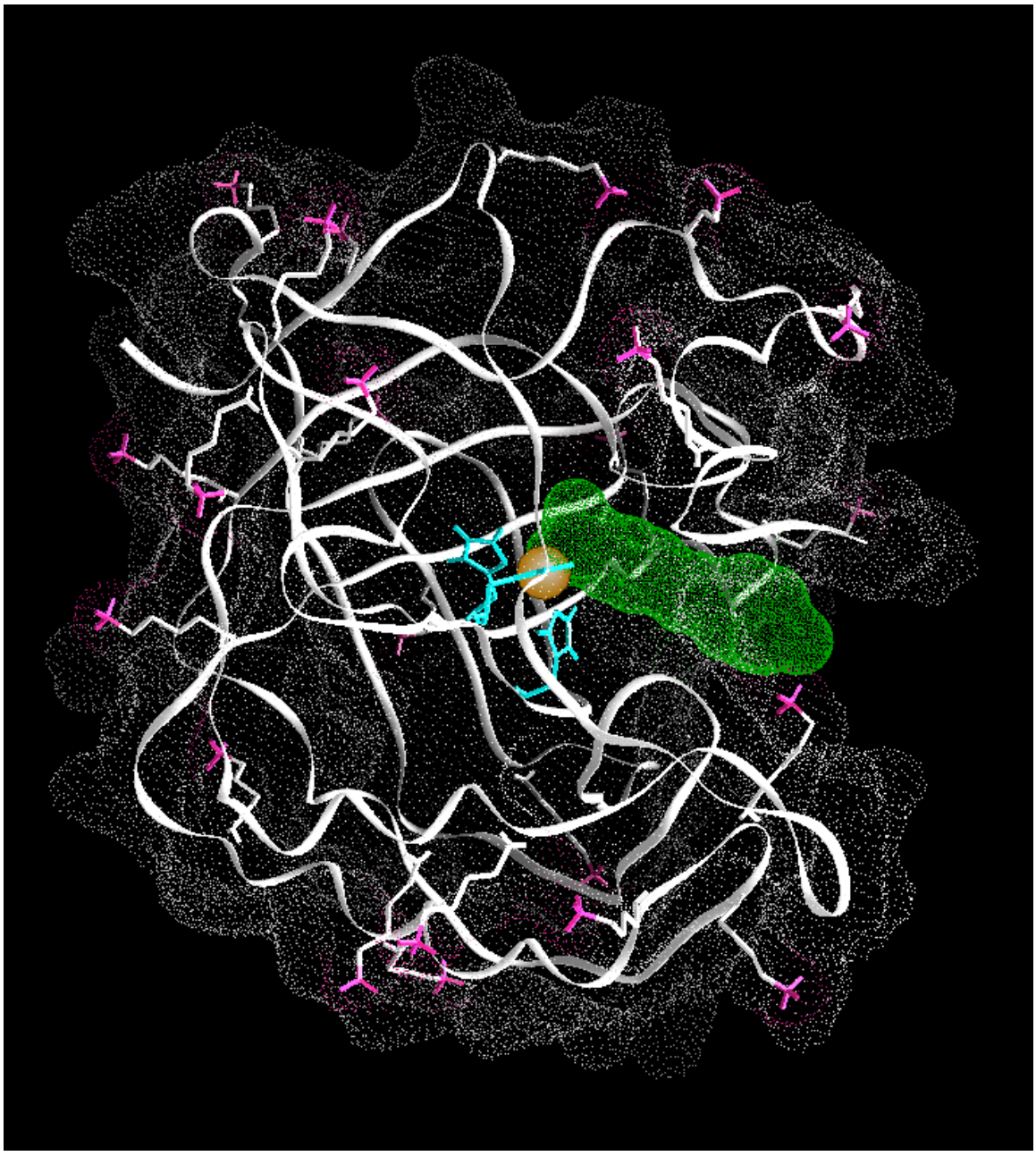
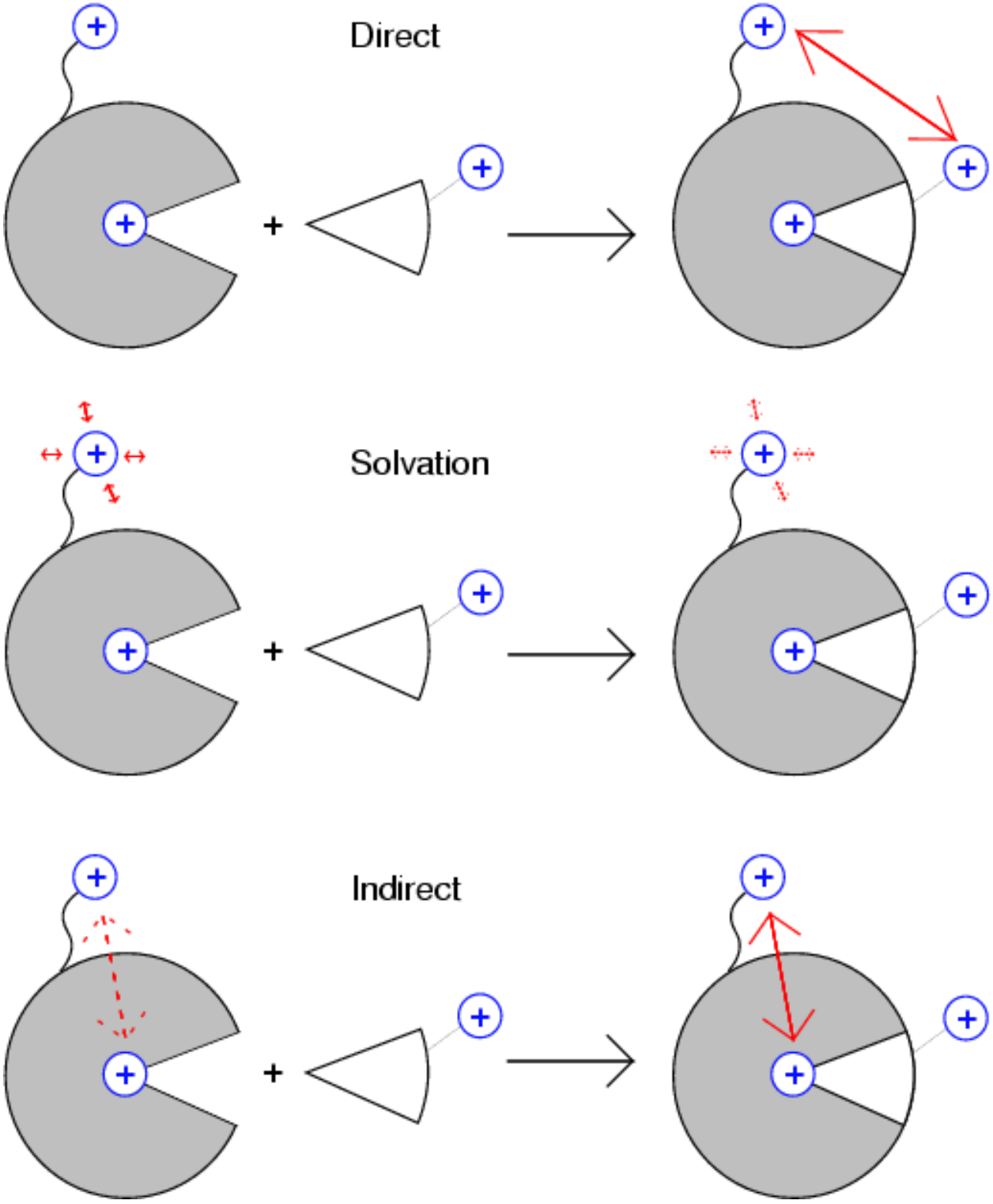


Figure 2.2: The structure of HCA II. The lysine side chain nitrogens and hydrogens are in pink. The active site zinc ion is shown as a gold sphere, and the side chains of the histidines ligated to it are shown in blue. The surface of the protein is depicted with white dots, and the surface of the bound inhibitor (inhibitor 1) is shown with green dots. All of the lysines are surface accessible, as indicated by the pink dots where the lysine amino groups contact the surface.



Scheme 2.2: An illustration of the three contributions to $\Delta\Delta G$ of binding.

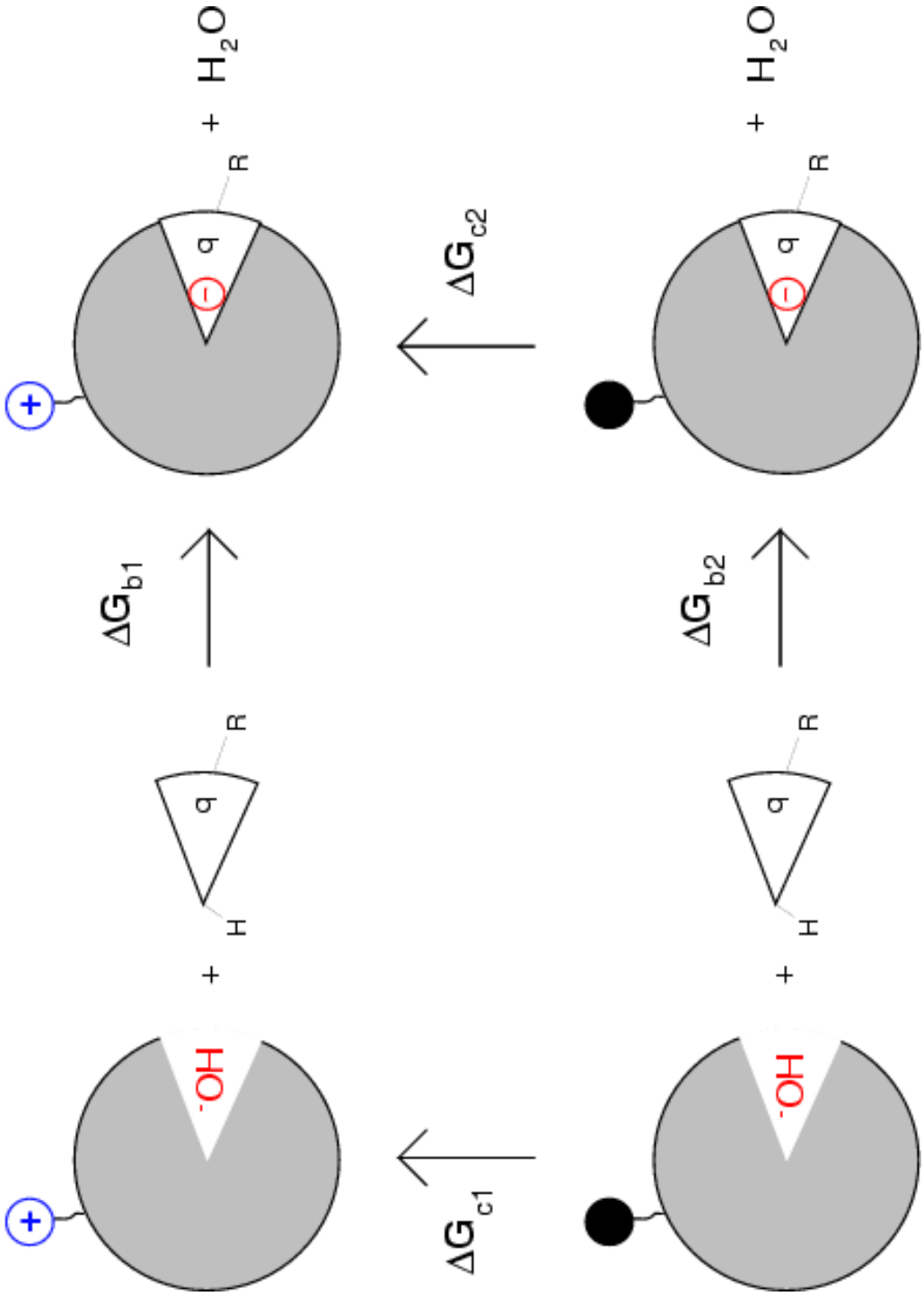


Figure 2.3: Thermodynamic cycle for the binding of a ligand to HCA II. HCA II is shown as a circle with a conical binding pocket. When it binds a sulfonamide ligand, hydroxide is displaced from the binding site. The lysine being acetylated is shown with a + charge in its native state and as neutral in its acetylated (uncharged) state. The difference in electrostatic binding free energy caused by acetylating the lysine can be determined by finding the energy of charging (deacetylating) the lysine in both the bound and unbound (hydroxide-bound) states.

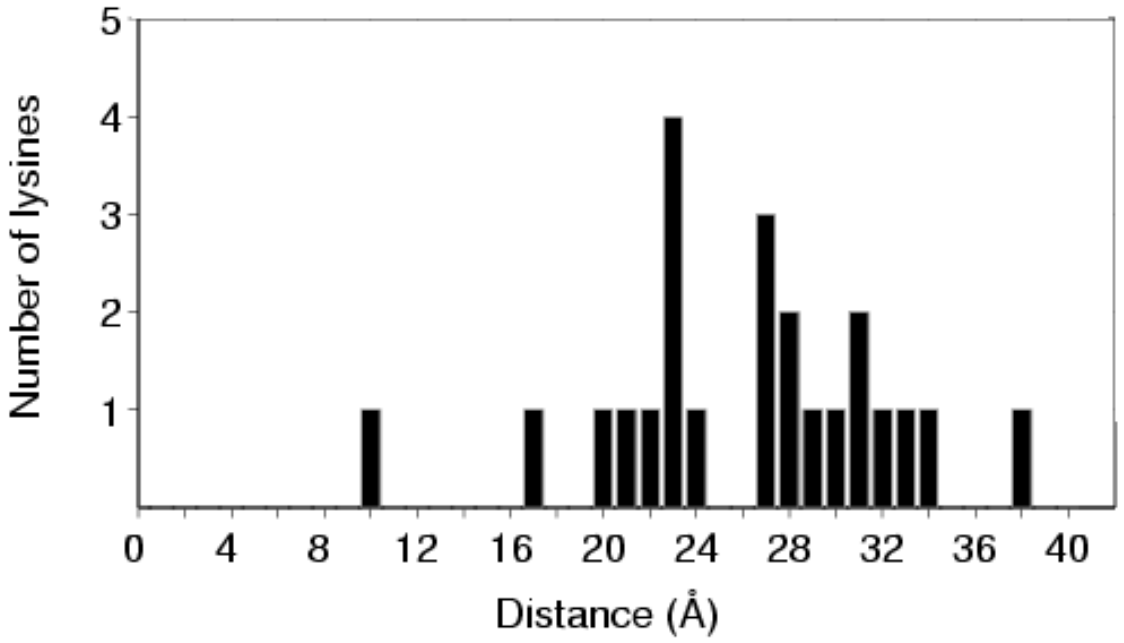


Figure 2.4: Histogram of distances from the Lys ϵ-amino groups in HCA II to the ammonium group of inhibitor 1 bound in the active site.

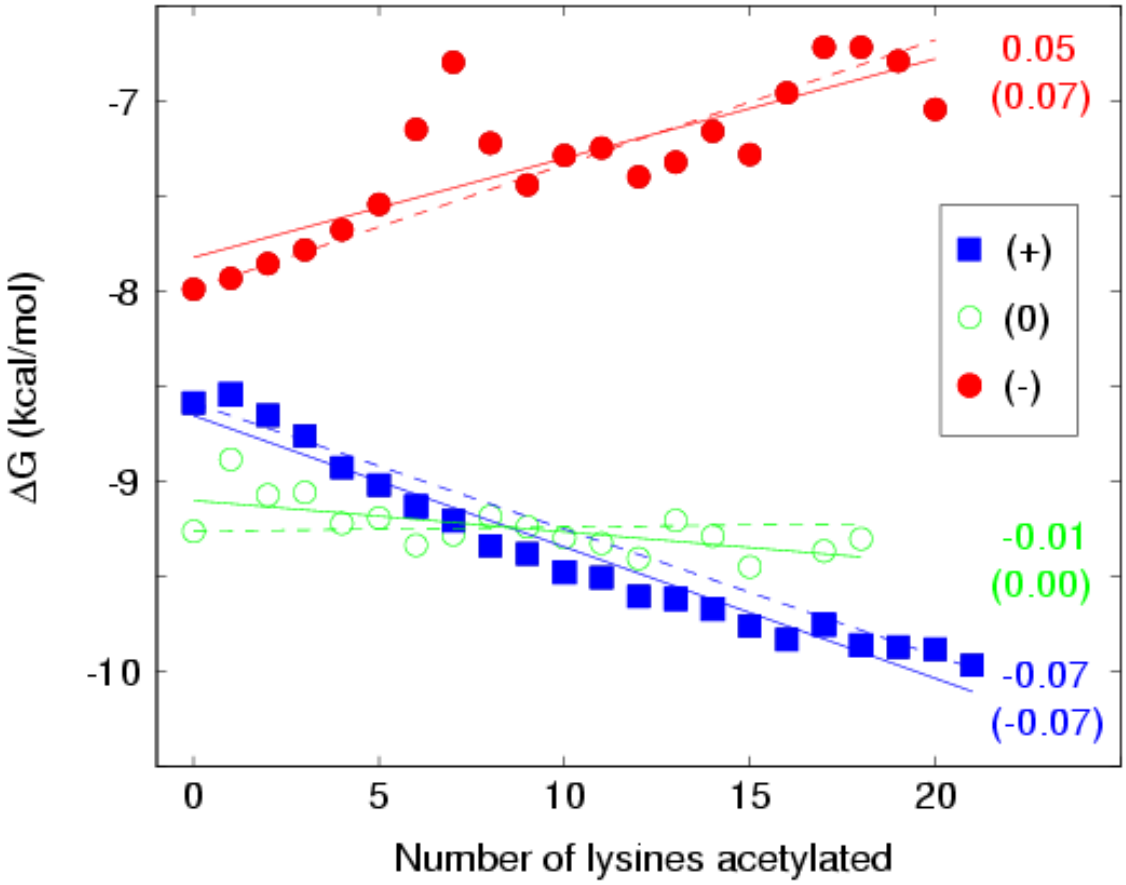


Figure 2.5: Plot of the free energy of binding (ΔG) determined by ACE *versus* the number of acetylated lysines (n) for each of the three inhibitors. The solid lines represent the least-squares fit to the data points, and the dashed lines represent the calculated values of the slopes as determined from the "null" model. The filled squares are the experimental data for inhibitor 1, the open circles are for inhibitor 2, and the filled circles are for inhibitor 3.

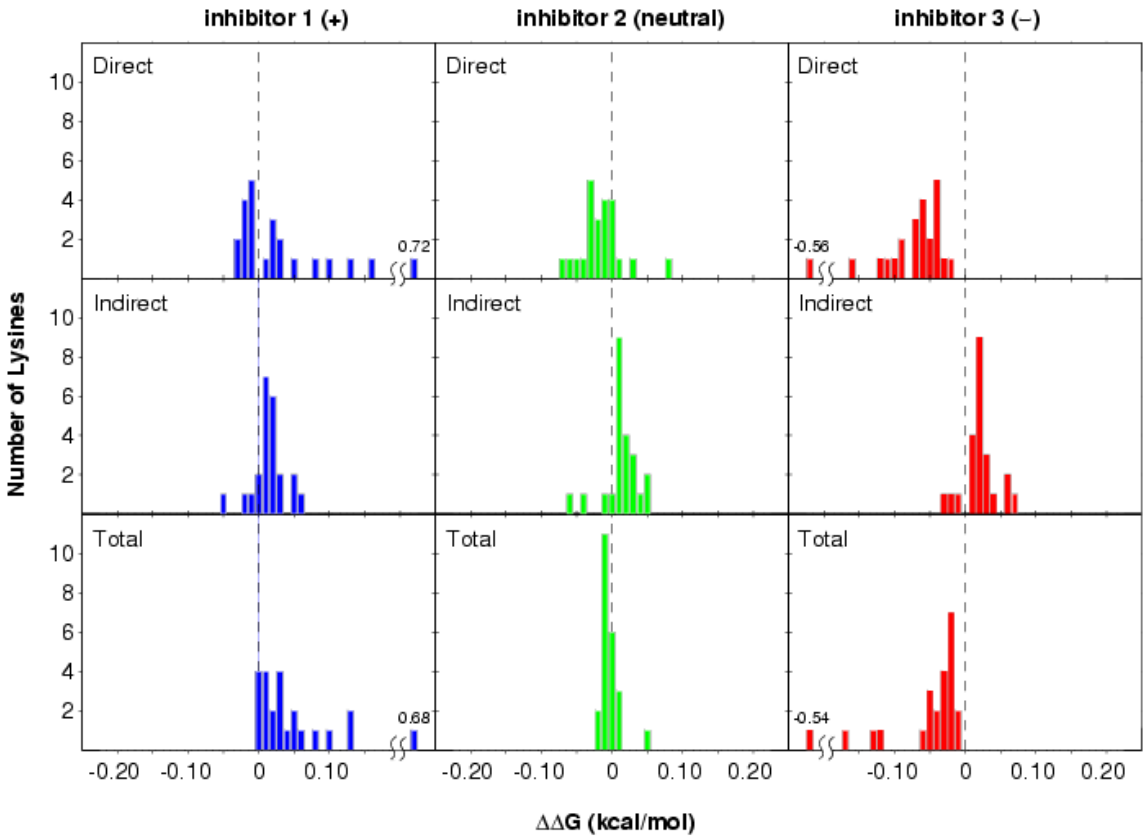


Figure 2.6: Histograms of $\Delta\Delta G_{dir}$, $\Delta\Delta G_{indir}$, and the total $\Delta\Delta G$ as calculated for each of the three inhibitors. The values of $\Delta\Delta G_{dir}$ and the total $\Delta\Delta G$ calculated for Lys-170 in the presence of inhibitors 1 and 3 are much larger in magnitude than the other $\Delta\Delta G$ values. These four values are printed on the histograms.

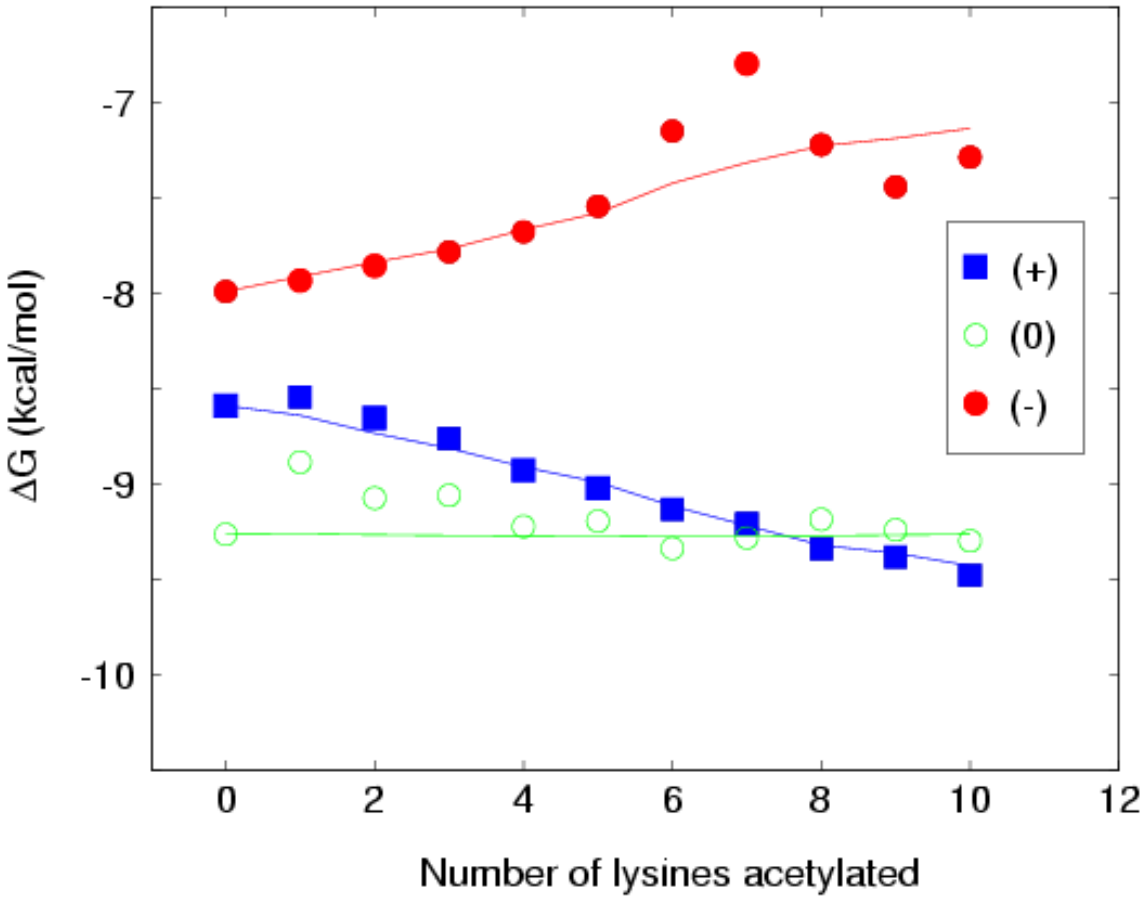


Figure 2.7: A simulated plot of ΔG *versus* n. The filled squares, open circles, and filled circles represent experimental data as in Figure 2.5. The lines represent a fit of the calculated $\Delta\Delta$G values to the experimental data. The rates of acetylation of each lysine were chosen so that the fit to the experimental data was optimized.

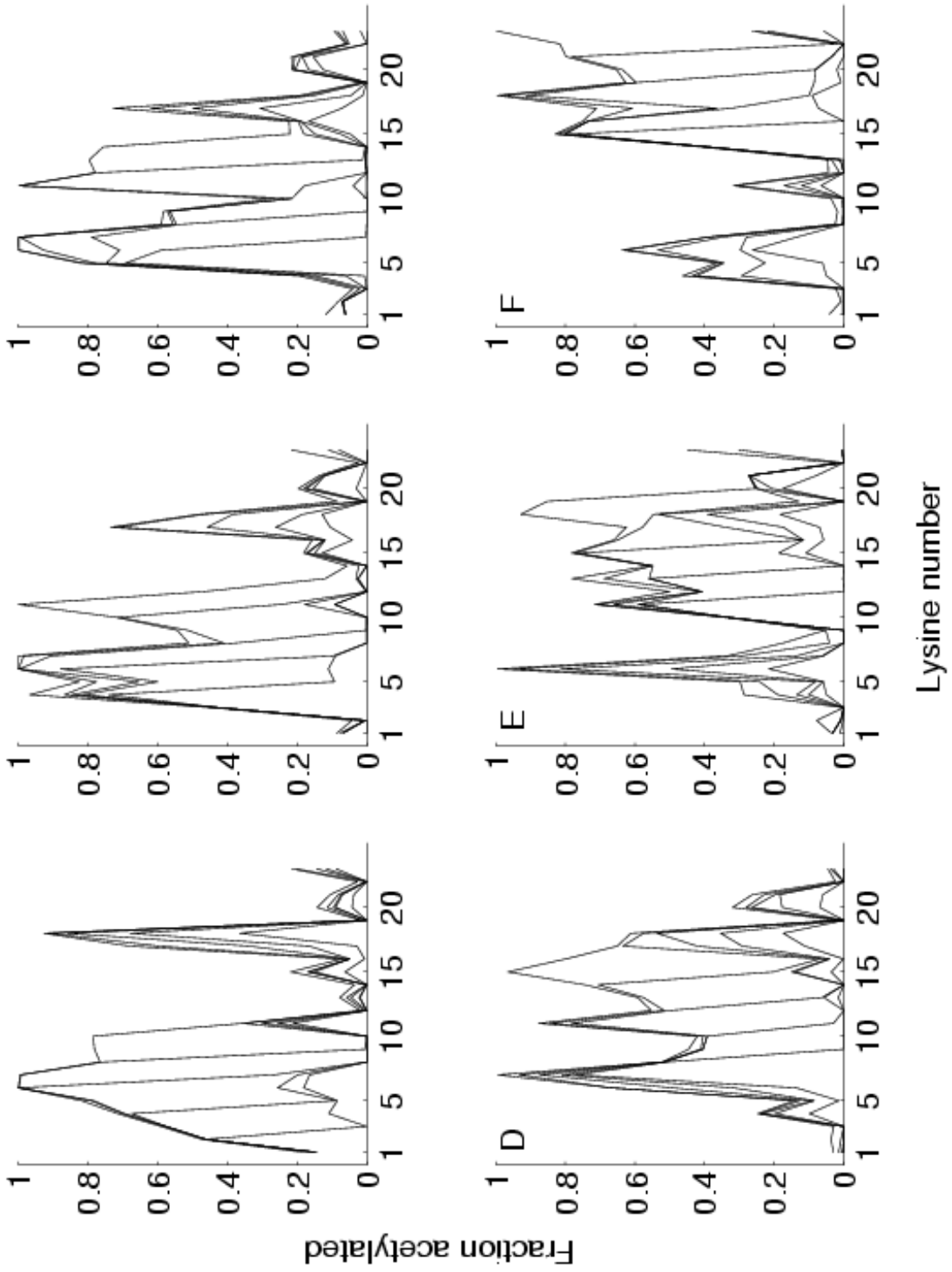


Figure 2.8: Six different patterns of acetylation that fit the data shown in Figure 2.7. In **A**–**F**, the bottom most line plots the fraction of each lysine that is acetylated in the second rung of the charge ladder. The second line from the bottom plots the fractions for the fourth rung, and so on. The patterns of acetylation are significantly different, yet they produce identical values of ΔG (see Figure 2.7).

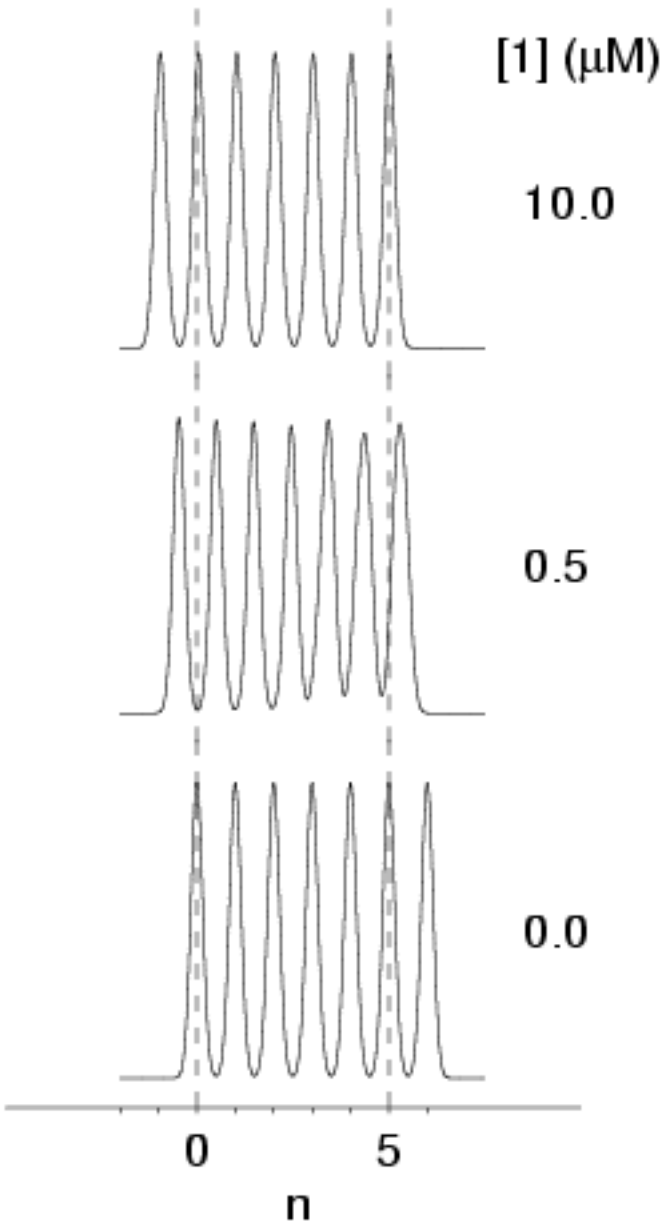


Figure 2.9: A simulated charge ladder experiment. The fraction of each possible protein derivative was chosen to be consistent with the ΔG *versus* n plot. The concentrations of inhibitor 1 are shown in the Figure. 0.5 μM was chosen in order to maximize the separation between protein derivatives. Note that each rung appears as a single peak, as in the experiments, despite the spread in values of $\Delta\Delta$G for each lysine. 10 μM is saturating, and the peaks are therefore shifted over one charge unit relative to their position with no ligand present.

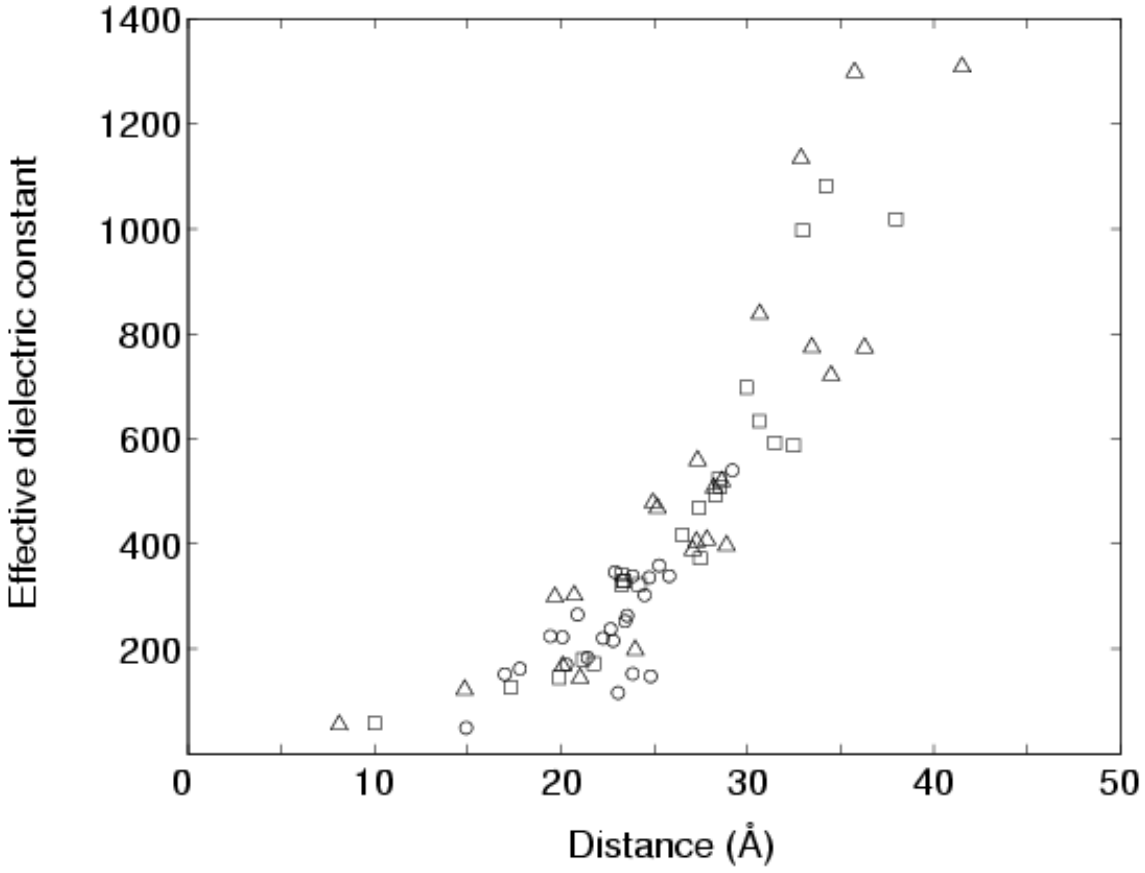


Figure 2.10: A plot of the effective dielectric constant for pairwise charge–charge interactions *versus* distance. The squares denote Lys–inhibitor 1 interactions, the triangles denote Lys–inhibitor 3 interactions, and the circles denote Lys-hydroxide interactions.

Chapter 3

Electrostatic Contributions to Zif268 Zinc Finger–DNA Binding

3.1 Introduction

Electrostatic interactions play an important role in determining the structure and function of biological molecules. The role of electrostatics is particularly important in the case of DNA-binding proteins, which generally interact with the negatively charged phosphate backbone of DNA in addition to its hydrogen-bonding bases. Although the coulombic interaction in a salt bridge or hydrogen bond is clearly favorable, the groups involved must be desolvated to make such interactions, and thus the net contribution of electrostatics to binding may be unfavorable. Previous calculations have shown that replacing some individual salt bridges with hydrophobic groups would lead to more stable proteins [59], although the results depend on the particular environment of the salt bridge. In a study using combinatorial mutagenesis to probe a salt bridge triad in Arc repressor, Waldburger et al. [141] found that simultaneous hydrophobic substitutions for all three residues stabilized the protein. Wimley et al. [144] also found that salt bridges were electrostatically destabilizing using a peptide model system. More recently, Albeck et al. [1] studied a larger set of electrostatic interactions in the TEM-1-β-lactamase–BLIP interface using double and higher-order mutant cycles. They found that salt bridges were either neutral

or unfavorable in isolation but favorable in a background of other complementary electrostatic residues, illustrating the importance of the protein environment in determining the effect of a salt bridge. As previously pointed out [60], charged side chains are less solvated in folded proteins than in unfolded proteins, and therefore one might expect a smaller desolvation penalty for binding than folding. This may account in part for the favorable salt bridges observed in the TEM-1-β-lactamase–BLIP complex [1]. In addition, protein–DNA binding interfaces are rich in charged and polar groups. A network of charged and polar groups may result in somewhat more favorable electrostatics since each group may be able to make more favorable interactions while not paying a substantially larger desolvation penalty.

Other theoretical studies have suggested that hydrogen bonds also tend to be unfavorable to folding. Using free energy simulation, Wang et al. [142] find that the electrostatic contribution of hydrogen bonds to α-helix formation is unfavorable. In addition, Yang and Honig [147–149] have used Poisson-Boltzmann electrostatics to predict that the electrostatic contributions to α-helix, β-sheet, and turn formation are all unfavorable.

In this work we have examined the role of electrostatics in a protein–DNA complex — the Zif268 zinc finger complexed to a consensus binding site. A continuum electrostatic model was used to account for interactions in the complex and in the unbound state. In this model, the solute atoms were treated as point charges in a low-dielectric medium whose boundary was defined as the molecular surface. The solvent was modeled as a high-dielectric continuum with a Debye-Hückel treatment of salt. The model has been applied to other protein–DNA complexes to study the effect of salt concentration [2, 93, 94, 151], the "steering" of ligands to receptors [2, 72, 124, 151], temperature effects [2] and pK_a shifts upon binding [94]. Here, the continuum model was used to study the electrostatic interactions involved in binding of Zif268 to DNA. Calculations on other complexes have shown that electrostatics tend to disfavor binding in protein complexes [60, 95, 100, 101, 123]. A study of this highly charged complex may provide additional insight into the role of electrostatics in binding. Moreover, the current study analyzes the changes in interactions between

pairs of chemical groups and between individual chemical groups and solvent that accompany binding.

The crystal structure of a Zif268-DNA complex was solved to 2.1 Å resolution by Pavletich and Pabo [107]. A second structure was solved to 1.6 Å resolution by Elrod-Erickson et al. [38]. The latter, which was used for this work, is shown in Figure 3.1. It consists of three zinc finger domains, each of which makes contacts to a three base-pair subsite on the DNA. The amino acid and DNA base numbering described here will follow the convention of Elrod-Erickson et al. The bases on the strand of DNA that makes the most protein contacts are numbered from 1 to 11, and the numbers 2' to 11' refer to the bases that are complementary to the unprimed bases. There are four protein residue positions in each finger that make direct base contacts, and these are labelled positions -1, 2, 3, and 6 (numbering relative to the start of the α-helix). In each zinc finger domain, amino acid position -1 contacts the base at the 3' end of the primary DNA strand. The amino acid at position 3 is closest to the central base, and amino acid position 6 is nearest the base at the 5' end of the subsite.

The role of each of these residues in the zinc finger's binding affinity has been studied experimentally. We will consider each group's role in terms of its solvation and direct electrostatic interactions across the interface. In addition, we consider "indirect" intramolecular electrostatic interactions, caused by reduced solvent screening in the bound state relative to the unbound state. In similar calculations on the GCN4 leucine zipper, substantial favorable electrostatic contributions were predicted from groups in the same molecule [60]. In a study of the electrostatic contributions to the binding of carbonic anhydrase II to several ligands, indirect effects were computed to be nearly equal in magnitude to direct electrostatic interactions [20]. Although individual electrostatic effects are difficult to resolve by experiment, the total predicted electrostatic effects in this system were consistent with experiment. Others have pointed out the potential importance of the indirect effect in binding DNA [35, 96, 140, 143]. Binding of protein to DNA causes decreased screening of the phosphates, and increased inter–phosphate repulsion. Any protein that binds DNA

must overcome this unfavorable contribution.

It is clear that a DNA binding protein should contain positively charged side chains to interact with the phosphates of the DNA. However, in the zinc finger (as well as a few other DNA binding proteins) the protein has a few negatively charged residues, even at the interface. When new zinc finger sequences were selected using phage-display methods, a preponderance of the new sequences also had negatively charged residues at the protein–DNA interface. Examples of this type of result include phage selections on an NRE binding site [53] as well as selections for variants that bind to the prototypical Zif268 DNA binding site [66, 67]. An analysis of the contributions of each group in the protein and DNA may give some insight into the reason for this preference.

3.2 Methods

Structure Preparation. The calcluations were carried out using the crystal structure of Zif268 bound to DNA from the Protein Data Bank [38]. (pdb code 1aay) Polar hydrogen atoms were positioned using the HBUILD facility of CHARMM [19]. All electrostatic calculations used the CHARMM PARAM19 parameters [18] for protein and an experimental set for DNA [146].

The electrostatic energy was analyzed by dividing the complex into chemical groups. The chemical groups all carried a charge of -1, 0, or $+1$, and they corresponded to intuitive chemical groupings. Each amino acid residue was divided into three groups: side chain, backbone carbonyl, and backbone amino (C^{α}–N–H). Each ribonucleotide residue was divided into base, ribose, and phosphate groups. The C1' atom (which has a total charge of 0.26) was included in both the base (with a charge of 0.06) and the ribose (charge of 0.20) so that both groups were neutral. The phosphate group had a charge of -1 and included all four oxygen atoms bonded to a phosphorous. By convention, we refer to the phosphate group on the 5' side of a base as the phosphate associated with that base. Thus, phosphate group guanine 7 is the phosphate on the 5' side of guanine 7. Each zinc ion and its coordinating cysteine

and histidine side chains were treated as one complete group with a total charge of 0.

Partial atomic charge assignment for zinc–Cys_2–His_2 complex. Charges were assigned to each of the three zincs in Zif268 using a fit to electrostatic potentials computed from *ab initio* methods. The geometry of zinc complexed with two methylimidazole molecules and two methanethiolate molecules (to model the His and Cys side chains, respectively) was optimized at the 3–21G level with the Wachters-Hay all-electron basis set for zinc as implemented in the program Gaussian 98 [41]. The dihedral angles were constrained to be the same as finger 1 in the crystal structure. The electrostatic potential was computed using the 6–31G* basis set, and the atomic charges were fit using the program RESP [7]. This procedure resulted in a charge of +1.160 on the zinc, a total charge of −0.704 on each Cys side chain, and a total charge of +0.124 on each His side chain.

Continuum Electrostatic Calculations. All electrostatic calculations were carried out using a locally modified version of the program DELPHI [46, 47, 126]. The interior dielectric constant was set to 4 and the exterior dielectric to 80. The linearized Poisson–Boltzmann equation was solved with an ionic strength of 0.145 M and a 2-Å Stern layer. Each calculation of the potential in the bound and unbound states was repeated with ten different translations of the grid with respect to the molecule (i.e., the molecular geometry was unchanged). Results for the total binding free energy from the linearized Poisson–Boltzmann equation agreed with those from the non-linear form to less than 10%. The linear form allows the results to be analyzed in terms of contributions from individual groups because of superposition. Contributions from individual groups were calculated on a 65x65x65 grid with four levels of focussing: 23%, 92%, and overfocussing at 184% and 368%. The final grid spacing was 4.9 grid units/Å. Short- to medium-range interactions between groups were computed from the overfocussed grid. Long-range interactions were computed using the finest grid upon which both groups fit. This scheme was validated by comparison to results from a 257x257x257 grid with no overfocussing, both for the total energy of the system and for solvation and interactions from selected groups. The agreement between the overfocussed results and the results from a 257x257x257 grid is excellent. The

difference in $\Delta\Delta G_{contrib}$ caused by overfocussing for each group tested was less than 0.01 kcal/mol, and the difference in the total electrostatic binding free energy was 0.03 kcal/mol. Hence overfocussing was used to increase computational efficiency without sacrificing accuracy.

For each group used in the electrostatic analysis, three energy terms were calculated for forming the protein–DNA complex. $\Delta\Delta G_{solv}$ is the difference in the group's interaction with solvent in the complex and the unbound state. The direct interaction, $\Delta\Delta G_{dir}$, is the group's solvent-screened interaction with groups across the interface. The indirect interaction, $\Delta\Delta G_{indir}$, is the change in solvent screening of a group's interaction with other groups in the same half of the complex. These individual terms were computed as in our previous work [60].

It is convenient to have a single number for each group to represent its overall contribution to binding. We define $\Delta\Delta G_{contrib}$ as the sum of $\Delta\Delta G_{solv}$ for a group plus one-half of its $\Delta\Delta G_{dir}$ and $\Delta\Delta G_{indir}$ with other groups. In this way, each interaction is divided equally between the two groups responsible, and the total of all contributions adds up to the total binding free energy. The mutation term, $\Delta\Delta G_{mut}$, is the sum $\Delta\Delta G_{solv} + \Delta\Delta G_{dir} + \Delta\Delta G_{indir}$ for a particular group. It corresponds to the change in binding free energy caused by mutating in the partial atomic charges for the group in question (i.e., for mutating from a hydrophobic isostere to the actual charge distribution).

Mutations. DNA mutations were made by simply deleting the base pair in the crystal structure and replacing it with the mutant base pair in standard geometry as defined by the parameter set. The four side chains of finger 2 that contact these base pairs were then allowed to relax. Subsequent partial relaxation of the local protein environment was achieved by a search of a discrete set of side chain rotamers using the dead-end elimination algorithm [32, 48]. The side chains chose the minimum energy structure from a library including their crystal structure coordinates. The rotamer library of Dunbrack and Karplus [34] was used and expanded with rotamers of $\pm 15^\circ$ about χ_1 and χ_2 and rotamers of $\pm 20^\circ$ about χ_3. The rotamers of histidine included the three different protonation states of the side chain. This resulted in a total of

2188 Arg rotamers, 82 Asp rotamers, 165 His protonation/rotamer states, and 82 Thr rotamers. In each of the control DNA mutations, where a DNA base pair was replaced with an identical rebuilt base pair, the crystal structure conformation of the four side chains had the lowest computed energy. When base pair 5 was mutated, each side chain again chose its crystal structure conformation. When base pair 6 was mutated to any of the three other base pairs, the His 49 side chain changed conformation slightly to avoid a poor electrostatic interaction with the mutated bases. When base pair 5 was mutated, Arg 46 changed conformation to avoid a steric clash with the mutated base pair, and His 49 moved to accomodate the new conformation of the Arg 46 side chain.

Mutations of protein side chains to Ala were carried out by simply removing the atoms of the side chain beyond C_β.

3.3 Results

The Zif268–DNA complex is shown in Figure 3.1 [38]. It consists of three zinc finger domains, each of which makes contacts to a three base-pair subsite on the DNA. The total electrostatic contribution to protein–DNA binding in the Zif268 complex was calculated using a continuum electrostatic model. The overall electrostatic contribution was +25.7 kcal/mol, which is unfavorable as in many other macromolecular complexes that have been studied by this method [59, 93, 95, 100, 101, 123, 127]. The cost of protein desolvation was 93.4 kcal/mol and that of DNA desolvation was 39.9 kcal/mol. These desolvation costs were not fully compensated by the total screened intermolecular interactions of −101.7 kcal/mol and the total intramolecular contribution of −5.9 kcal/mol; hence the unfavorable overall electrostatic contribution to binding.

The total electrostatic binding free energy was decomposed into a sum of terms representing desolvation and pairwise interaction terms between groups of atoms in the protein. The protein was divided into amino groups (consisting of H–N–C_α), carbonyl groups (C–O), and side chains. The DNA was divided into phosphate

(including the O5' and O3'), ribose, and base groups. Each zinc center was considered together with the four side chains that coordinated it as a single group. The groups were chosen to have integer charge. We define three types of terms: solvation, direct, and indirect. The solvation term is the loss of solvent interactions upon binding, which is always unfavorable. A direct interaction is simply an intermolecular solvent–screened coulombic interaction between two groups. An indirect term results from an intramolecular interaction between two groups in the same portion of the complex. Because the two groups interact in both the bound and unbound states, only the *difference* in interaction energy between bound and unbound states contributes to the overall binding free energy. We define this difference as the indirect term. Since the protein and DNA were treated as rigid in this calculation, the indirect terms simply reflect a difference in solvent screening. Thus, we might expect the indirect term for an arginine and asparate pair to be favorable, because binding typically reduces the solvent screening, thereby giving an interaction energy that is larger in magnitude.

3.3.1 Solvation Effects

The total desolvation free energy contribution of 133.7 kcal/mol was dominated by contributions from charged groups. The charged side chains of the protein contributed 82.3 kcal/mol and the phosphates contributed 27.5 kcal/mol for a total of 109.8 kcal/mol from charged groups. Most of the remaining desolvation penalty was from the DNA bases, which accounted for 11.9 kcal/mol, and from the polar side chains of the protein, which contributed 9.3 kcal/mol.

The zinc centers in fingers 1 and 2, when considered with their coordinating residues, paid desolvation penalties of 1.7 and 1.5 kcal/mol, respectively. The bulk of this solvation free energy ($> 90\%$) was from His 25 and His 53, respectively, which contact phosphates in the primed strand of the DNA. It is likely that the zinc of finger 3 would also be desolvated in a native complex; however, the DNA in the crystal structure does not have a phosphate on residue 12' (i.e., the 5' end of the primed DNA strand). This phosphate would be expected to desolvate His 81, but its

absence resulted in a desolvation penalty for the finger 3 zinc center of less than 0.1 kcal/mol.

3.3.2 Effective Pairwise Interactions

Table 3.2 lists all effective pairwise interactions involving protein and DNA and greater than 2.0 kcal/mol in magnitude. Direct interactions are listed at the top of the Table, and indirect interactions are listed below the line. Both sets are sorted according to magnitude. All of the larger favorable interactions (< -4.0 kcal/mol) involve group pairs that formed two hydrogen bonds in the structure. In addition, there are five pairs that are computed to interact strongly but do not form hydrogen bonds. The side chains of Arg 24, Arg 46, and Arg 74 each made strong interactions with one or two phosphates but only made direct contacts with a guanine base. Several phosphate groups also made strong interactions with Arg side chains in the protein through hydrogen bond contacts. Zinc groups 1 and 2, which include His 25 and His 53, also form hydrogen bonds with the phosphate backbone that result in substantial favorable interactions.

Five Arg–Gua pairs in the complex were among the strongest interactions, even though they represent charged–polar, as opposed to charged–charged, interactions. Three of the Arg residues (at the -1 position in each zinc finger helix) were hydrogen-bonded to Asp side chains; indirect contributions of these salt bridges were nearly as strong as the strongest direct interactions, which is remarkable considering that only the change in their mutual solvent screening upon binding contributed.

Overall, indirect interactions in the complex made a small contribution compared to direct interactions and solvation; however their role was still significant. The total $\Delta\Delta G_{indir}$ for DNA was +8.3 kcal/mol and for protein was -14.2, for a total $\Delta\Delta G_{indir}$ of -5.9 kcal/mol (Table 3.1). The unfavorable contribution of DNA to $\Delta\Delta G_{indir}$ was caused mainly by enhanced phosphate–phosphate repulsions, which accounted for +16.4 kcal/mol (partially compensated by other indirect interactions within the DNA). The favorable contribution of the protein to $\Delta\Delta G_{indir}$ resulted from the three Arg–Asp pairs, which each contributed 4.8–5.2 kcal/mol to binding affinity. Other

substantial indirect contributions were made by Arg 74–Glu 77 and Arg 24–Asp 48 pairs. While neither formed a hydrogen bond, each contributed −2.4 kcal/mol to $\Delta\Delta G_{indir}$. No individual contributions greater than 2 kcal/mol in magnitude resulted from pairwise interactions of a polar, uncharged group pair or from a single pairwise phosphate–phosphate interaction.

By far the largest unfavorable $\Delta\Delta G_{indir}$ contribution was +5.3 kcal/mol, made by Arg 24–Arg 46. These two arginine side chains were in adjacent fingers, stacked on top of one another. In the unbound state they were both exposed to solvent, but in the bound state, they became fully buried, which increased the strength of their electrostatic repulsion. However, this unfavorable interaction was compensated by several favorable interactions made by each of these groups, some of which are listed in Table 3.2.

The protein backbone had an interaction with the phosphates of −11.2 kcal/mol. The interactions did not include any direct hydrogen bonds between the α-helices and the phosphate. They are favorable because the positive N-terminal end of the α-helix backbone dipoles is directed toward the DNA. The protein backbone's interactions with the other groups of the complex are small compared to this effect (unfavorable by a total of 2.2 kcal/mol).

3.3.3 Contributions of individual chemical groups

We define the contribution (denoted $\Delta\Delta G_{contrib}$) of a group to binding by adding its full desolvation penalty to half of its direct and indirect interactions (each interaction has half its value assigned to each group in the pair). In this way, the group contributions are additive so that the sum of contributions is equal to the total electrostatic binding free energy. The $\Delta\Delta G_{contrib}$ values do not correspond to a simple experiment — rather, they provided a convenient deconstruction of the electrostatic binding free energy energy among the constituent groups in the system. A complementary method of describing the role of each group is given by the mutation free energy (denoted $\Delta\Delta G_{mut}$), which is the sum of a group's desolvation penalty and its full interactions. The mutation energies do not add to give the full binding free

energy, but they do correspond to at least a conceptual experiment. The $\Delta\Delta G_{mut}$ of a group represents the effect on binding of mutating from a group with no partial atomic charges (a hydrophobic isostere) to the fully charged group.

Table 3.3 shows all groups with $\Delta\Delta G_{contrib}$ or $\Delta\Delta G_{mut}$ larger than 1.0 kcal/mol in magnitude. These groups are also illustrated in Figure 3.2. The five most favorable contributions were all from the five Gua bases hydrogen-bonded to Arg side chains from the zinc fingers. Since the bases were electrostatically neutral, their desolvation penalty was relatively low (0.8–1.0 kcal/mol for each of the five Gua bases). The net result was that their desolvation penalties were easily compensated by their favorable interactions with arginines (see Table 3.2.) The only other interactions larger than 1.0 kcal/mol in magnitude involving these Gua bases were the unfavorable interactions of Gua 10, 7, and 4 with the Asp side chains that are hydrogen-bonded to their respective Arg partners. (Each of these repulsions was +1.1–1.2 kcal/mol.)

The only other favorable contribution larger than 1.0 kcal/mol in magnitude was that of Arg 55. This is somewhat surprising considering that Arg 55 did not contact the DNA directly. In fact, its failure to contact the DNA was part of the reason that its contribution was favorable. Its desolvation penalty was only 0.5 kcal/mol, but it made favorable interactions with phosphates 4, 7', and 8' (−0.5, −1.0, and −1.1, respectively). Removing the charge from this Arg side chain resulted in a computed loss of 2.8 kcal/mol of binding affinity. The two side chains that occupied homologous positions in fingers 1 and 3, Arg 27 and Lys 83, also did not contact the DNA. These groups had favorable values of $\Delta\Delta G_{mut}$: −1.2 kcal/mol and −1.3 kcal/mol, respectively.

We are not aware of experiments that have probed the role of these residues directly. To determine the possible significance of these residues, we aligned 1949 different Cys_2–His_2 zinc finger sequences from SWISS-PROT. The alignment showed that at this position in the zinc finger, an Arg or Lys residue was present in 71% of the sequences, consistent with previous statistical studies [65]. This observation supports the conclusion from the electrostatic calculation, which suggests that a positively charged residue at this position contributes strongly to DNA binding affinity, despite

the fact that in this structure, the side chains do not contact the DNA directly.

There are two other positions in a zinc finger domain where Arg/Lys is even more highly conserved. The first is Arg 3/Lys 33/Lys 61 in the Zif268 structure, a position occupied by Arg/Lys in 80% of the sequences in the database. Arg 3 had a computed $\Delta\Delta G_{mut}$ of -3.5 kcal/mol, and it contacted phosphate 8. Lys 33 and Lys 61 did not contact the DNA directly, but were also predicted to have favorable $\Delta\Delta G_{mut}$ values of -1.0 and -1.3 kcal/mol. These two lysines were part of a conserved TGEKP linker [65], which is known to contribute to binding affinity [23, 74]. This analysis suggests that the Lys contributed to binding affinity due to its interactions with phosphates, which more than compensated for its desolvation penalty. The other highly conserved, positively charged position is Arg 14/42/70 in Zif268. These residues also interacted favorably with phosphates. Arg 14 and 42 had $\Delta\Delta G_{mut}$ values of -2.7 and -1.5. Arg 70 had a computed $\Delta\Delta G_{mut}$ of 0.1 kcal/mol, but phosphate 1 was not present because the 5' end of the DNA in the crystal structure was not phosphorylated. Hence it may make better interactions with a long strand of DNA than with the oligonucleotide in this structure. There are a total of 15 Arg/Lys residues in Zif268 that are conserved in at least 20% of the sequences in the database. All of these have substantially favorable $\Delta\Delta G_{mut}$ values (with the exception of Arg 70) shown in Table 3.3. Conversely, there are 5 Arg/Lys residues in Zif268 that are not conserved in the sequence database. Of the five (Arg 15, Arg 38, Lys 71, Arg 78, Arg 87), only Arg 78 is predicted to have a substantial interaction with the DNA. Leucine occurs at this position in 86% of the zinc finger sequences, which may indicate a different type of structural or functional role for this residue. In any event, there is a reasonably good correlation between positively charged residues that are predicted to contribute to binding affinity and residues that are conserved as Arg or Lys.

There was a larger number of residues with an unfavorable $\Delta\Delta G_{contrib}$ than with a favorable contribution. Five of the seven most unfavorable contributors were the five arginines partnered with guanines in the DNA. While the Gua's were neutral and had small desolvation penalties, the Arg's were charged (and perhaps more highly desolvated), and hence paid a much larger desolvation penalty, but recover the

same Arg–Gua interaction. The Arg contribution terms included other significant interactions with phosphates and nearby negatively charged side chains, but these were not enough to fully compensate the desolvation penalties of 6.0–10.8 kcal/mol incurred by these five arginines (all the desolvation penalties are 8.3 kcal/mol or more except for Arg 80, at the outside of finger 3). Other groups with an unfavorable $\Delta\Delta G_{contrib}$ included the side chains of Glu 21 and 77, at position 3 of the helices in fingers 1 and 3. These two residues may affect DNA specificity, as discussed below. Additional unfavorable $\Delta\Delta G_{contrib}$'s came from the Asp side chains at position 2 of each helix, and from a few phosphates in the DNA. Although their $\Delta\Delta G_{contrib}$ was unfavorable, the removal of charge from most of these groups would not greatly enhance binding affinity (as reflected in $\Delta\Delta G_{mut}$). The scarcity of positive $\Delta\Delta G_{mut}$ values indicates that the electrostatic interactions are networked to a significant extent: removing all charge from the complex would substantially favor binding, but removing charge from any single group is likely to hamper binding.

3.3.4 Pairwise Mutation

In previous work, it has been shown that electrostatics generally tend to be destabilizing. In particular, results have shown that salt bridges and hydrogen bonds are often unfavorable to protein folding and binding [59, 60]. In the Zif268–DNA complex, the effects of several pairwise mutations from hydrophobic isosteres to actual polar and charged chemical groups are shown in Table 3.4 for pairs of groups that spanned the binding interface. Table 3.4 lists all the pairs of groups that formed an intermolecular hydrogen bond involving a charged group. The first ten entries are charged–neutral group interactions, and the last three are salt bridges. Zinc groups 1 and 2 refer to to the zinc ions in fingers 1 and 2 and their four coordinating side chains. The groups hydrogen bond to DNA phosphates via His 25 and His 53 side chains, respectively, and these His side chains are responsible for the electrostatic interactions of these groups. If we consider the charges only on these His side chains instead of the zinc and all surrounding side chains, then none of the entries in Table 3.4 changes by more than 0.1 kcal/mol. In Table 3.4, $\Delta\Delta G_{bridge}$ is the solvent-screened

electrostatic interaction free energy between the two bridging groups. $\Delta\Delta G_{env}$ is the total contribution to binding from the group pair that was not counted in either $\Delta\Delta G_{solv}$ or in $\Delta\Delta G_{bridge}$. It includes the solvent-screened interaction each of the groups made with its binding partner plus the indirect contributions to binding from within the same partner. $\Delta\Delta G_{total}$ is the total effect on binding of the charges in the two groups.

The results show that many pairs of electrostatic groups favored binding slightly, in contrast to previous results which showed a general trend for electrostatics to disfavor binding. The reasons for this difference are threefold: (1) the analysis for the Zif268 complex involved binding rather than folding, and many of the groups were already partly desolvated in the unbound state; (2) there were many charged residues in the complex, and these tended to make additional favorable interactions that stabilized each group pair through network effects; and (3) many of the interactions in the complex were charged–polar, and the polar group only had a small desolvation penalty. In this complex, the charged–polar interactions were more favorable on average than the salt bridges. (Here we are ignoring the interactions with phosphate 7, which is complicated because it is a three-way network of charged–charged–polar interactions.) Interestingly, when $\Delta\Delta G_{env}$ is ignored, all of the interactions Table 3.4 become unfavorable. All pairs require favorable interactions with other charged and polar groups in the complex in order to be favorable overall.

3.3.5 DNA Mutations

To consider the basis of zinc finger DNA-binding specificity, we built several DNA mutants of the Zif268 zinc finger complex. Table 3.5 shows the results of electrostatic calculations, given as $\Delta\Delta G$ with respect to the values calculated for the crystal structure. The structure of finger 2 is shown in Figure 3.3. Semi-quantitative binding data for these mutants has been obtained by Nardelli et al. [97]. Their experiments were conducted on Krox-20, a zinc finger transcription factor which has 94% sequence identity to Zif268 in its zinc-finger regions. The two sequences are 100% identical in finger 2, the finger closest to the DNA subsite that is considered here. The

experimental results of Nardelli et al. are included in Table 3.5. Here, +++ indicates binding to the mutant within 2.5-fold range of wild-type, + indicates binding 5- to 20-fold lower than wild-type, and − indicates binding at least 20-fold lower than wild-type. Additional experimental results are available from Swirnoff and Milbrandt [138], who selected DNA sequences with affinity to either Zif268 (also known as NGFI–A) or Krox-20. Their selection experiments demonstrated a preference for the sequence T(G/A)G. However, a more quantitative gel shift assay showed that at position 5, the DNA in this structure had only a 1.5-fold higher affinity for Krox-20 than when G was substituted at position 5, and only a 4-fold higher affinity than when C was substituted. Despite this relatively weak preference, T was observed in 100% of the oligonucleotides selected for binding to Krox-20. In a similar experiment with Zif268, the preference for T was not as pronounced: T was selected with a frequency of 83.6%, and G had a frequency of 14.5% at this position. If the difference in these two results were not due to random variation in the oligonucleotide pools, they must be due to sequence differences far from the binding site, as Krox-20 and Zif268 have identical sequences in finger 2.

The computational results predicted that G was slightly favored over T at position 5, but otherwise agreed with the results of Nardelli et al. that all bases can be tolerated at position 5. The computational results also showed that only G or A could be tolerated at position 6, as shown experimentally by both Nardelli et al. and Swirnoff and Milbrandt. The calculated structures placed adenine or guanine so that N7 could form a hydrogen bond with His 49. The substitution of a pyrimidine base forced the His side chain to move, thereby disrupting any hydrogen bond that could form and decreasing the affinity.

At position 7 of the DNA, the calculations predicted a clear preference for G over any other base. The experimental results of both Swirnoff and Milbrandt and Nardelli et al. predicted this preference. However, the results of Nardelli et al. showed that the preference for G over the other three bases was only 5- to 20-fold in magnitude — not as large a difference as the calculations suggest. In the cases of the GGC or GGA sequences, the exocyclic amines of Cyt or Ade made an unfavorable interaction

with Arg 46 in the protein. The way in which the complex avoids this unfavorable interaction may be difficult to determine without allowing more flexibility at the protein–DNA interface.

3.3.6 Protein Mutations

To further examine the importance of electrostatic interactions in this system, we compared the computed electrostatic interactions of several protein mutants to experiments of Elrod-Erickson and Pabo [37]. The mutations involved the four side chains of finger 1 that contacted DNA bases. These side chains are illustrated in Figure 3.4. Table 3.6 lists the contributions to the electrostatic binding free energy of each of these mutants. $\Delta\Delta G_{mut}$ is as defined above—the energy required to charge the residue without changing the shape of the protein. $-\Delta\Delta G_{mut}$ is given in the Table because the charges are removed from each residue when they are mutated to Ala. We list $\Delta\Delta G_{shape}$ as the change in electrostatic interactions of all other groups caused by the change in shape when the residue is mutated to Ala. $\Delta\Delta G_{elec}$ is the total change in electrostatic binding free energy due to the mutation, or simply $-\Delta\Delta G_{mut} + \Delta\Delta G_{shape}$. $\Delta\Delta G_{surf}$ is the free energy of the two hydrophobic cavities associating in solution. This is typically taken to be proportional to the change in solvent accessible surface upon binding. We use a proportionality constant of 25 cal/mol/Å^2 favoring hydrophobic burial [125]. Finally, $\Delta\Delta G_{total}$ is the total contribution of these terms to the change in binding free energy due to mutation.

Table 3.6 shows that mutating either Arg 18 or Arg 24 was unfavorable to binding. Removing the charges from either residue was unfavorable, but this was compensated to some extent by solvating the complex (indicated by a favorable $\Delta\Delta G_{shape}$ term). However, the Ala mutants buried less surface area than their Arg-containing counterparts, and thus $\Delta\Delta G_{total}$ increased to 3.3 and 3.4 kcal/mol for Arg 18 and Arg 24, respectively. This agreed well with experimental values. Mutation of Asp 20 was computed to slightly decrease binding affinity due to its favorable indirect interactions, particularly with Arg 18. The experiment indicated that it was slightly favorable, but the results agree to within less than 1 kcal/mol.

The largest discrepancy between calculation and experiment occurred with the Glu 21 mutant. Glu 21 has a desolvation penalty of 4.1 kcal/mol and its interactions with the remainder of the complex were slightly unfavorable. In the structure, it made no hydrogen bonds in the bound state. Thus, one might expect that changing it to an uncharged residue would be favorable. The calculation did predict that mutation to Ala should be favorable by 4.5 kcal/mol; however, the experiment showed that it was slightly unfavorable. The reason for this difference is unclear. One possibility is that Glu 21 has an unusually high pK_a, and therefore does not have to pay the desolvation cost computed here. A pK_a calculation on the bound state of the molecule gave a pK_a shift of 0.5 pK_a units, which is too small to account for the observed discrepancy between the experiment and the calculation.

Elrod-Erickson and Pabo also discuss the role of Glu 21 in specificity. They demonstrate that wild-type Zif268 had 13-fold higher affinity for the wild-type DNA sequence GCG than for a mutant GAG site. They also showed that the E21A mutant Zif268 lacked this specificity and actually had slightly higher affinity for the GAG site. The calculations provide a rationale for this specificity. Glu 21 had a favorable interaction of -1.3 kcal/mol with Cyt 9 (the middle base pair of the subsite). Despite the fact that there was no hydrogen bond formed, the negative charge on the Glu side chain interacted favorably with the amino group of the cytosine. If the cytosine was changed to adenine, this interaction would be lost (and there would likely be a slight repulsion of the adenine's hydrogen bond acceptors), and thus the Glu side chain is responsible for specificity. Here we are assuming that the desolvation penalty paid by Glu upon DNA binding would not be changed substantially by the DNA base sequence—a reasonable assumption if the overall structure of the complex remained the same.

Asp 20 has also been shown to have a role in specificity [37], as its mutation to Ala resulted in a 5-fold decrease in specificity. The negative charge on the Asp was computed to have a favorable interaction of -1.0 kcal/mol with cytosine 10', despite the absence of a good hydrogen bond. This interaction would be lost upon mutation of cytosine-10' (the GCG subsite) to guanine (the GCC subsite). Therefore

the specificity at this position can also be explained by an electrostatic effect.

The R18A/D20A double-mutant allows a potentially interesting observation. The role of interacting residue pairs has been measured in the past by means of a double-mutant cycle [21, 63, 120, 122]. To measure the interaction between two side chains, the effect of mutating both simultaneously to a reference amino acid is subtracted from the effects of the two individual mutations. The difference gives an effective cooperativity between the sites that may be viewed as an interaction free energy, although the analysis may be complicated by structural changes in the various mutants. In this system, the interaction in question is actually the *change* in interaction free energy between the side chains of Arg 18 and Asp 20 caused by DNA binding.

Experimental mutation of Arg 18 and Asp 20 to Ala simultaneously caused a decrease in binding affinity of 1.9 kcal/mol. The sum of the individual effects was (2.7 + (−0.3)) 2.4 kcal/mol. The difference of −0.5 kcal/mol may be attributed to an indirect interaction between Arg 18 and Asp 20. The hydrogen bonds that they form were predicted to be stronger upon binding of DNA, since the DNA displaces the solvent that screens their interaction with one another. The electrostatic interaction between these two residues was predicted to contribute −5.3 kcal/mol to binding affinity (see Table 3.2). When shape effects of mutation to Ala were considered in the calculation, then the predicted interaction between the two side chains became −3.5 kcal/mol. Part of the discrepancy between this value and the experimental value may be due to conformational changes — especially the additional flexibility of the singly-mutated proteins when they are not bound to DNA. Nevertheless, the experimental result suggests that this indirect effect may contribute to binding affinity.

3.3.7 Alternative Definition of $\Delta\Delta G_{contrib}$

As noted earlier, we define $\Delta\Delta G_{contrib}$ so that the sum of contributions over the whole complex adds up to the total electrostatic binding free energy of the system. This was originally done by assigning half of each interaction to each group in the interacting pair. Then each group was examined to see if its interactions undercompensated or

overcompensated its desolvation penalty. A different way to answer this question would be to divide the interaction of two groups so that the group that has a larger desolvation has a larger share of the interaction.

The contribution of each group when the interactions are weighted is shown in Figure 3.5. This definition results in fewer highlighted groups (9 instead of 21), and some of the groups make substantially different contributions. For example, Arg 24 and Arg 46 were unfavorable by 3.1 kcal/mol using the old definition, but have favorable contributions of -1.6 and -1.9 kcal/mol, respectively, using the weighted definition. This is not because of a change in the net result of the calculation; it is simply a different way of adding up the numbers. This picture shows that most of the groups in the complex do manage to nearly compensate their desolvations with favorable interactions. The fact that binding is unfavorable overall is due to small contributions (less than 1.0 kcal/mol in magnitude) from many groups in the complex.

3.4 Discussion

The electrostatic contribution to the free energy of binding in the Zif268–DNA complex is unfavorable by 25.7 kcal/mol. The electrostatic contribution is simply the difference in binding affinity between this complex and a hypothetical, hydrophobic pair of molecules with the same shape. The net unfavorable electrostatics should not be taken to mean that binding of this complex is unfavorable, especially since binding of two completely hydrophobic molecules would be expected to be quite favorable. It also does not indicate that electrostatic properties of the molecules are unimportant. If, for example, either the DNA or the protein were made completely hydrophobic, the electrostatic contribution to binding would increase to either 79.3 kcal/mol or 48.1 kcal/mol, respectively. Failing to compensate groups that are desolvated on binding is clearly destabilizing to the complex.

In fact, there are relatively few groups which destabilize the complex relative to their hydrophobic isosteres. Only three groups are predicted to have an unfavorable $\Delta\Delta G_{mut}$. Two of these are Glu side chains in position 3 of the α-helix–Glu 21 and

Glu 77. Since these residues are in similar positions on similar fingers that recognize the same DNA subsite, their roles are likely to be similar. Despite the fact that Glu 21 does not contribute to binding affinity, Elrod-Erickson and Pabo have shown that it contributes to specificity for a particular DNA sequence. Our calculations suggest a basis for this specificity. The charge on Glu 21, combined with the fact that the side chain packs closely against cytosine 9 and that both groups are buried, makes a favorable electrostatic interaction with the exocyclic amine of cytosine 9. This interaction would not be possible with other bases in this position, and thus Glu 21 contributes to specificity. Glu 77 has similar geometry and similar calculated energetics, and likely plays the same role. The only other group computed to have an unfavorable $\Delta\Delta G_{mut}$ is phosphate 3. It appears that the Zif268 is well-optimized for electrostatic binding affinity.

The electrostatic calculations provide a rationale for much of the observed data on the roles of individual bases and side chains in binding affinity. In fingers 1 and 3, the amino acids RDER (at positions -1, 2, 3, 6 of the α-helix) recognize the base sequence GCG. It was already recognized [38] that Arg 18 and Arg 24 recognize guanines 10 and 8 by making a pair of hydrogen bonds to each of them. The calculations here correctly account for the cost of mutating each Arg side chain to within a few tenths of a kcal/mol. The desolvation penalty each Arg side chain pays (more than 8 kcal/mol in both cases) is compensated by favorable interactions. In the case of Arg 18, interactions with Asp 20 and guanine 10 (each more than -5.0 kcal/mol favorable) are important. In the case of the more buried Arg 24, interactions with phosphates 6, 7, and 8 are predicted to be significant (totaling -6.2 kcal/mol).

In finger 2, the amino acids RDHT recognize the base sequence TGG. Arg 46 coordinates guanine 7 (the third base listed) and is again responsible for affinity and specificity, as previously observed. His 49 is expected to make a modest contribution to binding affinity ($\Delta\Delta G_{mut} = -0.6$) and appears to specify A or G at base position 6. Thr 52 does not appear to play a role in either affinity or specificity, and there appears to be little base discrimination at position 5 according to either calculations or experiment [97].

Some aspects of recognition in this complex were straightforward to observe from the crystal structure. One example is all of the Arg–guanine pairs, which were thought to be important for specificity and affinity as soon as the crystal structure was solved. It is not particularly surprising to find that Arg specifies guanine over other bases, but it is somewhat novel to suggest, as these results do, that Arg side chains only contribute to affinity because of other charged groups in the complex, such as Asp 20 and the more distant phosphates. The computed results also rationalize other experimental observations that are less apparent from the crystal structure. For example, Glu 21 is predicted to specify cytosine *via* a non-hydrogen bonded electrostatic interaction. Asp 20 is predicted to contribute to base specificity *via* a non-hydrogen bonded, water-bridged hydrogen bond with cytosine 10'. Another example of a novel prediction from these results is that Arg 55, which makes no contacts with the DNA at all, is expected to make a substantial contribution to DNA binding ($\Delta\Delta G_{mut} = -2.8$ kcal/mol). This residue is far enough from the interface to pay a small desolvation penalty of 0.5 kcal/mol, but close enough to make substantial electrostatic interactions with the phosphates in the "primed" DNA strand. This sort of positioning of a residue with an appropriate charge near a binding interface may be a simple way to engineer affinity. Mutation of residues in a similar position has been observed to enhance affinity in a β-lactamase–inhibitor complex [1].

Comparison with 434 repressor, Arc repressor. Similar electrostatic calculations have been performed on two other protein–DNA complexes in the past—the 434 repressor–OR1 operator complex and the Arc repressor–operator complex [57]. In both of these systems, electrostatic binding free energy was found to be unfavorable (by 31.0 and 35.8 kcal/mol, respectively) because electrostatic interactions did not fully compensate for the desolvation penalties paid by the two molecules. Charged residues that do not contact the DNA made significant contributions to affinity in all three complexes.

One trend that emerges in all three complexes is that salt bridge formation tends to be favorable on average. In the Zif268 complex, two out of three interfacial salt bridges are stabilizing, and in the 434 and Arc complexes, seven out of the eight total

salt bridges in the two complexes are stabilizing. All of the stabilizing salt bridges in the three complexes are stable by virtue of interactions with other groups in the complex. Burying a positively charged protein side chain is typically predicted to be unfavorable if it only interacts with one hydrogen bonded negative charge, but in DNA, there are always other phosphates nearby, and these tend to tip the balance so that forming salt bridges is favored.

Acknowledgements

We thank Barry Honig for making the DELPHI computer program available and Martin Karplus for making the CHARMM computer program available. J.A.C. is a National Science Foundation Predoctoral Fellow and a Merck/MIT Fellow.

Table 3.1: Contributions to Electrostatic Binding Free Energy in Zinc Finger–DNA Complex.

	$\Delta\Delta G_{solv}$	$\Delta\Delta G_{dir}$	$\Delta\Delta G_{indir}$	Total
DNA base	11.9	−14.9	−3.3	−6.2
DNA backbone	28.0	−36.0	11.5	3.5
Side chain	91.7	−46.4	−14.2	31.1
Protein backbone	1.8	−4.5	0.0	−2.7
Total	133.3	−101.7	−5.9	25.7

All free energy values are in kcal/mol. Direct and indirect interactions between different groups are divided equally between the two groups.

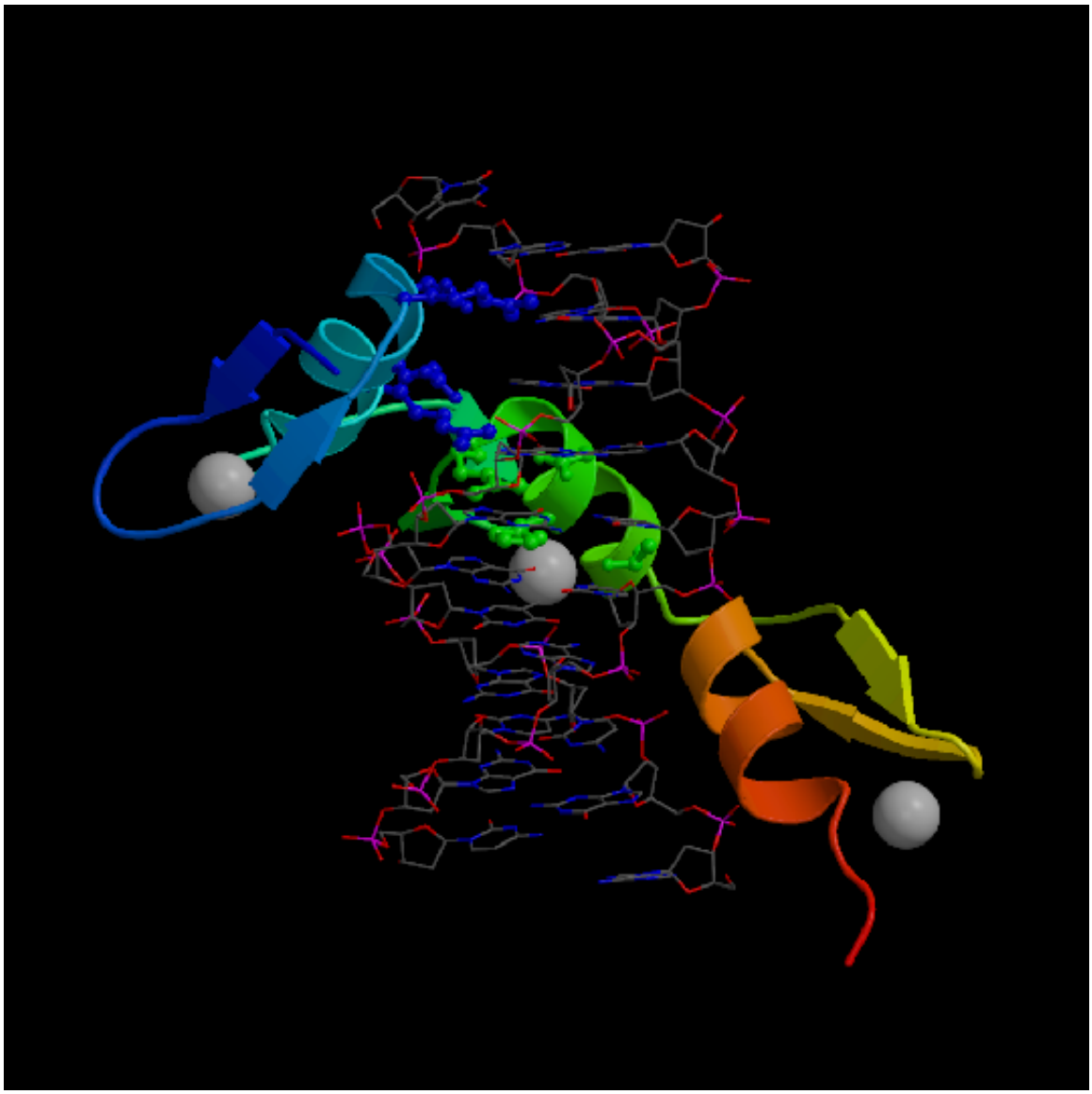


Figure 3.1: The Zif268–DNA complex [38]. Fingers 1, 2, and 3 are shown as cartoons with Finger 1 at the top of the figure. The zinc centers of the three fingers are shown as spheres. The DNA is shown as bonds with atoms colored according to element type. Figures 3.1–3.5 produced using the program MOLSCRIPT [73].

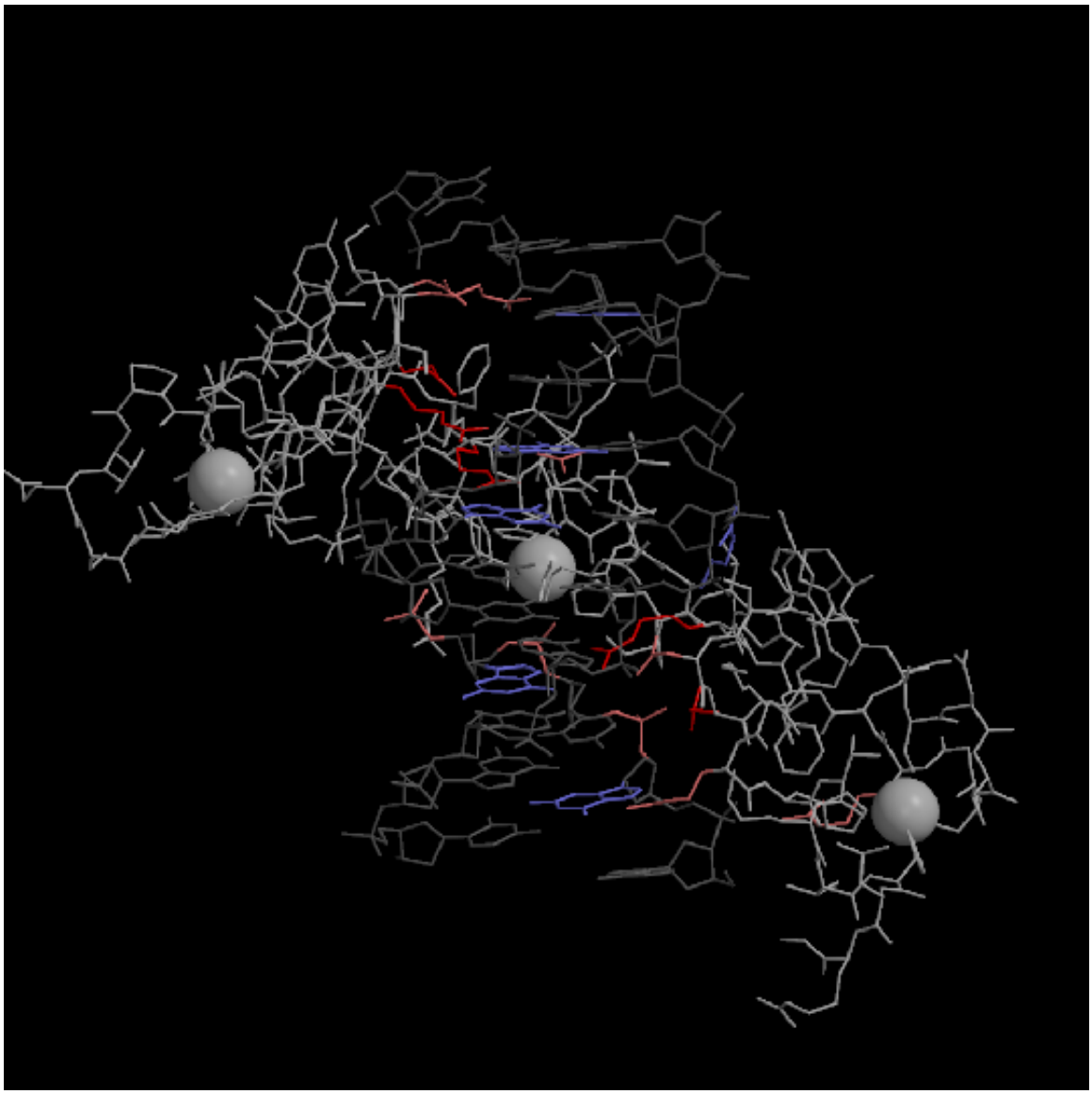


Figure 3.2: The Zif268–DNA complex shown from the same point of view as in Figure 3.1. The DNA atoms are shown in dark gray, and the protein atoms are shown in off-white. Groups with a $\Delta\Delta G_{\text{contrib}}$ of more than 1.0 kcal/mol in magnitude are highlighted: $> +3.0$ red; $+1.0$–3.0 pink; (-3.0) – (-1.0) light blue. No groups have a $\Delta\Delta G_{\text{contrib}}$ lower than -3.0.

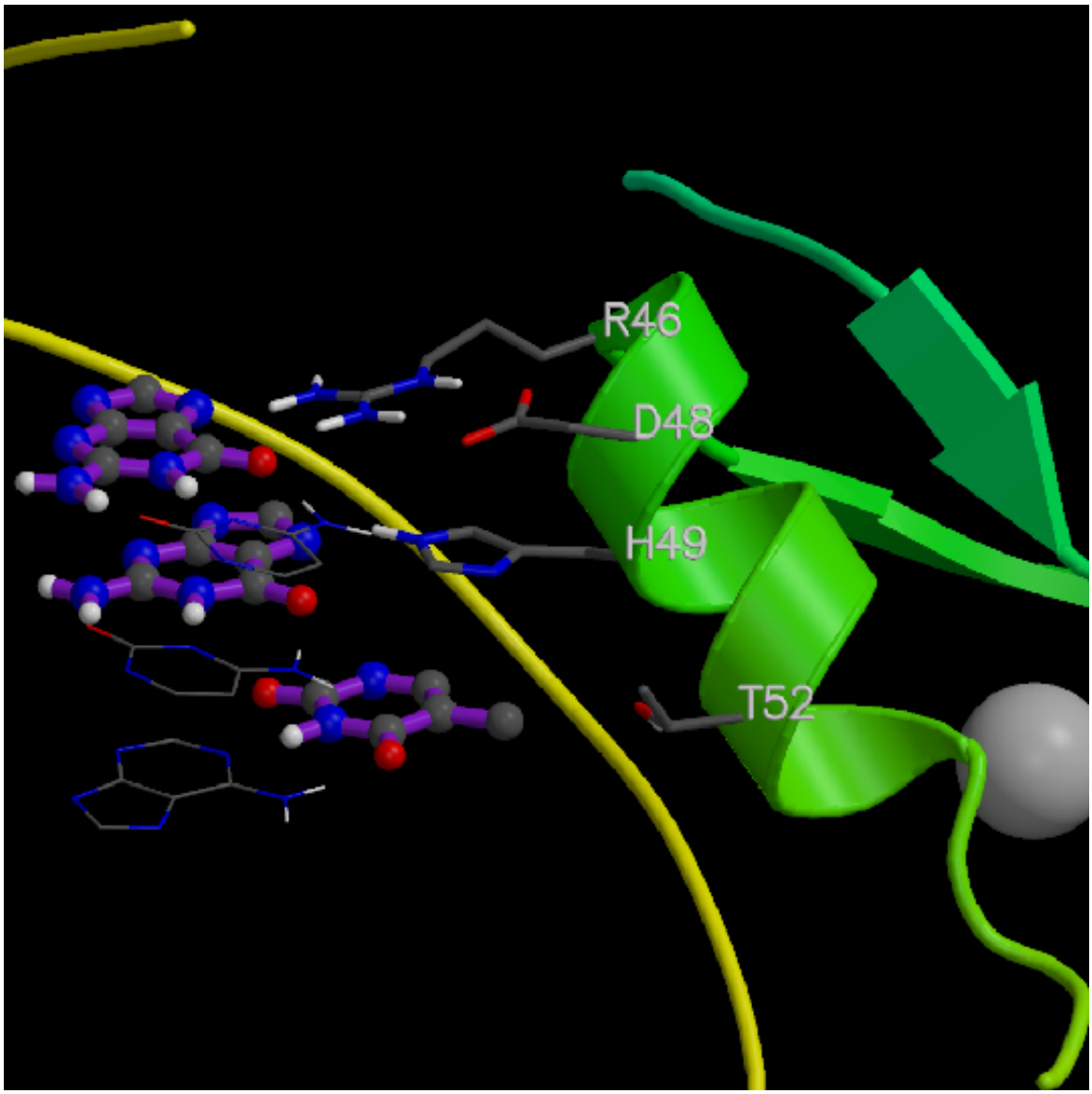


Figure 3.3: Finger 2 of the Zif268–DNA complex. The side chains that contact DNA bases are shown. The three bases contacted by these side chains are shown with balls and purple sticks. The bottom base in the figure is Thy 5, and above that are Gua 6 and Gua 7. The bases complementary to the contacted bases are shown with narrow bonds.

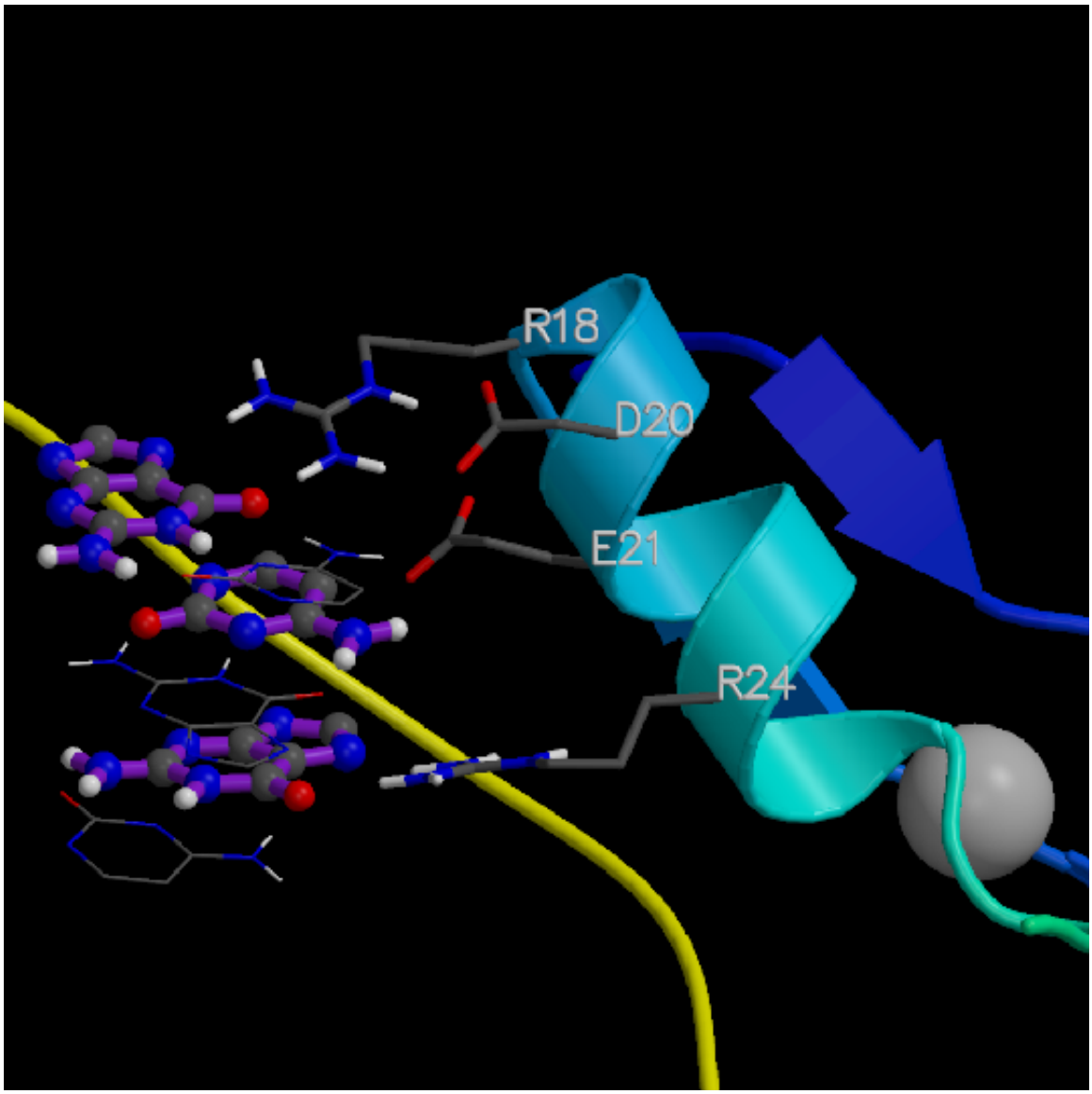


Figure 3.4: Finger 1 of the Zif268–DNA complex. The side chains that contact DNA bases are shown. The three bases contacted by these side chains are shown with balls and purple sticks. The bases complementary to the contacted bases are shown with narrow bonds.

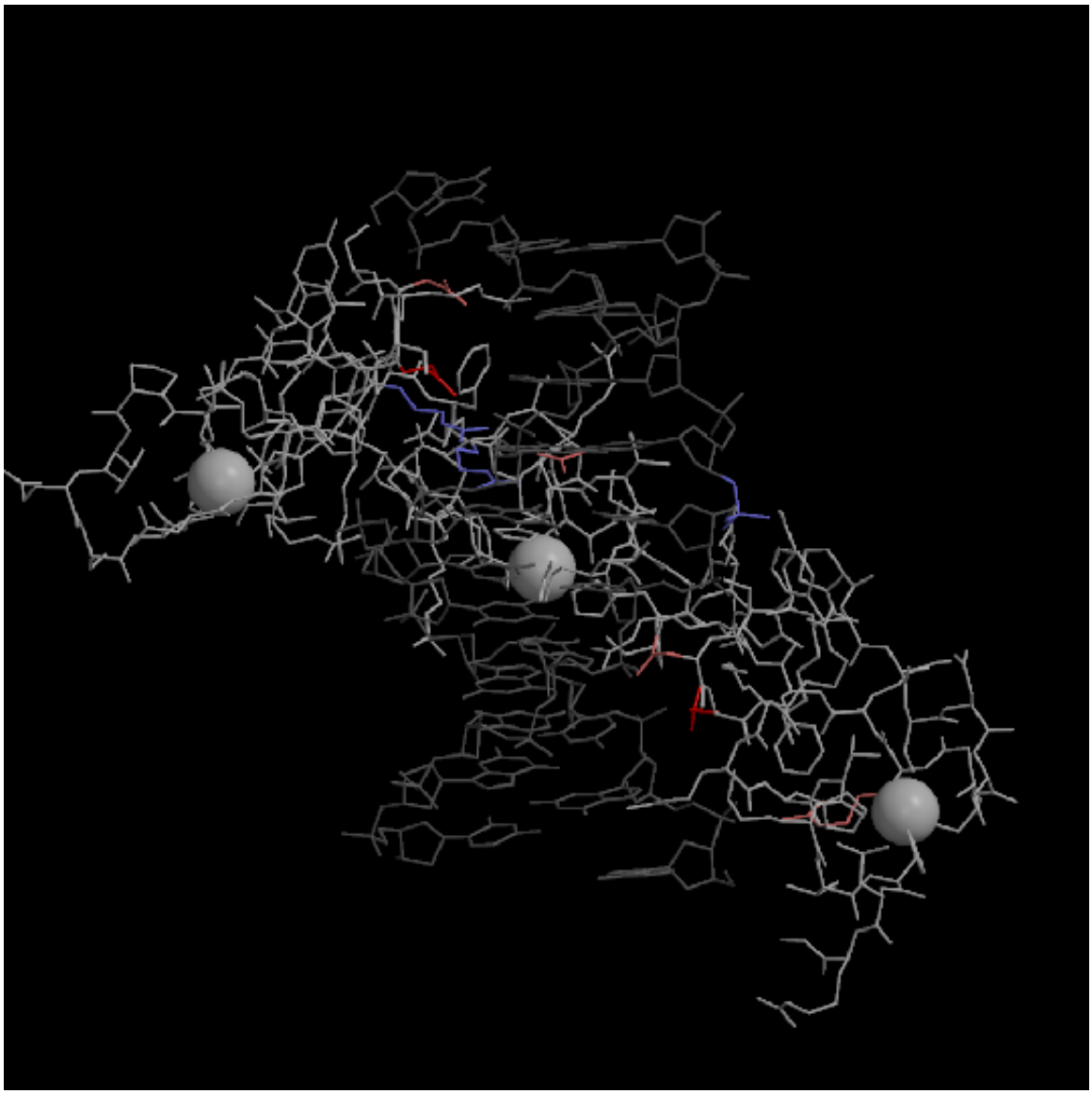


Figure 3.5: The Zif268–DNA complex shown from the same point of view as in Figure 3.1. The DNA atoms are shown in dark gray, and the protein atoms are shown in off-white. This figure shows the contributions after group–group interactions are divided asymmetrically. If group i has solvation energy $\Delta\Delta G_{solv,i}$ and group j has solvation energy $\Delta\Delta G_{solv,j}$, then group i gets a share of the i–j interaction multiplied by $\Delta\Delta G_{solv,i}$ / ($\Delta\Delta G_{solv,i}$ + $\Delta\Delta G_{solv,j}$). The color scheme for $\Delta\Delta G_{contrib}$ is the same as in Figure 3.2.

Table 3.2: Group Pair Interactions in Zinc Finger–DNA Complex.

	Interaction	$\Delta\Delta G^a$
Gua 7 –	Arg 46	−6.1
Gua 8 –	Arg 24	−5.7
Gua 4 –	Arg 74	−5.7
Gua 10 –	Arg 18	−5.6
phosphate Gua 8 –	Arg 3	−5.3
Gua 2 –	Arg 80	−4.7
phosphate Gua 7 –	Arg 14	−3.9
phosphate Gua 4 –	zinc group[b] 2	−3.8
phosphate Gua 7 –	Arg 24	−3.5
phosphate Gua 7 –	zinc group[b] 1	−3.5
phosphate Gua 2 –	Arg 70	−3.5
phosphate Gua 4 –	Arg 74	−3.0
Cyt 8' –	Asp 48	−2.9
phosphate Gua 6 –	Arg 46	−2.7
phosphate Gua 7 –	Arg 46	−2.5
phosphate Gua 6 –	Arg 24	−2.3
phosphate Cyt 7' –	Ser 75	−2.2
phosphate Gua 8 –	Arg 14	−2.1
phosphate Cyt 3 –	Glu 77	+2.4
phosphate Gua 8 –	Glu 21	+2.7
Arg 18 –	Asp 20	−5.3
Arg 74 –	Asp 76	−5.1
Arg 46 –	Asp 48	−4.8
Arg 74 –	Glu 77	−2.4
Arg 24 –	Asp 48	−2.4
Arg 24 –	Arg 46	+5.3

[a] All free energy values are in kcal/mol. Values for intermolecular interactions (above the line) are the solvent-screened coulombic energy between the groups. Values for intramolecular interactions (below the line) are differences between their interaction energy in the bound and unbound states. Negative values favor complex formation. All values greater than 2 kcal/mol in magnitude are listed.

[b] The zinc group includes the zinc center, the two His side chains, and the two Cys side chains that coordinate the zinc. Groups 1 and 2 contain the zinc ions in fingers 1 and 2, respectively.

Table 3.3: Contribution of Individual Groups to Binding Free Energy.[a]

Interaction	$\Delta\Delta G_{\text{cont}}$ [b]	$\Delta\Delta G_{\text{mut}}$ [c]
Gua 8	−2.2	−5.4
Gua 7	−1.8	−4.5
Gua 2	−1.7	−4.2
Gua 4	−1.6	−4.1
Gua 10	−1.6	−4.0
Arg 55	−1.2	−2.8
Lys 79	−0.9	−2.0
phosphate Gua 7	−0.8	−5.8
Lys 83	−0.6	−1.3
Arg 78	−0.6	−1.6
phosphate Gua 2	−0.5	−2.6
phosphate Cyt 7'	−0.5	−2.4
zinc group 2	−0.5	−2.4
Lys 61	−0.3	−1.3
Gua 6	−0.1	−1.2
Ser 75	0.0	−1.0
zinc group 1	0.1	−1.5
Arg 3	0.2	−3.5
phosphate Gua 8	0.3	−2.6
Arg 42	0.3	−1.5
phosphate Gua 6	0.4	−2.1
Arg 14	0.4	−2.7
Arg 27	0.5	−1.2
Lys 33	0.5	−1.0
phosphate Thy 5	1.3	−0.1
phosphate Gua 4	1.3	−2.0
Asp 48	1.4	−0.7
phosphate Cyt 3	1.4	1.0
Asp 76	1.7	0.0
Arg 70	2.1	0.1
Asp 20	2.2	−0.6
Arg 80	2.5	−0.9
Arg 18	2.5	−3.3
Arg 46	3.1	−3.2
Arg 24	3.1	−2.7
Glu 77	3.2	3.3
Arg 74	3.5	−3.8
Glu 21	4.2	4.3

[a] All free energy values are in kcal/mol. All group contributions larger than 1.0 kcal/mol in magnitude for either value are listed.

[b] The contribution of a group was the sum of its desolvation penalty and one-half of its interactions.

[c] The mutation term of a group is the sum of its desolvation penalty and all of its interactions at full strength.

Table 3.4: Contribution of Interacting Pairs to Binding Free Energy in Zinc Finger–DNA Complex.[a]

Group 1	Group 2	Group 1 $\Delta\Delta G_{solv}$	Group 2 $\Delta\Delta G_{solv}$	$\Delta\Delta G_{bridge}$	$\Delta\Delta G_{env}$	$\Delta\Delta G_{total}$
Gua 10	Arg 18	0.8	8.3	−5.6	−5.4	−1.7
Gua 8	Arg 24	1.0	8.9	−5.7	−6.6	−2.5
Gua 7	Arg 46	1.0	9.3	−6.1	−5.9	−1.6
Gua 6	His 49	1.0	0.9	−1.6	−0.3	−0.1
Gua 4	Arg 74	0.8	10.8	−5.7	−8.1	−2.2
Gua 2	Arg 80	0.8	6.0	−4.7	−2.5	−0.4
phosphate 4	zinc group 2	4.5	1.5	−3.9	−2.6	−0.5
phosphate 7	zinc group 1	4.3	1.7	−3.5	−6.1	−3.7
phosphate 7'	Ser 75	1.4	1.0	−2.2	−1.4	−1.2
phosphate 6	Ser 45	2.9	0.9	−1.4	−3.4	−1.0
phosphate 2	Arg 70	1.5	4.1	−3.5	−1.1	1.0
phosphate 5	Arg 42	2.7	2.1	−1.8	−2.9	0.1
phosphate 7	Arg 14	4.3	3.5	−3.9	−8.4	−4.5
phosphate 8	Arg 3	3.2	4.0	−5.3	−2.6	−0.8

[a]All free energy values are in kcal/mol. Charged–polar pairs are listed above the line and charged-charged pairs are listed below the line

Table 3.5: Change in Binding Free Energy Due to DNA Mutations in Zinc Finger Binding Site.[a]

Sequence	DNA hydration	$\Delta\Delta G_{dir}$	Protein hydration	Total Elec	Surface Area	Total Result	Expt
TGG (x-ray)	0.0	0.0	0.0	0.0	0.0	0.0	+++
AGG[c]	−0.4	1.7	−1.0	0.2	−0.3	−0.0	+++
CGG[c]	0.0	1.4	−0.8	0.7	0.1	0.8	+++
GGG[c]	1.3	−1.5	−1.1	−1.3	0.3	−1.1	+++
GGG[d]	1.4	−1.7	−1.1	−1.4	0.2	−1.1	+++
GAG[e]	−0.1	1.3	−0.9	0.3	0.4	0.7	+++
GCG[e]	0.9	3.9	−2.1	2.7	0.9	3.6	−
GTG[e]	0.2	3.7	−2.1	1.8	0.3	2.1	+
GGG[f]	1.3	−1.6	−1.1	−1.4	0.3	−1.2	+++
GGA[e]	0.1	5.8	2.0	7.9	0.2	8.2	+
GGC[e]	0.7	9.2	−5.4	4.5	0.8	5.3	+
GGT[e]	−1.0	7.7	−3.9	2.9	0.1	3.0	+

[a] All free energy values are in kcal/mol.
[b] TGG is the wild-type Zif268 sequence. This base was produced by deleting the Thy 5 base and rebuilding it as described for the other mutations.
[c] These structures were built by mutating the TGG sequence from the crystal structure.
[d] This structure was built by deleting and rebuilding Gua 6 just as if a mutation had been made at that position.
[e] These structures were built by mutating the GGG sequence, which was made by mutating base Thy 5 from the crystal structure.
[f] This structure was built by deleting and rebuilding Gua 7 just as if a mutation had been made at that position.

Table 3.6: Change in Binding Free Energy Due to Mutations in Zinc Finger.[a]

Mutation	$-\Delta\Delta G_{mut}$	$\Delta\Delta G^{b}_{shape}$	$\Delta\Delta G_{elec}$	$\Delta\Delta G_{surf}$	$\Delta\Delta G_{total}$	$\Delta\Delta G^{c}_{expt}$
R18A	3.3	−2.3	1.0	2.5	3.4	2.7
D20A	0.6	−0.2	0.4	0.0	0.5	−0.30
E21A	−4.3	−0.6	−4.9	0.4	−4.5	0.24
R24A	2.7	−1.1	1.6	1.7	3.3	3.6
R18A.D20A	−1.4	−1.5	−2.9	3.3	0.4	1.9

[a] All free energy values are in kcal/mol.
[b] The change in interaction and solvation of all groups besides the mutated group due to the change in shape of the dielectric boundary.
[c] Experimental values from Elrod-Erickson and Pabo

Chapter 4

Electrostatic Complementarity Applied to the Pairing Problem in Functional Genomics

4.1 Introduction

Genomic sequencing efforts reveal homologous protein families whose members are expected to share certain features of structure and function yet exhibit different specificities. An important step in understanding genome function is to learn which members of one family bind to which members of another — a question which we term the pairing problem. The goal of this study is to develop a computational method that will address this problem. Such a method will be practical for helping assign a function to the overwhelming number of sequences that are currently available.

Homology modeling will be used to produce structures of the complexes from their sequences, since homology modeling is currently the most accurate method of structure prediction. If we know that two molecules A and B bind to one another, then we will model each of A's homologues in complex with each of B's homologues in order to determine which complexes can actually form. Each modeled complex will then be evaluated by calculating the complementarity of the two molecules in

an appropriate way. Shape complementarity could in principle be a useful measure; however, assessments of shape complementarity are very sensitive to the detailed placement of individual atoms. Small errors in homology modeling can dramatically degrade complementarity. Instead, we will apply a novel definition of electrostatic complementarity to the pairing problem.

Previous work by McCoy et al. [88] has shown that several different protein–protein complexes exhibit electrostatic complementarity. Their definition accounted for the electrostatic potential produced by each molecule, but it did not account for interactions between the molecules and solvent. Other work has shown that it is possible, within the framework of continuum electrostatics, to determine an electrostatically optimal ligand for any given receptor [69, 79]. The predicted optimal charge distribution has been shown to be similar to a known tight-binding ligand [80]. The optimum charges are selected because they strike an appropriate balance between the cost of desolvating the molecule and the benefit of interacting favorably with the target receptor [22, 81]. We will make use of a complementarity measure that accounts for how well such a balance is struck in each hypothetical complex, and thus predict which A and B molecules are likely to form complexes.

The idea will be tested on the complexation of two halves of myoglobin. Myoglobin is a well-studied protein for which several crystal structures from homologous species are available. It folds by early formation of a core consisting of its A, G, and H helices, which may precede even hydrophobic collapse [44, 106]. In this study, myoglobin will be treated as a complex between the core of A, G, and H helices and the rest of the molecule, subsequently referred to as the A portion and the B portion of the molecule, respectively. Electrostatic complementarity is used as a basis for choosing which of several B portions binds best to a particular A portion. First the method will be tested on the known structures. Then, one structure will be used to produce homology models of all the possible A–B pairs so that we may evaluate the usefulness of the method when only sequence information is available.

4.2 Theory

A procedure for illustrating and evalulating electrostatic complementarity is a direct result of charge-optimization theory [69]. In the optimal ligand, the potential produced by the ligand must cancel the potential produced by the receptor when they are defined as follows [81]. As shown in Figure 4.1, the interaction potential in molecule A is defined as the potential produced by charging only molecule B. The desolvation potential is the bound- minus unbound-state potential produced by the ligand. If molecule A is the optimal ligand for molecule B, then the desolvation and interaction potentials sum to zero everywhere inside of molecule A.

Complexes may be evaluated according to how nearly they satisfy the above condition. We sample the desolvation and interaction potentials at points on the surface of the ligand, giving vectors of potentials $\mathbf{\Phi_D}$ and $\mathbf{\Phi_I}$ respectively. We define the complementarity metric as [68]:

$$m = \frac{(\mathbf{\Phi_D} + \mathbf{\Phi_I})^2}{|\mathbf{\Phi_D}|^2 + |\mathbf{\Phi_I}|^2} \tag{4.1}$$

This metric is normalized so that it does not depend explicitly on the number of points where the potential is sampled or on the absolute magnitudes of the vectors. The value of m can be between 0 and 2. $m = 0$ means that the ligand is optimal, since the potentials cancel exactly. $m = 1$ means that $\mathbf{\Phi_D}$ is much larger in magnitude than $\mathbf{\Phi_I}$ or that the vectors are perpendicular. $m = 2$ implies that $\mathbf{\Phi_D} = \mathbf{\Phi_I}$, so the ligand and receptor are not at all complementary. We may compare different complexes using this metric in order to predict which complexes are more likely to form.

4.3 Methods

Structure Preparation. Each of the structures with pdb codes 1emy [11], 1wla [87], 1myg [104], 1mbd [109], 2mm1 [64], 1myt [10], and 1mba [12] was obtained from the Protein Data Bank. Hydrogens were added using the HBUILD facility of the

program CHARMM [18, 19]. The B portion of myoglobin is defined as residues 20–97 in elephant myoglobin, and the A portion consists of the remaining residues (helices A, G, and H). In other myoglobins, the B portion consists of the residues that align with these residues in elephant myoglobin. Sequence alignment was performed using the program Clustalx [139].

Continuum Electrostatic Calculations. All electrostatic calculations were performed using a locally modified version of the program DELPHI [46, 47, 126]. The interior dielectric constant was set to 4 and the exterior dielectric constant was set to 80. The linearized Poisson–Boltzmann equation was solved with an ionic strength of 0.145 and a 2 Å Stern layer. Each calculation of the potential in the bound and unbound states was repeated with ten different translations of the grid with respect to the molecule. Each molecule was placed on a 129x129x129 grid with focusing so that the molecule filled 23% then 92% of the grid. All molecules were aligned and so that the same grid spacing of 2.3 grids/Å was used for every molecule. Calculations on the crystal structures in which the grid spacing was increased to 4.6 grids/Å changed the complementarity scores by less than 0.5%.

Both desolvation and interaction potentials were computed for the A and B portions of each myoglobin. The set of structures to be compared was aligned to minimize the distances between corresponding C_α atoms. Alignment was performed using the McLachlan algorithm [89] as implemented in the program ProFit (Martin, A. C. R., `http://www.biochem.ucl.ac.uk/ martin/programs/#profit`). Once the structures were aligned, the molecular surfaces of the A and B portions of the molecules were computed and mapped on to a grid. The surface points that fell inside all of the A portions were found, giving the surface of the region enclosed by all A portions. Any of these surface points that fell within 2.0 Å of an atom from the B portion were pushed back to avoid measuring especially strong potentials from bonded atoms. The final result of this procedure was a uniform sampling of points near the surface of the A portion that was enclosed by all A portions of myoglobin. (The points where the potential is sampled must be inside the ligand in order for the desolvation and interaction potentials to cancel in the optimal ligand [69].) This

was the set of points where the desolvation and interaction potentials were computed using DELPHI. An identical procedure was performed for the B portions.

Homology Modeling. The program MODELLER [118] was used to build homology models of each myoglobin as well as chimeric myoglobins consisting of the A portion of one species and the B portion of another species. The spring constant for the angle C–C–O(amide) in Asn and Gln side chains was increased from 15.0 kcal mol^{-1} rad^{-2} to 60.0 kcal mol^{-1} rad^{-2} in order to prevent these side chains from adopting highly unfavorable covalent geometries during the course of homology modeling.

4.4 Results and Discussion

Complementarity of Crystal Structures. Seven structures of myoglobin from different species were used for this study. The structures were from elephant (pdb code 1emy) [11], horse (1wla) [87], pig (1myg) [104], whale (1mbd) [109], human (2mm1) [64], tuna (1myt) [10], and a sea hare (*Aplysia Limacina*, pdb code 1mba) [12]. The first five of these are mammalian species, and the other two are less similar species. The percent identities of the myoglobin sequences are shown in Table 4.4. The mammalian sequences all share at least 80% sequence identity with one another, while they share 40–45% identity with tuna myoglobin, and 20–25% identity with sea hare myoglobin.

The structure of elephant myoglobin is shown in Figure 4.2. It folds by forming a core of its A, G, and H helices [44, 106], which is labelled the A portion of the molecule and is shown in yellow. The B portion consists of the remainder of the molecule and is shown in purple. This figure also shows the surfaces of elephant myoglobin with its desolvation potential projected on the surface. In the upper and lower right of Figure 4.2, the A portion is shown with its desolvation potential projected on the surface. The desolvation potential is a difference between the bound and unbound state, and because solvent screening is reduced upon binding, the potential is typically larger in magnitude in the bound state. Therefore, the desolvation potential is positive

(blue) near positive charges and negative (red) near negative charges.

The crystal structures were aligned with one another so that the RMSD of their C_α atoms was minimized. The desolvation and interaction potentials were computed in the A and B portions of each of the structures listed above. For each portion, the potentials were mapped on to a region enclosed by all the molecular surfaces. This procedure allowed comparison of homologous points within each protein, and controlled for slight variations in the shapes of the homologous proteins. The elephant myoglobin was used as a training set (see below) for the electrostatic potentials, and the other structures were compared to see if the correct A–B pairings more complementary to one another than incorrect pairings (for example, whale A with tuna B).

The desolvation potentials of the A portion of elephant myoglobin are plotted versus the interaction potentials in Figure 4.3. Each point in the plot represents a position on the surface of the A portion, and x and y coordinates of the point represent the desolvation and interaction potentials at the position. The black line in the figure shows where $\Phi_D = -\Phi_I$, which is where all the points in the plot would fall if portion A were an electrostatically optimal ligand for portion B in this model. It is apparent from this plot that the A portion of elephant myoglobin is not predicted to be a perfectly optimal "ligand" for its B portion. Nevertheless, we know that the complementarity is good enough for the protein to fold. Some of the points do not need to fall so close to the line in order to have good complementarity. We use this structure to train the comparison of other structures. Our measure of complementarity is computed in the other structures using only the points that have good complementarity in the elephant structure. The points with good complementarity are defined as satisfying the conditions

$$|\Phi_D + \Phi_I| \leq |\Phi_D| \tag{4.2}$$

$$|\Phi_D + \Phi_I| \leq |\Phi_I| \tag{4.3}$$

The points that satisfy these conditions fall between the purple lines in Figure 4.3.

These points account for approximately 38% of the surface of the A portion of elephant myoglobin. All of the selected points are at the interface between the A and B portions.

All possible pairings of the 6 A portions with the 6 B portions were evaluated by comparing the complementarity metrics for each pairing. The results are shown in Figure 4.4. In the top panel, the B portion is treated as the ligand to be optimized. The desolvation and interaction potentials inside of this portion were computed for each pairing. The first grouping of bars represents the results for each possible B ligand when the A receptor is fixed to be A1 (horse myoglobin). The first bar represents the complementarity of the interaction potential from charging portion A1 and the desolvation potential from charging B1 (the B portion of horse myoglobin. The second bar represents the complementarity of A1 and B2, and so on. The second grouping of bars represents each possible B portion binding to the A2 receptor. In the bottom panel, the definitions of ligand and receptor are reversed. The A portions of each hypothetical complex are treated as the ligand, and the desolvation and interaction potentials are computed in this portion. The A–B pairings that are actually observed (for example, A portion from horse with B portion from horse) are shown as red bars in the figure. The vertical axis is inverted in the plots so that the tallest bars have the best complementarity—i.e. the m value closest to zero.

Charge optimization theory gives the best ligand for a fixed receptor, so when comparing the complementarity results in Figure 4.4 it is appropriate to compare the bars in the same grouping. The correct A–B pair has the best or nearly the best computed complementarity in each of the groupings shown. In each of the cases where a correct pairing is outscored by an incorrect pairing, the incorrect pairing consists of two mammalian sequences. The mammalian sequences (1–4) tend to be conserved fairly well both in terms of their sequence and their structure. Hence in several cases, there is little difference between a correct A–B pairing of mammalian sequences and an incorrect pairing. For instance, in the bottom plot, B3–A1 has nearly the same score as B3–A3. With this minor exception, the complementarity metric picks the correct A–B pairing of myoglobin portions in each possible comparison.

Complementarity of Homology Models. The electrostatic complementarity metric was used to evaluate potential myoglobin A–B pairings given knowledge of just one structure and only the sequences of the remaining myoglobin A portions and B portions. Homology models of each possible complex were built using the elephant myoglobin structure and the sequences of the 6 A portions and the 6 B portions of the other known structures. This gave a total of 36 structures, 6 of which were models of known myoglobins, and 30 of which were chimeric myoglobins contaning mismatched A and B portions. Each of the complexes was evaluated by calculating the complementarity metric as described for the crystal structures.

The results are shown in Figure 4.5. As before, the top graph shows the results where portion B is treated as the ligand to be optimized, and the bottom graph results from treating portion A as the ligand. When the B portion is the ligand, the correct pairing typically scores higher than the incorrect pairing in each grouping. As with the calculations on the crystal structures, there are cases where a correct mammal–mammal pairing is outscored slightly by an incorrect mammal–mammal pairing. Overall, the scores vary somewhat from the scores of the crystal structure complexes. The differences are due to conformational rearrangements of the complexes, especially in the side chains, which have fewer restraints than the backbone in homology modeling [118]. When the A portion is considered as the ligand to be optimized, there is one case in which the conformational variations actually cause an incorrect pairing to be scored higher than a correct pairing. The A5–B4 (tuna-human chimera) complex scores higher than the A4–B4 (human myoglobin) complex (lower panel of Figure 4.5). This difference in complementarity arises from the accumulation of several errors in side chain positioning during homology modeling that are illustrated in Figures 4.6 and 4.7.

Figure 4.6 shows the desolvation potentials in the human myoglobin crystal structure, and in the homology models of human myoglobin and the tuna-human chimera. The structures are shown from the front just as in the bottom right of Figure 4.2 and from the side as in the top right of Figure 4.2. There are two regions of the surface where the difference in potentials tends to favor the incorrect pairing

over the correct pairing. The first set of changes is highlighted by a box in the top row of surfaces in Figure 4.6. In the human crystal structure, the carbonyl O of the side chain of Asn 145 points outward in the upper left corner of the box. Met 142 is oriented so its sulfur is on the surface. The negatively charged ends of these two polar groups are buried upon binding to the B portion, and they contribute to the red patch highlighted in the crystal structure. In the homology model of this sequence, Asn 145 rotates 180° about its χ_2 angle so that its amide NH_2 points toward the surface in the upper left of the box. Similarly, Met 142 reorients during homology modeling so that its sulfur is buried and the partially positive C_γ and C_ϵ are exposed. This results in the blue patch where the red patch was in the crystal structure. In the tuna-human chimera, Met 142 is mutated to a Leu. Although Asn 145 is in a similar orientation to the human homology model, this side chain is not buried to the extent that it is in the human homology model. Thus we see a smaller desolvation potential throughout the yellow box in the incorrect pairing. Complementarity is better with a hydrophobic residue than with an incorrectly oriented polar residue.

The second highlighted area of desolvation potential is seen in the side views of the three structures. In the crystal structure, a backbone carbonyl carbon is desolvated by hydrogen bonding with Lys 42 (in the B portion), producing the red spot shown. In the homology model of human myoglobin, this Lys side chain changes its conformation so that it is buried by two different side chains, causing the red spot to disappear. Lys 98 is not fully extended, and its partial burial by the B portion causes a more intense blue patch to appear on the left side of the yellow box in the figure. However, in the tuna-human homology model, the conformations of both Lys 42 and Lys 98 are closer to those of the crystal structure. As a result, the desolvation potentials in the incorrect pairing resemble the crystal structure more than in the incorrect pairing.

In Figure 4.7, the interaction potentials of the three A portions are shown. Recall that these potentials are produced by charging the atoms in the B portion of the molecule, even though only the A portion is shown. The boxed region again has different potentials because of side chain conformational changes. In both the human crystal structure and the tuna-human chimera, the Lys 87 amino group points away

from the interface and is solvent exposed. However, in the human homology model Lys 87 adopts a conformation that allows it to salt bridge with Glu 148 (left side of box). The amino N of Lys 87 is only 2.7 Å from the carboxylate O of Glu 148 in the homology model, but it is 6.2 Å away in the crystal structure. Lys 87 is also buried by contacting Leu 149. The result is a large blue patch in the homology model that is only partially compensated by the small red patch of desolvation potential produced by Glu 148 (Figure 4.6).

A second feature of the interaction potentials is shown in the bottom row of Figure 4.7. In both the human myoglobin crystal structure and the incorrectly paired homology model, Lys 42 forms a hydrogen bond with a backbone carbonyl O. In both cases, this results in a region of positive interaction potential on the left edge of the boxes in Figure 4.7, and a corresponding red patch of desolvation potential as already described in Figure 4.6. In the human homology model, however, this Lys also adopts a different conformation, in which it is buried by the side chains of Pro 100 and Tyr 103. This large region of blue (boxed in Figure 4.7) is not complemented by any desolvation potential from portion A, and therefore Lys 42 contributes to the lower complementarity of the correct human A–B pairing.

Each of the contributions to the higher score of the incorrect pairing results from errors in side chain placement due to homology modeling. Errors exist in other predicted structures as well—the example shown here was one in which the errors tended to accumulate to strongly favor the incorrect pairing. Some modifications have been made to the modeling procedure; for example, an electrostatic energy term can be added, or one portion of the molecule can be constrained so that it has the same conformation in all pairings. However, each of these modifications to the procedure have resulted in lower quality homology models (as judged by comparison to the known structures) and lower accuracy in choosing correct pairings by electrostatic complementarity.

Comparison to Other Metrics. We wished to compare the results for the electrostatic complementarity metric to the ability of other possible metrics to choose which A–B pairs are most suitable. First, we find a simple way to measure how

well each measure discriminates the correct pair from the other pairs. We define n as the number of possible comparisons of A–B pairs in which the correct pair scores higher than an incorrect pair. For example, consider the electrostatic complementarity results for the crystal structures. In each group of possible ligands, 5 comparisons can be made between the correct pair and each of the 5 incorrect pairs. We multiply by the number of possible fixed receptors (12) to find that there are a total of 60 possible comparisons between correct and incorrect pairings.

Several measures of complementarity were used on the homology models to determine which would be likely to pair up. The CHARMM energy function was used to compute the van der Waals interaction energy and a cheap electrostatic energy ($4r$ dielectric) of interaction between the two halves. Both the total surface area buried upon complexation and the buried hydrophobic surface area were also computed for each complex. In addition, the MODELLER objective function, which includes molecular mechanics energies as well as terms for how well the homology models fit their restraints, was tested as a measure of the quality of each possible complex.

We used a measure of shape complementarity to evaluate each complex. A number of methods have previously been put forward for evaluating shape complementarity [25, 31, 42, 50, 56]. Here, we sample a set of points on the surface of two portions and compute the following metric:

$$\frac{\sum_i^A \sum_j^B \exp(-r_{ij}^2/\alpha^2)}{\sqrt{[\sum_i^A \sum_j^A \exp(-r_{ij}^2/\alpha^2)][\sum_i^B \sum_j^B \exp(-r_{ij}^2/\alpha^2)]}}$$

This normalized metric has larger values for A and B surfaces that have points close to one another and hence are more complementary. The summations are over the surface points of portions A or B, as indicated; r_{ij} is the distance between points i and j. α is an adjustable parameter with units of Å that describes exactly how close the points from the surfaces must be in order to be considered complementary. We allow this parameter to vary from 0.1 Å to 2.0 Å in order to allow a slight extra bias to shape complementarity as a way of distinguishing correctly matched A–B pairs. The value of $\alpha = 0.5$ Å was determined to be the best value as judged by the number

of correct comparisons.

The data for the performance of shape complementarity as well as each of the other measures is shown in Table 4.2. For each measure, the number of correct comparisons n is shown. Since the mammalian sequences share a high degree of sequence similarity, we also show the quantity n_2, which is defined as the number of correct comparsions when mammal–mammal comparisons are ignored. There are a total of 36 possible comparisons of this sort. When electrostatic complementarity is used (as in Figure 4.5), there are 35 correct comparisons out of a possible 36. The one incorrect comparison between the tuna-human and human homology model has already been described. The other possible measures fare worse at discriminating the correct A–B pairs by this standard. The metrics in Table 4.2 are sorted according to their value of n_2 when applied to the myoglobin homology models. The charmm electrostatic energy does substantially worse than the electrostatic complementarity metric, with a n_2 value of 28. The other metrics shown in the table are hardly better than picking structures at random. Despite the fact that the shape complementarity metric included an adjustable parameter, it performs only slightly better than random choices. The poor performance does not necessarily indicate that molecular shape is unimportant in molecular recognition. It is more likely that such factors are important, but that the process of homology modelling does not provide high enough resolution to evaluate the structures in this way.

Conclusion. Electrostatic complementarity appears to be a useful descriptor for determining which pairs of proteins may be able to form a complex. When applied to crystal structures, the correctly paired portions always gave a higher complementarity score than incorrectly paired portions. When the measure was applied to homology models, errors in side chain placement led to identification of one incorrectly paired complex as better than its correctly paired counterpart. The electrostatic metric here does appear to perform substantially better at choosing potential binding partners than other readily available measures. Future improvements in modeling of protein structures may make this measure of electrostatic complementarity more useful.

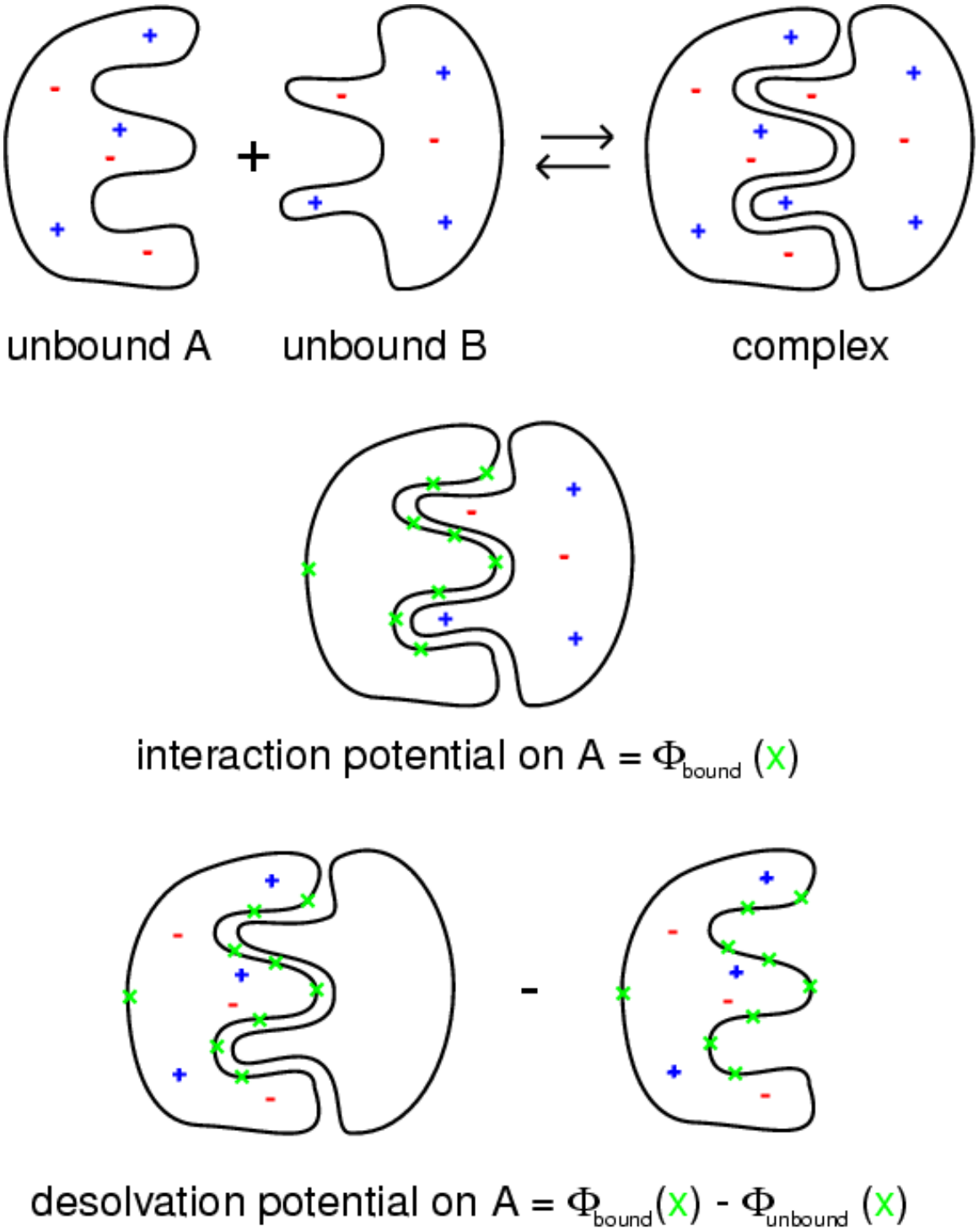
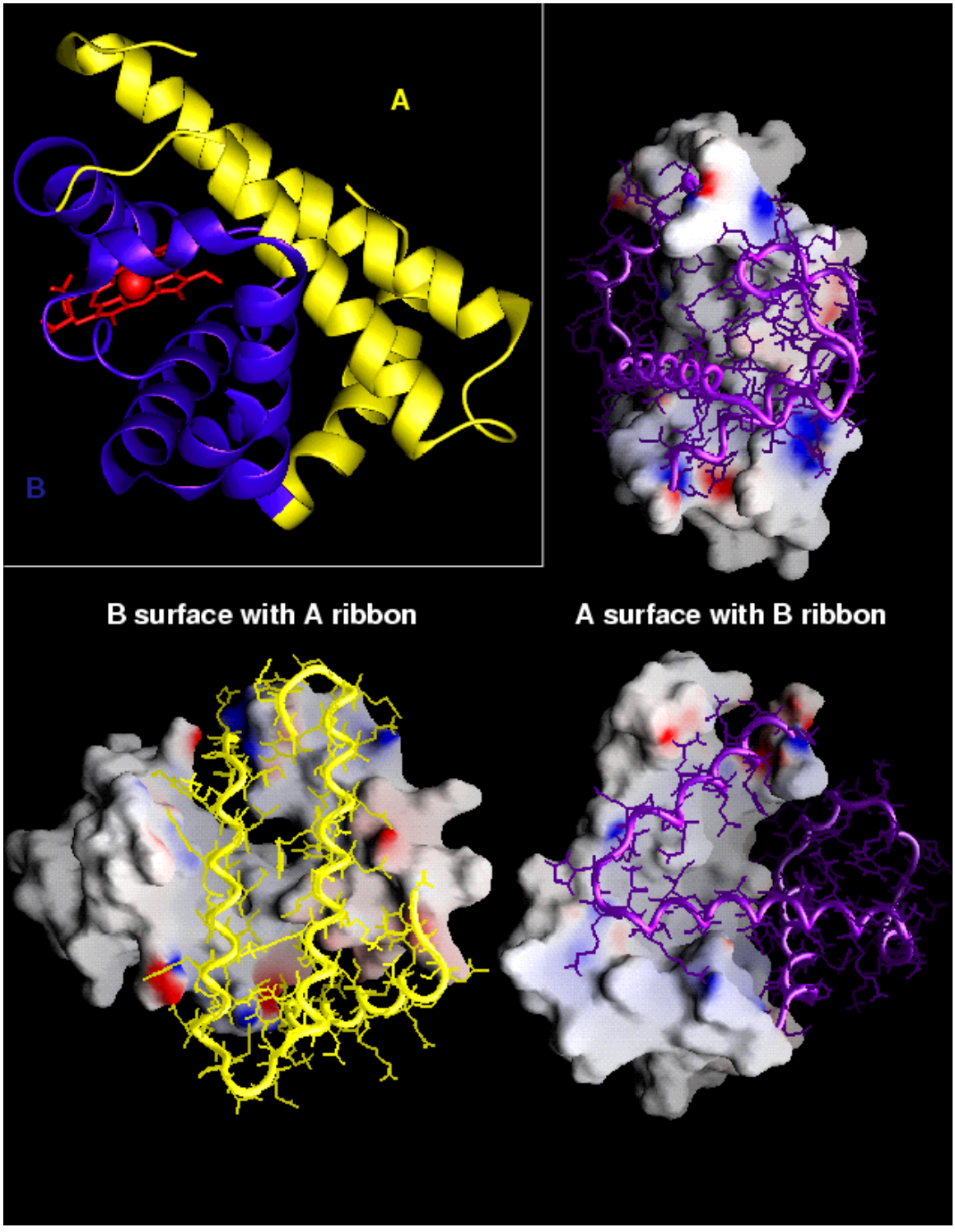


Figure 4.1: The definition of interaction potentials and desolvation potentials. When two molecules A and B form a complex (top line), their potentials are defined as shown. (middle) The interaction potential is the potential on molecule A that results from charging molecule B. (bottom) The desolvation potential is the difference between the bound and unbound state of the potential produced by charging molecule A.

Table 4.1: Percent identity of myoglobin sequences.

	elephant	horse	pig	whale	human	tuna	sea hare
elephant	–	87.7	85.2	81.3	84.5	41.3	23.4
horse		–	91.0	87.7	88.4	44.5	24.1
pig			–	86.5	93.5	43.9	23.4
whale				–	85.8	43.2	25.3
human					–	43.9	22.2
tuna						–	23.7
sea hare							–

Figure 4.2: (following page) The structure of elephant myoglobin. (upper left) The structure of elephant myoglobin with its A, G, and H helices (A portion) shown in yellow, and with the remainder of the protein (B portion) shown in purple. (upper right) The A portion shown as a molecular surface with interaction potentials projected on it, and the B portion shown as a ribbon with side chains in purple. On the surface, blue indicates positive potential, and red indicates negative potential. (lower right) Same representation as upper right, with the molecule rotated by 90°. (lower left) The B portion of myoglobin shown as molecular surface with interaction potentials projected on it, and the A portion shown as a yellow ribbon and side chains. Figure produced using the program GRASP [98] and the program MOLSCRIPT [73].


A
B
B surface with A ribbon
A surface with B ribbon

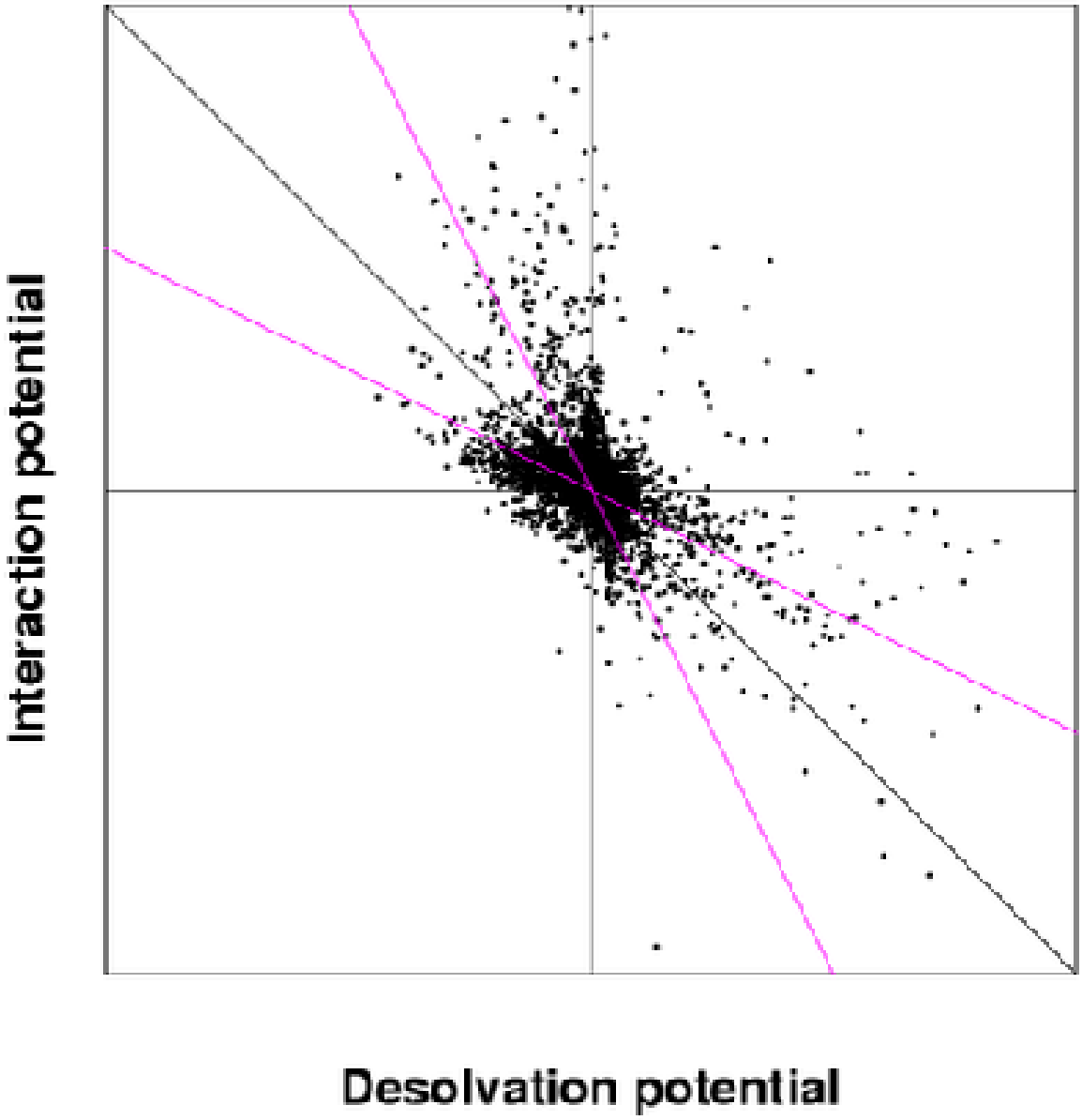


Figure 4.3: A plot of desolvation vs. interaction potential on the surface of the A portion of elephant myoglobin. The solid black line is $y = -x$, where the potentials cancel perfectly. The purple lines illustrate the points selected for calculating complementarity in the other structures.

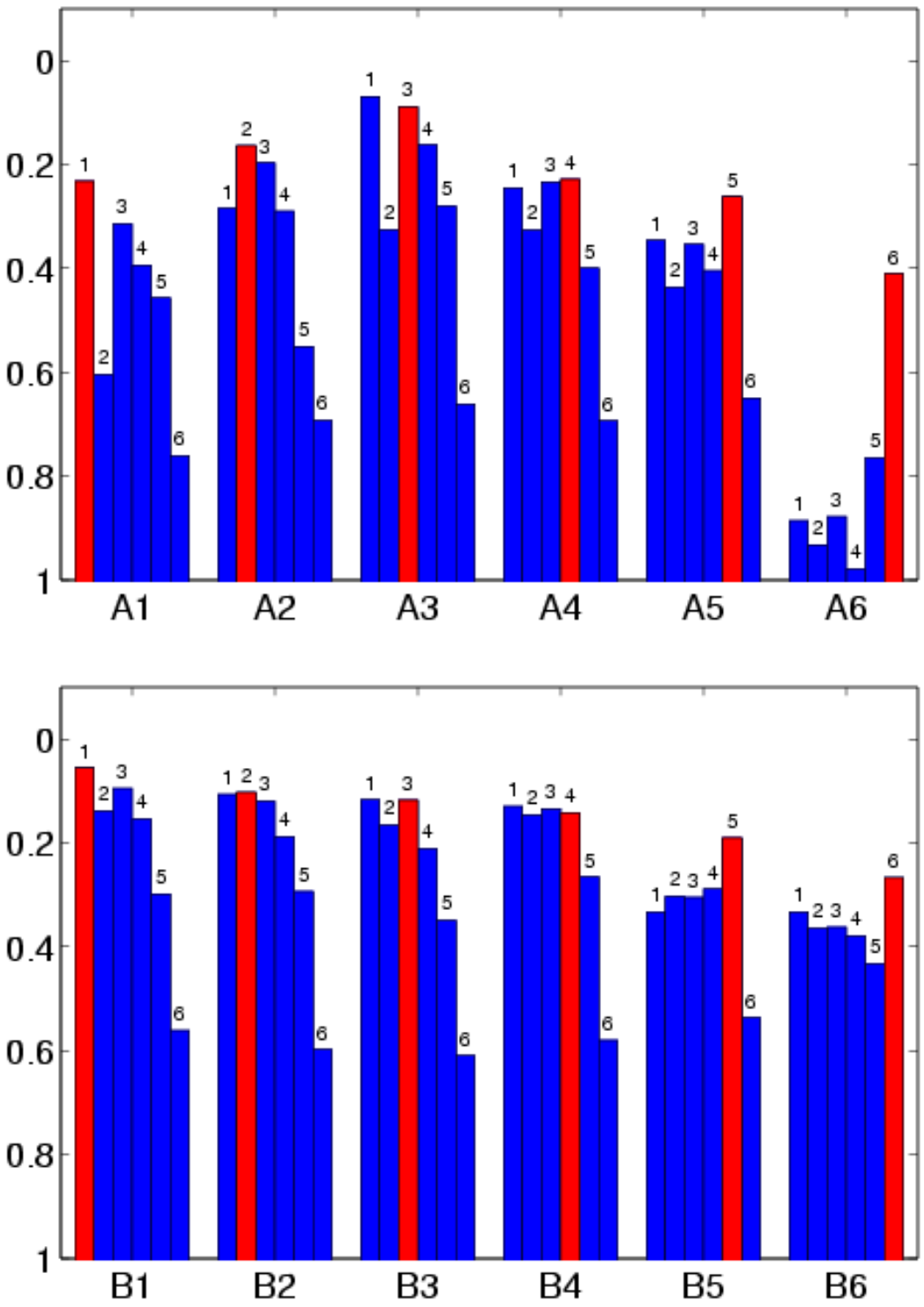


Figure 4.4: The electrostatic complementarity in myoglobin crystal structures. In the top graph, portion B is treated as the ligand, and the bars represent the complementarity score for the complex of the receptor (A1, A2, etc.) with the ligand. One group of bars represents all possible ligands for a given receptor. The numbers correspond to the different myoglobins as follows: 1 – horse, 2 – pig, 3 – whale, 4 – human, 5 – tuna, 6 – sea hare.

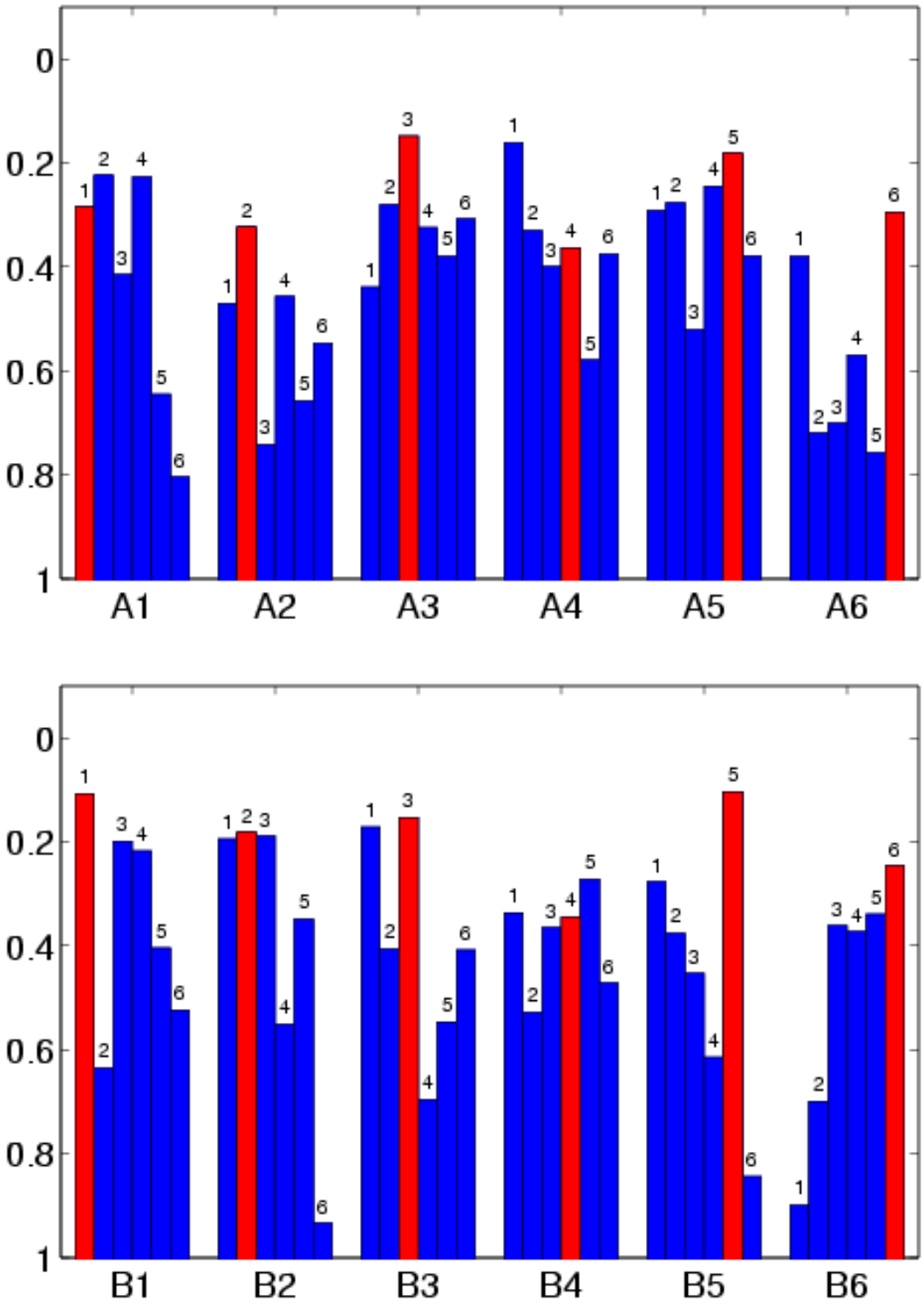


Figure 4.5: The electrostatic complementarity in myoglobin homology model structures. As in Figure 4.4, the numbers correspond to the following sequences: 1 – horse, 2 – pig, 3 – whale, 4 – human, 5 – tuna, 6 – sea hare.

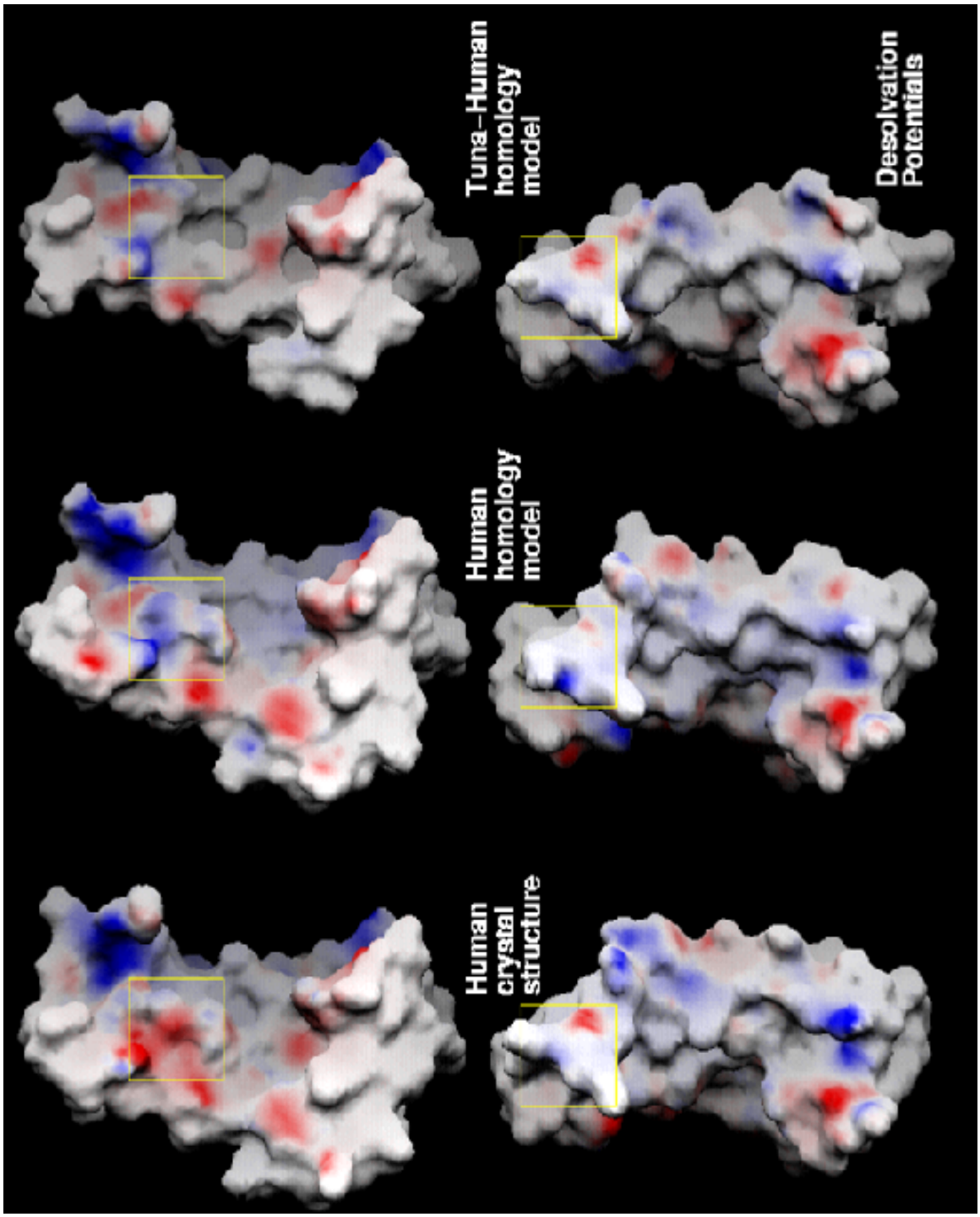


Figure 4.6: The desolvation potentials in the A portion of myoglobin in the human crystal structure, the human myoglobin homology model, and a model of a hypothetical tuna-human chimeric myoglobin. Red indicates indicates negative potential, and blue indicates positive potential. The three structures at the top are viewed from the same perspective as in the top right of Figure 4.2, and the three structures at the bottom are rotated 90°. Figure produced using the program GRASP [98].

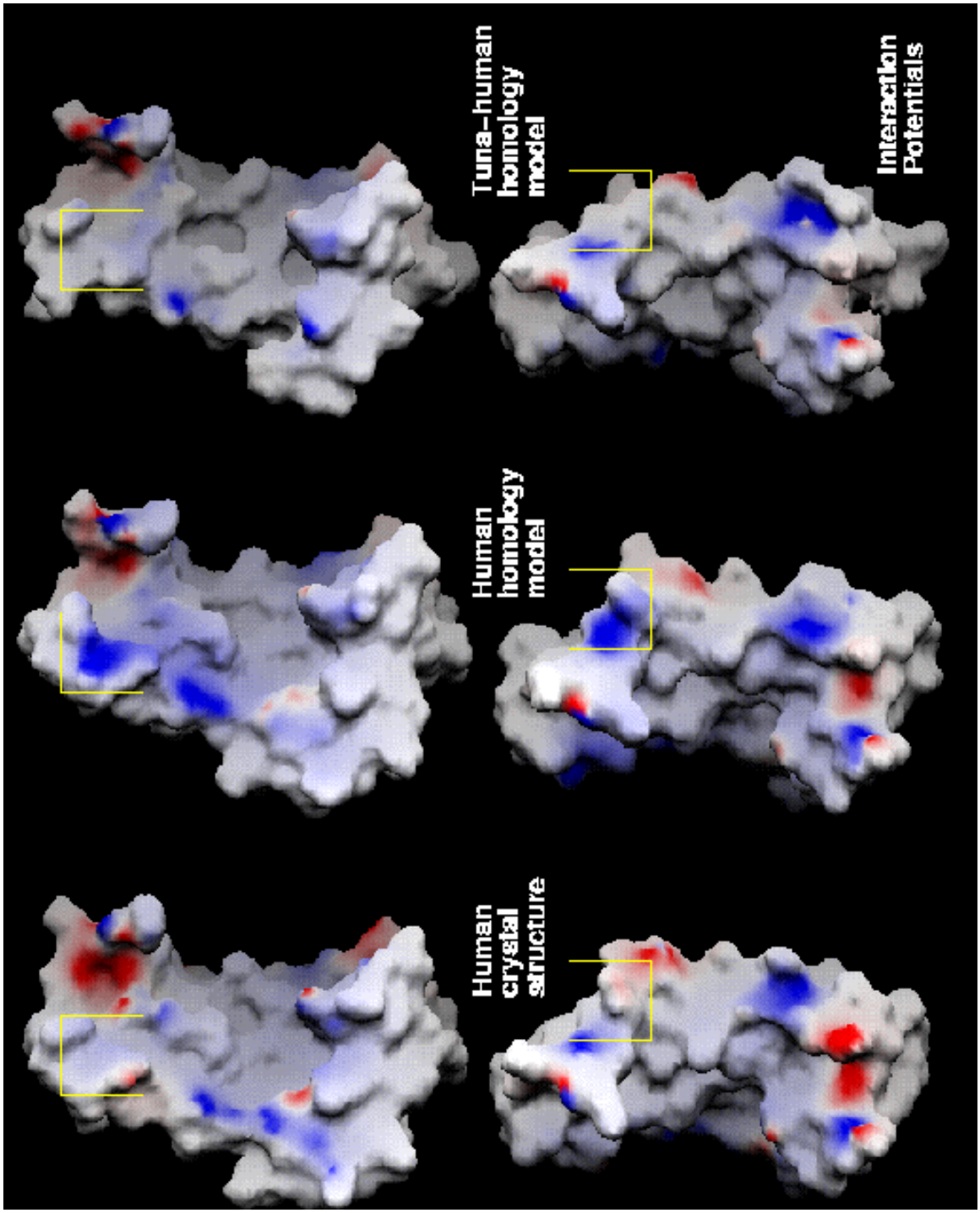


Figure 4.7: The interaction potentials in the A portion of the three myoglobins in Figure 4.6. Figure produced using the program GRASP [98].

Table 4.2: Evaluation of Possible Complementarity Measures.

Complementarity measure	n	n_2^a
Maximum possible	60	36
Electrostatic on crystal structures	56	36
Electrostatic complementarity	54	35
CHARMM electrostatic interaction[b]	51	28
Buried surface area	32	20
Shape complementarity	31	20
CHARMM total interaction energy[c]	36	19
Buried hydrophobic surface area	32	19
CHARMM van der Waals interaction	33	18
Random scoring[d]	30	18
MODELLER objective function	41	17

[a] n_2 is the number of correct comparisons when comparisons between two similar mammalian species are neglected.
[b] The energy of interaction was evaluated using CHARMM charges and an effective dielectric constant $\epsilon = 4r$ for each atom-pair interaction across the A–B interface.
[c] The total energy of interaction is simply the sum of the van der Waals and electrostatic interactions for each A–B pair.
[d] Random scoring simply means that each comparison has a 50% chance of being decided in favor of the correct A–B pair.

Chapter 5

Discrete Conformational Search Methods for Biomolecules

If the side chains of a protein are restricted to discrete conformational states, called rotamers, then the protein's conformations can be explored systematically. The systematic search algorithms of dead-end elimination (DEE) and A*, described here, have a couple of advantages over other types of search algorithms. First, they allow us to know in advance the range of motion that will be allowed in the search. Monte Carlo or simulated annealing methods have elements of randomness, and search results can depend on random choices, meaning that it is difficult to determine when a search is complete. Second, the systematic search algorithms allow us to examine many more states of the system. Recent efforts have demonstrated an efficient search of more than 10^{40} conformations of a protein [84, 128], a space which is much larger than that accessible by other search methods. This chapter includes a review of the DEE and A* search methods and describes some additions to the approach that allow a more efficient search of a more finely sampled conformational space.

5.1 Dead-End Elimination

Consider a protein consisting of a set of flexible amino acid side chains i, each in a rotamer state i_r, and a set of fixed atoms including the protein backbone. The energy

of the protein may be written

$$E_{total} = E_{fixed} + \sum_{i=1}^{p} E(i_r) + \sum_{i=1}^{p} \sum_{j=i+1}^{p} E(i_r, j_s) \tag{5.1}$$

where p is the number of variable residue positions, E_{fixed} is the energy of the fixed atoms, $E(i_r)$ is the self energy of rotamer i_r, including interactions within the rotamer and interactions between the rotamer and the fixed atoms, and $E(i_r, j_s)$ is the interaction energy between rotamers i_r and j_s. Here we assume that the energy of the system is pairwise additive, i.e. the interaction energy between two groups depends only on the positions of the two groups, and not on the placement of the other atoms in the system.

Even given the simplification of discrete rotamers and a fixed backbone, the search for low-energy conformations of the system can become intractable. If there are p variable side chains in the protein and each side chain has n rotamers, then there are n^p rotameric states of the system. In order to find the lowest energy state of the system, we need a search algorithm that will reduce the number of states that have to be evaluated. One such algorithm is called dead-end elimination (DEE), and was proposed in its original form by Desmet et al. [32]. Using this algorithm, it is best to compute the interaction energy between all pairs of rotamers in advance and store the energies for later use. Consider two rotamers of the same residue i_r and i_t. Rotamer i_r is not part of the global minimum energy conformation (GMEC) if

$$E(i_r) + \sum_{j \neq i} \min_s E(i_r, j_s) > E(i_t) + \sum_{j \neq i} \max_s E(i_t, j_s) \tag{5.2}$$

This condition says that if we arrange all the side chains so that they interact with i_r as favorably as possible, then arrange the side chains so that they interact with i_t as *un*favorably as possible, then the total interactions that i_r makes are still less favorable than those of i_t. Hence i_t is a better rotamer than i_r, and i_r can be eliminated. This condition may be evaluated quickly for the rotamers in the system since it only requires calculating the interactions between residue i and the other residues.

Goldstein [48] improved upon this idea by substituting the min and max operators in equation 5.2 with a single min operator:

$$E(i_r) - E(i_t) + \sum_{j \neq i} \min_s [E(i_r, j_s) - E(i_t, j_s)] > 0 \tag{5.3}$$

The Goldstein criterion says that if changing from rotamer i_t to i_r always leads to an increase in energy, then i_r can not be part of the GMEC. This criterion can be applied to all the pairs of rotamers at a particular position, and in many cases, rotamers will be eliminated, thus reducing the number of conformations of the system.

In equation 5.3, rotamer i_r is eliminated because i_t is always a better conformation, regardless of the way that the other residues are arranged. In some cases, there is no single rotamer that is always better than i_r, but it may be possible to eliminate i_r using multiple rotamers at position i. As described by Pierce et al. [110] and Looger and Hellinga [84], the conformational space of the system may be partitioned by choosing a position k (or multiple positions) and fixing it in turn to each of its rotamer states k_v. If, for every rotamer k_v, there exists a rotamer i_t such that

$$E(i_r) - E(i_t) + \sum_{j \neq i \neq k} \min_s [E(i_r, j_s) - E(i_t, j_s)] + [E(i_r, k_v) - E(i_t, k_v)] > 0 \tag{5.4}$$

then i_r can be eliminated. Using this criterion, the identity of the eliminating rotamer i_t may vary according to the identity of k_v. This simply means that in some arrangements of the system, one rotamer at position i is better than i_r, and in some other arrangements of the system, a different rotamer is better than i_r, but in all cases a better alternative to i_r exists.

It is relatively straightforward to extend any of the above criteria to elimination of pairs of rotamers. When applied to pairs, the basic DEE criterion becomes

$$E([i_r j_s]) + \sum_{k \neq j \neq i} \min_t E([i_r j_s], k_t) > E([i_u j_v]) + \sum_{k \neq j \neq i} \min_t E([i_u j_v], k_t) \tag{5.5}$$

In this equation, $E([i_r j_s])$ is the self energy of the pair of rotamers i_r and j_s, and $E([i_r j_s], k_t)$ is the interaction between that pair of rotamers and the rotamer k_t. If

the above criterion holds, then the pair of rotamers i_r and j_s can not be in the GMEC, although either one or the other of these rotamers may be present. The criterion for pairs that is analogous to the Goldstein criterion for single rotamers is

$$E([i_r j_s]) - E([i_u j_v]) + \sum_{k \neq j \neq i} \min_t [E([i_r j_s], k_t) - E([i_u j_v], k_t)] > 0 \qquad (5.6)$$

It is often desirable to obtain several low-energy structures, rather than the one structure that has the lowest energy. Each of the above criteria can be easily modified so that a rotamer (or rotamer pair) is eliminated if it can not occur within some energy cutoff E_{cut} of the minimum energy conformation. For example, equation 5.3 would become

$$E(i_r) - E(i_t) + \sum_{j \neq i} \min_s [E(i_r, j_s) - E(i_t, j_s)] > E_{cut} \qquad (5.7)$$

In the case where $E_{cut} = 0$, the above equation simply reduces to equation 5.3, and rotamer i_r is then eliminated if it is not part of the global minimum conformation.

There have been additional incremental improvements to the DEE procedure [51], but the time required to carry out DEE can be seen from the above equations. Again suppose there are p residue positions and n rotamers per position. We can create an $n \times n$ table in which each entry represents a possible assignment of i_r and i_t in the DEE criterion. There are $(n^2 - n)$ possible comparisons in this table. The most often-repeated part of the calculation is the evaluation of the min and max operators in equation 5.2. The same min and max values may be used repeatedly in each row or column of the table, and therefore the number of times the extrema are calculated is proportional to n rather than n^2. Since the DEE criterion is applied to every residue position, the number of extrema calculated is $O(np)$ ("of order" np). Calculating any particular min or max requires a search of n rotamers at each of the other $(p-1)$ positions in the protein. Therefore the time required for one evaluation of the criterion is $O(np)$. The total time required for an entire cycle of DEE using equation 5.2 is the product (number of applications of DEE criterion) x (time required for one application), which is $O(np) \times O(np) = O(n^2 p^2)$.

Application of the Goldstein criterion (equation 5.3) does not allow one calculation

to be applied to an entire row or column of the table. The min operator in equation 5.3 varies according to the row and column of our comparison table. Thus the number of times that the equation is applied using this criterion increases to $O(n^2p)$, and the overall cost of a complete DEE cycle is $O(n^3p^2)$. This information is summarized in Table 5.1.

Table 5.1: DEE run time dependence on the number of variable positions (p) and on the number of rotamers per position (n).

Criterion	no of min calculated	time to find a min	Total time
Simple DEE (eqn. 5.2)	np	np	n^2p^2
Goldstein DEE (eqn. 5.3)	n^2p	np	n^3p^2
Split DEE (eqn. 5.4)	n^2p	np	n^3p^2
Simple Pair DEE (eqn. 5.5)	n^2p^2	np	n^3p^3
Goldstein Pair DEE (eqn. 5.6)	n^4p^2	np	n^5p^3

For the conformational splitting (equation 5.4), there is still an $n \times n$ table of possible comparisons. This criteria adds the variation of the splitting position (residue k), which is allowed to be any variable residue in the protein, and must be placed in each of its n rotamer states for the criterion to be applied. The min values in equation 5.4 can be precomputed for each comparison of rotamers i_r and i_t [110]. The criterion can then be applied using all possible splitting positions k with the same time dependence as Goldstein DEE, $O(n^3p^2)$. In practice, split DEE takes longer than regular Goldstein DEE, but the time required by the two methods grows similarly as the number of residues/rotamers increases.

For DEE applied to rotamer pairs, similar reasoning applies. When we make a table for purposes of comparing pairs of rotamers, there are now n^2 rows and n^2 columns, since the number of possible rotamer pairs $[i_r j_s]$ is n^2. The comparison table is $n^2 \times n^2$, and the number of pairs of residue positions (i and j) where DEE may be attempted is $p(p-1)$. As shown in Table 5.1, the total time for simple pair DEE is $O(n^3p^3)$ and that for Goldstein pair DEE is $O(n^5p^3)$.

In the current implementation of DEE, the Goldstein singles criterion (equation 5.3) is applied first, since it is computationally inexpensive. It is applied repeatedly at every residue until no rotamers can be eliminated at any position. Then, the slightly more expensive split DEE criterion (equation 5.4) is applied at every position until it too is no longer effective. Finally, the Goldstein pairs criterion is applied until it converges. If the pairs criterion eliminates all the pairs that a rotamer i_r can make with some residue j, then i_r may be discarded [75]. If one or more rotamers are eliminated in this step, then the process continues at the beginning with application of singles DEE.

5.2 A* Search (Branch and Bound)

Even after application of the above DEE criteria, a large number of possible conformations of the system may remain. This is particularly true when a criterion with an energy cutoff is used, as this weakens each of the DEE criteria. A search algorithm called the A* algorithm [145] can be used to search the remaining conformations of the system [78]. The rotameric states of the protein may be represented as a tree, as shown in Figure 5.1. The first level of nodes in the tree represents the rotameric states of the first variable residue. If there are n rotamers at each position, then there will be n nodes on the first level. In addition, every node on the first level of the tree will have n branches from it, representing each of the rotamers of the second residue of the protein. The tree continues branching in this manner until the bottom level, which will have a node for each possible conformation of the system (n^p total nodes).

We again assume that the energy of the system is pairwise (as in equation 5.1). The energies may be computed and stored in advance as in DEE. The A* algorithm is a method for finding the optimal path from the root node to a goal node of a search tree. In this problem, the optimal path represents the lowest energy conformation of the system. The algorithm uses a function called $\mathbf{f}^*$ to evaluate the nodes of the tree,

where

$$\mathbf{f}^* = \mathbf{g}^* + \mathbf{h}^* \tag{5.8}$$

$\mathbf{g}^*$ is the lowest-cost path from the root to the current node, and $\mathbf{h}^*$ is a heuristic estimate for the cost of going from the current node to a goal node (a complete conformation of the protein). The sum $\mathbf{g}^* + \mathbf{h}^*$, defined as $\mathbf{f}^*$, represents an estimate of the total energy of the system given the set of rotamers that has been placed at the node in the tree. $\mathbf{g}^*$ is simply the energy of the residues that have been placed at the current point in the tree. (The residues at or above the current node in the tree have been placed; the others are still variable.) If we let p_f equal the number of variable positions that have been fixed at the current node in the search tree,

$$\mathbf{g}^* = E_{fixed} + \sum_{i=1}^{p_f} E(i_r) + \sum_{i=1}^{p_f} \sum_{j=i+1}^{p_f} E(i_r, j_s) \tag{5.9}$$

The $\mathbf{h}^*$ function is only an estimate of the cost of reaching a goal node. The algorithm requires that $\mathbf{h}^*$ always underestimates the cost of reaching a goal node from the current node. At the goal nodes, where all residues have been placed, $\mathbf{g}^*$ is equal to the total energy of the conformation, and $\mathbf{h}^*$ is zero.

In the basic version of A*, a list of nodes is stored and sorted according to the value of $\mathbf{f}^*$. One step of the algorithm consists of taking the node with the lowest value of $\mathbf{f}^*$ and expanding it—i.e. finding the values of $\mathbf{f}^*$ for all the nodes immediately below that node in the tree. The newly expanded nodes are then placed in the sorted list of nodes, and the next-lowest $\mathbf{f}^*$ node is expanded. (The level of the node in the tree is unimportant—values of $\mathbf{f}^*$ at different levels of the tree are compared.) The process continues until the node with the lowest value of $\mathbf{f}^*$ is a goal node. This node represents the lowest energy conformation of the protein, and its $\mathbf{f}^*$ value represents the energy of the conformation. The first goal node is guaranteed to be lowest in energy because $\mathbf{h}^*$ for all the other nodes underestimates their energies. Thus the energy of the first goal node is less than any energy that could be found by following any of the unexpanded nodes.

It is difficult to predict in advance how much time the A* algorithm will take to

find a solution. In the worst case, it will have to visit every node in the tree, requiring $O(n^p)$ time, just as if all the structures were enumerated. In the best case, A* will follow a direct path down the tree to the correct structure, and the number of nodes visited will be $O(np)$. A typical case is between these two extremes. The success of the algorithm depends on the quality of $\mathbf{h}^*$, the estimate of the minimum energy required to complete the conformation from the current node. $\mathbf{h}^*$ is required to underestimate the energy, but higher values of $\mathbf{h}^*$ make it more likely that the algorithm will ignore the unproductive branches of the search tree. It is also important to be able to calculate $\mathbf{h}^*$ fairly rapidly, since it must be determined at each node visited. Two methods for calculating the $\mathbf{h}^*$ bound have been used for the type of problem being considered here. The first was proposed by Leach and Lemon [78]. Again let p_f equal the number of variable side chains that have been placed at the current location in the search tree. In the notation used here, their value of $\mathbf{h}^*$ is given by

$$\mathbf{h}^* = \sum_{j=p_f+1}^{p} \min_s \left[E(j_s) + \sum_{i=1}^{p_f} E(i_r, j_s) + \sum_{k=j+1}^{p} \min_t E(j_s, k_t) \right] \tag{5.10}$$

In this equation, the first term inside the square brackets represents the self energies of positions that are still variable. The second term represents interactions between the positions that are still variable and the fixed residues, and the third term represents interactions among the variable positions. The min operators ensure that the value of $\mathbf{h}^*$ is no greater than any energy that could possibly result from the placement of the variable side chains. This $\mathbf{h}^*$ function can be computed quickly enough to make the search efficient. Since the last term inside the brackets may be computed for each j_s and stored ahead of time, the $\mathbf{h}^*$ calculation scales as $O(np)$ at each node.

A second way of computing $\mathbf{h}^*$ was used by Gordon and Mayo [52] in an algorithm that is similar to A*. The Gordon and Mayo bound differs from the bound in equation 5.10 in two relatively minor ways. The first difference is that a new definition for the energy is introduced, defined as E_{pair}:

$$E_{pair}(i_r) \;=\; 0$$

$$E_{pair}(i_r, j_s) \;=\; \frac{E(i_r) + E(j_s)}{p-1} + E(i_r, j_s)$$

Here the self energy of each rotamer is simply divided up and put into pair interaction terms. Using this definition, the total E_{pair} for any conformation is exactly equal to the total E for that conformation. This is just a new way of accounting for the energy. When substituted into equation 5.10, this gives

$$\mathbf{h}^* = \sum_{j=p_f+1}^{p} \min_s \left[\sum_{i=1}^{p_f} E_{pair}(i_r, j_s) + \sum_{k=j+1}^{p} \min_t E_{pair}(j_s, k_t) \right] \tag{5.11}$$

Because the min operators are evaluated with respect to self and pair terms simultaneously, the bound computed from equation 5.11 is almost always higher—and therefore better—than the bound calculated using equation 5.10. Tests performed with the two bounds show that the use of equation 5.11 consistently results in a faster search than the use of equation 5.10.

The second difference in the Gordon and Mayo bound can be seen by careful inspection of the last term in the above equation. The residue k varies from $j+1$ to the last residue p. This implies that there is an order of placement in the yet unplaced side chains of the system. The value of $\mathbf{h}^*$ depends on the order in which the residues j and k are summed. The Gordon and Mayo bound avoids choosing an order for these residues by calculating the energy a little bit differently. First note that the total energy of the system can be rewritten as

$$\begin{aligned} E_{total} &= E_{fixed} + \sum_{i=1}^{p} E(i_r) + \sum_{i=1}^{p} \sum_{j=i+1}^{p} E(i_r, j_s) \\ &= E_{fixed} + \sum_{i=1}^{p} E(i_r) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j \neq i}^{p} E(i_r, j_s) \end{aligned}$$

The factor of $1/2$ makes up for the double-counting of the pair interactions in the last term of the above equation. If we rewrite the interactions between variable residues

in equation 5.11 in this manner, we obtain the Gordon and Mayo bound:

$$\mathbf{h}^* = \sum_{j=p_f+1}^{p} \min_s \left[\sum_{i=1}^{p_f} E_{pair}(i_r, j_s) + \frac{1}{2} \sum_{k \neq j}^{p} \min_t E_{pair}(j_s, k_t) \right] \tag{5.12}$$

This equation for the bound does not depend on the ordering of the j and k residues. Some test cases have been run using equation 5.11 as a bound with all possible orderings of the residues j and using equation 5.12 as an alternative. Equation 5.12 performed approximately as well as the average of the possible bounds from equation 5.11. There is a heuristic method for calculating the order of residues j that was used by Leach and Lemon [78] to improve the performance of the algorithm. For each rotamer the following quantity is calculated

$$V(i_r) = E(i_r) + \sum_{j \neq i}^{p} \min_s E(i_r, j_s) \tag{5.13}$$

The two lowest values of $V(i_r)$ for each residue position are identified and their difference computed. The residue with the largest difference is expanded first, followed by the residue with the second largest difference, *etc.* Use of this ordering appears to result in better A* performance with the bound in equation 5.11. This bound was used throughout this work. However, it remains difficult to predict which of the two bounds performs better for any given system.

Depth-First A* Search. After the A* algorithm finds the best solution, it can continue expanding nodes as described above to find the next-lowest energy solution, then the next-lowest, *etc.* In principle, this could continue indefinitely, and it could, for example, find all solutions within 10 kcal/mol of the minimum energy conformation. However, in practice, the array of nodes often becomes too large for available memory before all solutions can be found.

There is a way to complete the search using very little memory once the minimum energy conformation is known. The depth-first A* search proceeds by simply doing a depth-first traversal of the search tree. A depth-first traversal simply means starting at the root, we follow the left-most branch down to its leaf, then take one step up,

follow each of that nodes branches to leaves, *etc.*, until we have visited every node of the tree. A complete traversal would be extremely time-consuming, but depth-first A* allows us to skip large parts of the search by examining the value of $\mathbf{f}^*$ at each node. As the depth-first traversal proceeds, each node's $\mathbf{f}^*$ value is compared to f_{limit}. If at any node $f* > f_{limit}$, then the traversal does not proceed downward from that node, but continues searching other parts of the tree. If the minimum energy of the system is X and we wish to find all solutions within Y kcal/mol of the minimum, we simply set f_{limit} to $X + Y$. If f_{limit} can be set appropriately in advance, then depth-first A* will visit exactly the same nodes as A* would to find the correct set of solutions, but the depth-first search will use virtually no memory. (It essentially only remembers the current node.)

We now have established a procedure for finding all rotamer states within Y kcal/mol of the minimum energy conformation. First, the suite of DEE criteria is used with $E_{cut} = 0$. Then, basic A* is used to identify the global minimum energy conformation since a large number of possible conformations may remain after DEE. Next, DEE is repeated from the beginning with $E_{cut} = Y$, and depth-first A* is run with $f_{limit} = X + Y$ (where X is the minimum energy of the system).

5.3 Additional Enhancements

Flexible rotamers. The use of discrete rotamer searches that have been described to this point can have limitations. Foremost among the limitations is that there is no possiblity of the system making a slight adjustment to relieve the strain caused by a clash between two rotamers or between a rotamer and the fixed atoms. The rotamer(s) must make a coarse change to a different discrete state, which can often effect the state of the rest of the system when a fine adjustment would have worked better.

The typical rotamer library has χ angles in the neighborhood of $+60°$, $-60°$, and $180°$ for rotatable bonds [34]. In order to allow side chains to make finer adjustments, one may simply add rotamers that have χ adjustments of $\pm 10°$ or $15°$ at each χ angle.

However, this leads to a large increase in the size of the search space. For example, a leucine side chain has 2 rotatable bonds, and in the Dunbrack & Karplus library, it has $3^2 = 9$ rotamers. Adding variations of $\pm 10^\circ$ to χ_1 and χ_2 increases the number of rotamers by a factor of 9. There essentially were 3 possible values of χ_1 and χ_2, and after the fine adjustments are added, there are 9 values of each angle. When these adjustments are added to all the side chains in the system, the total number of states increases dramatically, and the search can take much longer. In addition, there are many states that only differ by small changes in a few atomic positions. Many similar low-energy conformations often exist, and since these conformations are all reasonable solutions, the search algorithms must enumerate them separately. This process further increases the time required for the search.

The problems caused by adding fine adjustments to the rotamer library can be addressed by using the flexible rotamer model suggested by Mendes et al. [90]. In this model, a rotamer and all its fine adjustments are grouped together in a flexible rotamer, or "fleximer." We denote this grouping of a set of rotamers $\{i_r\}$ with the symbol $i_{\boldsymbol{\mathcal{R}}}$. The individual rotamers in the set are referred to as subrotamers or rigid rotamers. The aim is to specify the state of the system in terms of the fleximer state at each position. In order to find the fleximer state of the system, we must have a way of evaluating the interactions between fleximers. We wish to have a function F of the fleximer state such that the free energy of the system is approximated by

$$F_{total} = E_{fixed} + \sum_i F(i_{\boldsymbol{\mathcal{R}}}) + \sum_i \sum_{j>i} F(i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}}) \tag{5.14}$$

where $F(i_{\boldsymbol{\mathcal{R}}})$ is the contribution of a single fleximer to the fleximer energy of the system, and $F(i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}})$ is the contribution of the pair of fleximers to the energy of the system. Mendes et al. develop an approximation for $p(i_r|i_{\boldsymbol{\mathcal{R}}})$, the probability that a particular subrotamer is occupied given that the residue is in a particular fleximer state. This probability actually depends upon the positions of all the other rotamers in the system and determining its value would require an exhaustive search of the states of the system. Instead, since all the subrotamers are fairly closely spaced, we

may assume that interactions with the rest of the system are approximately constant. In this approximation, the probability depends only on the self energy:

$$p(i_r|i_{\boldsymbol{\mathcal{R}}}) \propto \exp\left[-\frac{E(i_r)}{RT}\right]$$

The right hand side of this equation is a Boltzmann factor where T is equal to the temperature and R is Boltzmann's constant. Similarly, the joint probability that two rigid rotamers are in a particular state depends on the total energy of the pair:

$$p(i_r, j_s|i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}}) \propto \exp\left[-\frac{E(i_r) + E(j_s) + E(i_r, j_s)}{RT}\right]$$

Given this set of probabilities, we can determine the free energy of a fleximer isolated from the variable parts of the system, $A(i_{\boldsymbol{\mathcal{R}}})$, and the free energy of an isolated rotamer pair, $A(i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}})$,

$$A(i_{\boldsymbol{\mathcal{R}}}) = -RT \ln \sum_{r \in \boldsymbol{\mathcal{R}}} \exp\left[-\frac{E(i_r)}{RT}\right] \tag{5.15}$$

$$A(i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}}) = -RT \ln \sum_{r \in \boldsymbol{\mathcal{R}}} \sum_{s \in \boldsymbol{\mathcal{S}}} \exp\left[-\frac{E(i_r) + E(j_s) + E(i_r, j_s)}{RT}\right] \tag{5.16}$$

In order to efficiently implement the search for the state with the best F_{total}, we wish to keep the expression for F_{total} pairwise additive, as shown in equation 5.14. Following the example of Mendes et al., we substitute

$$\begin{aligned} F(i_{\boldsymbol{\mathcal{R}}}) &= A(i_{\boldsymbol{\mathcal{R}}}) \\ F(i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}}) &= A(i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}}) - A(i_{\boldsymbol{\mathcal{R}}}) - A(i_{\boldsymbol{\mathcal{S}}}) \end{aligned}$$

The contribution of individual fleximers to the F_{total} is simply the free energy of the isolated fleximer. When we consider the interaction of two fleximers, their contribution to F_{total} is the *change* in their total free energy caused by bringing them in proximity to one another. In principle, this approach could be extended to include higher order terms, but this pairwise decomposition of the free energy of the

system allows us to use DEE and A* as described for rigid rotamers.

Each fleximer has an entropy reward built into equations 5.15 and 5.16 if it is able to make small adjustments in its position. We wish to use flexible rotamers to make fine adjustments to the system, so that low energy structures can be found. In order to aid in finding low energy structures, we set $T = 0$ in equations 5.15 and 5.16, yielding

$$A(i_{\boldsymbol{\mathcal{R}}}) = \min_{r \in \boldsymbol{\mathcal{R}}} E(i_r) \tag{5.17}$$

$$A(i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}}) = \min_{\substack{r \in \boldsymbol{\mathcal{R}} \\ s \in \boldsymbol{\mathcal{S}}}} [E(i_r) + E(j_s) + E(i_r, j_s)] \tag{5.18}$$

These equations effectively neglect the entropy contribution from fleximers that can adopt many slightly different conformations.

Freezing Flexible Rotamers. The use of fleximers allows adjustments to discrete rotamers to be made and solutions to be found much more quickly than a search over all the individual subrotamer conformations. The drawback of the fleximer approach is that the free energy F_{total} is not strictly pairwise additive. The pairwise decomposition of Mendes et al. can give conformations that are predicted to be low in energy but actually do not have accessible low-energy states. For instance, suppose that some residue 1 has only one subrotamer (subrotamer a, shown in red in Figure 5.2) that can interact favorably with residue 2. Subrotamer a's presence will result in a favorable contribution to F. But suppose that residue 1 must adopt a different subrotamer (subrotamer b, shown in green in Figure 5.2) to interact favorably with residue 3. Again, there will be a favorable contribution to F. The difficulty is that the two favorable contributions can not be realized, since side chain A can not be in two places at once. This effect often results in a value of F which is substantially less than any physically meaningful value of the energy of the system.

There is a fairly straightforward way to address this problem. Once a fleximer state with a low value of F is identified, the fleximers may be "frozen" so that each one adopts only one of its subrotamer conformations. The freezing is simply another DEE/A* search with a fairly small search space. It proceeds fairly rapidly, so thousands of possible solutions may be frozen in a reasonable amount of time.

Each of the conformations is then reordered according to the energy E of the frozen conformation. E is typically greater than F for its fleximer conformation's counterpart because freezing effectively prevents any side chain from being in two places at once. This procedure helps ensure that conformations with a low F are in fact low energy conformations.

Alternative Method of Computing Fleximer Energies. One advantage of using a search of discrete conformations is that, using DEE/A*, the conformations found are guaranteed to be the lowest energy states available. When fleximers are added, one must be careful that low-F states actually correspond to low energy structures. Suppose we search all fleximer states with an F less than some threshold—say Y kcal/mol—and find that there are some frozen structures with an energy better than this threshold. If it were true that a fleximer conformation's E can never be lower than F, then we would be guaranteed to have the lowest E conformation among those already searched. Any other conformations with F above the threshold would necessarily have a frozen E above the threshold. Thus we would like $F < E$. Although this is true in many cases, a few cases arise in which $F > E$.

To illustrate the cause of this problem, let us consider a case in which $F > E$. Suppose there are three variable residues in the system, each with several subrotamers, as shown in Figure 5.3. For each fleximer, the minimum self energy occurs in the blue rigid rotamer (number 1), and this minimum self energy is -5 (arbitrary units). For residue $i = 1,\ 2,$ or 3,

$$\begin{aligned} F(i_{\boldsymbol{\mathcal{R}}}) &= A(i_{\boldsymbol{\mathcal{R}}}) \\ &= \min_{r \in \boldsymbol{\mathcal{R}}} E(i_r) \\ &= E(i_1) = -5 \end{aligned}$$

Now suppose that rotamer 1_2 (i.e. residue 1, rotamer 2—shown in red in Figure 5.3) has favorable interactions with rotamers 2_2 and 3_2 (also in red) that are each worth -8. Suppose also that the self energies of all red rotamers are 0, and all other interactions are worth 0. When we consider residues 1 and 3 as a pair (Figure 5.3B),

it is better to have them both as blue rotamers (total energy -10) than both as red rotamers ($0 + 0 + (-8) = -8$ energy). The state with both blue rotamers is likewise better than any other arrangement of the two residues. For the pairwise F values, we have

$$\begin{aligned} F(1_{\boldsymbol{\mathcal{R}}}, 3_{\boldsymbol{\mathcal{S}}}) &= A(1_{\boldsymbol{\mathcal{R}}}, 3_{\boldsymbol{\mathcal{S}}}) - A(1_{\boldsymbol{\mathcal{R}}}) - A(3_{\boldsymbol{\mathcal{S}}}) \\ &= \min_{\substack{r \in \boldsymbol{\mathcal{R}} \\ s \in \boldsymbol{\mathcal{S}}}} [E(i_r) + E(j_s) + E(i_r, j_s)] - (-5) - (-5) \\ &= 0 \end{aligned}$$

Similarly, the pair F interaction between residues 1 and 2 (Figure 5.3C) will give an energy of 0, for an F_{total} of -15 for the three residues. However, we can see that simply setting all positions to red rotamers would give an energy of 0 for the self energies, and two interactions worth -8 for a total E of -16, which is lower than the observed value of F. This frozen E is lower than the corresponding F because of the way that one pair of residues is considered at a time. When residue 1 is considered with rotamer 3, both have to pay an energy penalty of 5 to get -8 in interaction, and thus the penalty appears to be too much to be worth paying. When residues 1 and 2 are considered, again they each have to pay a penalty of 5 to get -8 in interaction, and again the penalty is not worth paying. The problem with this approach is that the penalty paid by residue 1 has been double-counted. To get promoted from rotamer 1 to rotamer 2, it only needs to pay the penalty once. It may be worthwhile to produce a new definition of F, thereby guaranteeing there can be no conformations with a high approximate F but a low actual E.

We therefore focus on the extra energy that each subrotamer needs to be promoted:

$$X(i_r) = E(i_r) - \min_s E(i_s) \tag{5.19}$$

This is the extra energy that was double-counted in the example above. We define a new energy of each subrotamer E_2 so that this excess energy is distributed over the

interactions that residue i makes with the other residues.

$$E_2(i_r) = \min_s E(i_s) \tag{5.20}$$

$$E_2(i_r, j_s) = E(i_r, j_s) + [X(i_r)W(i_r, j)] + [X(j_s)W(j_s, i)] \tag{5.21}$$

$W(i_r, j)$ is a weighting factor for the excess energy when considering rotamer i_r's interactions with residue j. Setting $W(i_r, j) = 1$ and calculating F as before will give exactly the same results as the original definition of E. In order to avoid double-counting the extra energy $X(i_r)$, we need to satisfy the condition

$$\sum_{j \neq i}^{p} W(i_r, j) = 1 \quad \text{for all } i_r$$

The simplest way to satisfy this condition is to divide the extra energy of each rotamer equally among the other residues, so that $W(i_r, j) = 1/(p-1)$. Although this guarantees that $F < E$, it may give values of F that greatly underestimate the energy. With this definition, if a residue needs to change subrotamer states to make a better interaction, it only needs to pay a small fraction $(1/[p-1])$ of the true penalty in self energy required to change rotamers. We may modify this approximation so that positions that interact more strongly with any given rotamer get a larger share of the rotamer's extra energy. A simple way of measuring how strongly a position interacts with a rotamer is the range of energies of interaction. Defining the range

$$R(i_r, j) = \max_s E(i_r, j_s) - \min_s E(i_r, j_s)$$

we may use the following weighting

$$W(i_r, j) = \frac{R(i_r, j)}{\sum_j R(i_r, j)} \tag{5.22}$$

The structure of switch Arc was used as a test of the method described above. For the NMR minimized average structure of switch, the conformation with minimum F was found to have $F = -97.0$ kcal/mol. When F was recomputed with the

above definition of weights, the structure with minimum F dropped to -107.6. The structure with the actual minimum energy had an energy $E = -90.5$. Although the second definition guaranteed that $F < E$, and that no low energy structures would be missed, its larger underestimation of the true energy meant that many more fleximer states had to be frozen in order to identify states with low true energies. In practice, it is therefore more efficient to use the original definition of F (weight $W = 1$) for most purposes. The strict lower bound is useful for simply verifying that no good structures are missed.

Divide and Conquer. In a few cases, the application of the DEE/A* does not eliminate enough rotamers, and the search becomes very slow. In many of these cases, an approach that has proven effective is one we term "divide and conquer." We select a residue i and fix it to a particular rotamer i_r. We then execute the DEE/A* procedure with this residue fixed, and record the solution. The search continues by fixing residue i to each of its available rotamer states, creating several smaller subproblems whose best energies can each be recorded. The overall solution is simply the best answer from all of the subproblems.

One advantage of this approach is that it allows each separate search with different fixed i_r's to be sent to a different processor—the problem is easily parallellized. Another advantage is that in each subproblem, DEE often eliminates more residues at other positions than when all i_r rotamers are available. This means that the total time spent on all the subsearches can be less than the time that would be spent on the original problem. In different searches, divide and conquer has yielded from 1.5–50-fold faster searches. The method appears to work best when the divided residue i can interact with several other positions. One can simply choose a residue by finding a position that is centrally located, or by defining a heuristic. A useful heuristic appears to be summing up the range of interactions that a residue can make with all other residues. The residue that can interact strongly with the most other residues is the best one to fix, since fixing it has more of a chance of constraining other residues, thus simplifying the search.

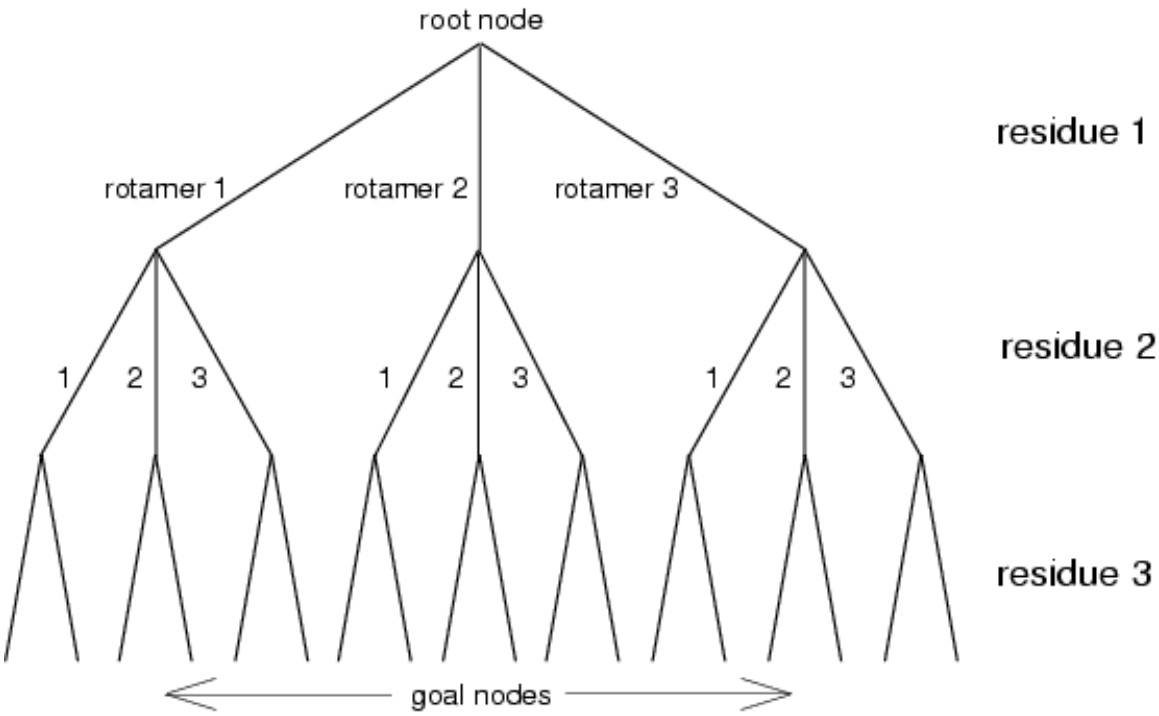


Figure 5.1: The tree representation of conformational search space. This figure is adapted from reference [78]. The top set of branches represents the possible rotamers of residue 1. The next set branches represents the rotamers for residue 2, and so on to the bottom of the tree, where there is one goal node for each possible conformation of the system.

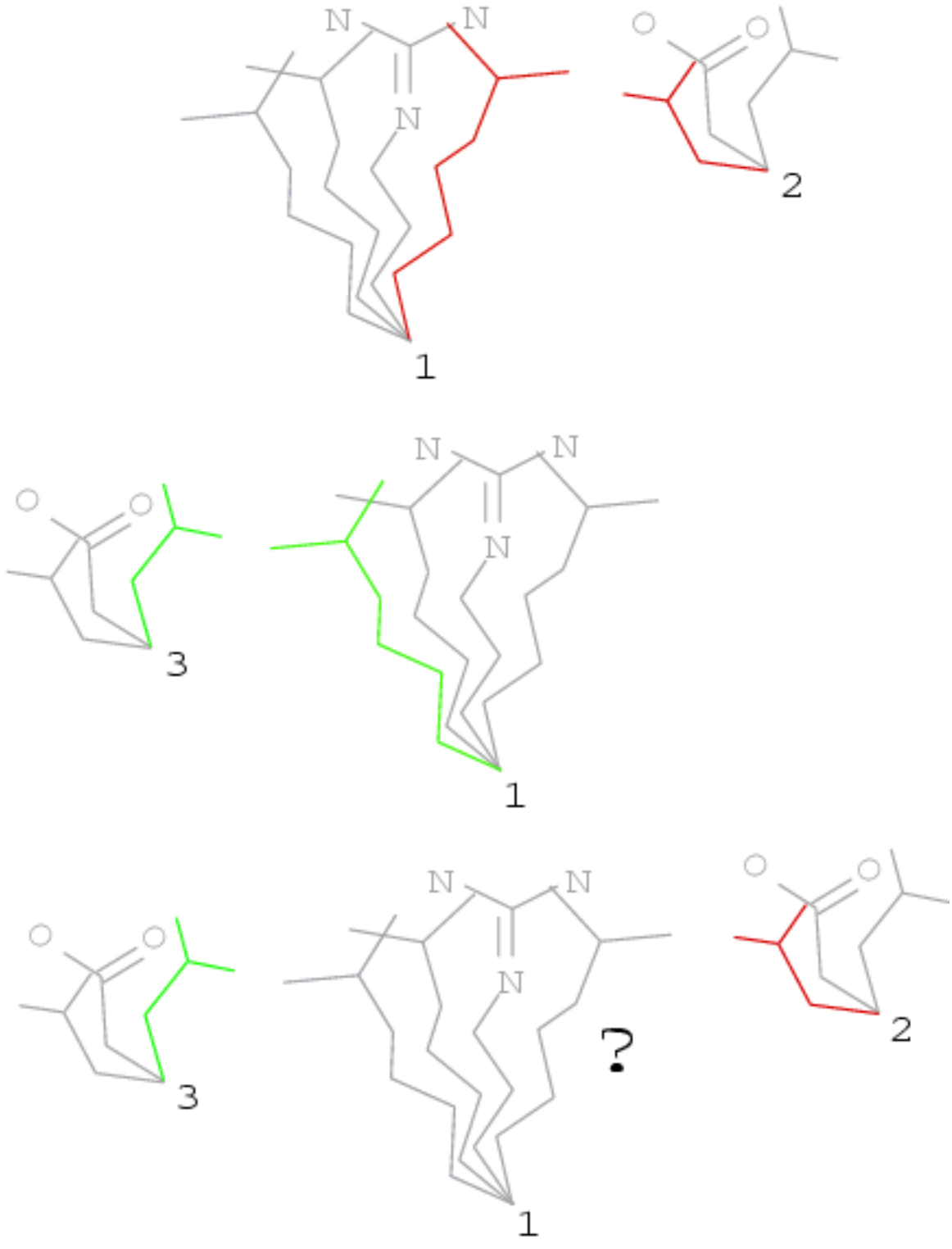


Figure 5.2: Illustration of how F_{total} can underestimate E of a true conformation. When residue 1 and 2 are considered together, their best conformation is the red subrotamers. When residue 1 and 3 are considered together, their best conformation is shown in green. When the subrotamers are frozen (bottom of figure) residue 1 can not simultaneously occupy both subrotamers, and therefore the favorable interactions that contributed to F can not all be fully realized.

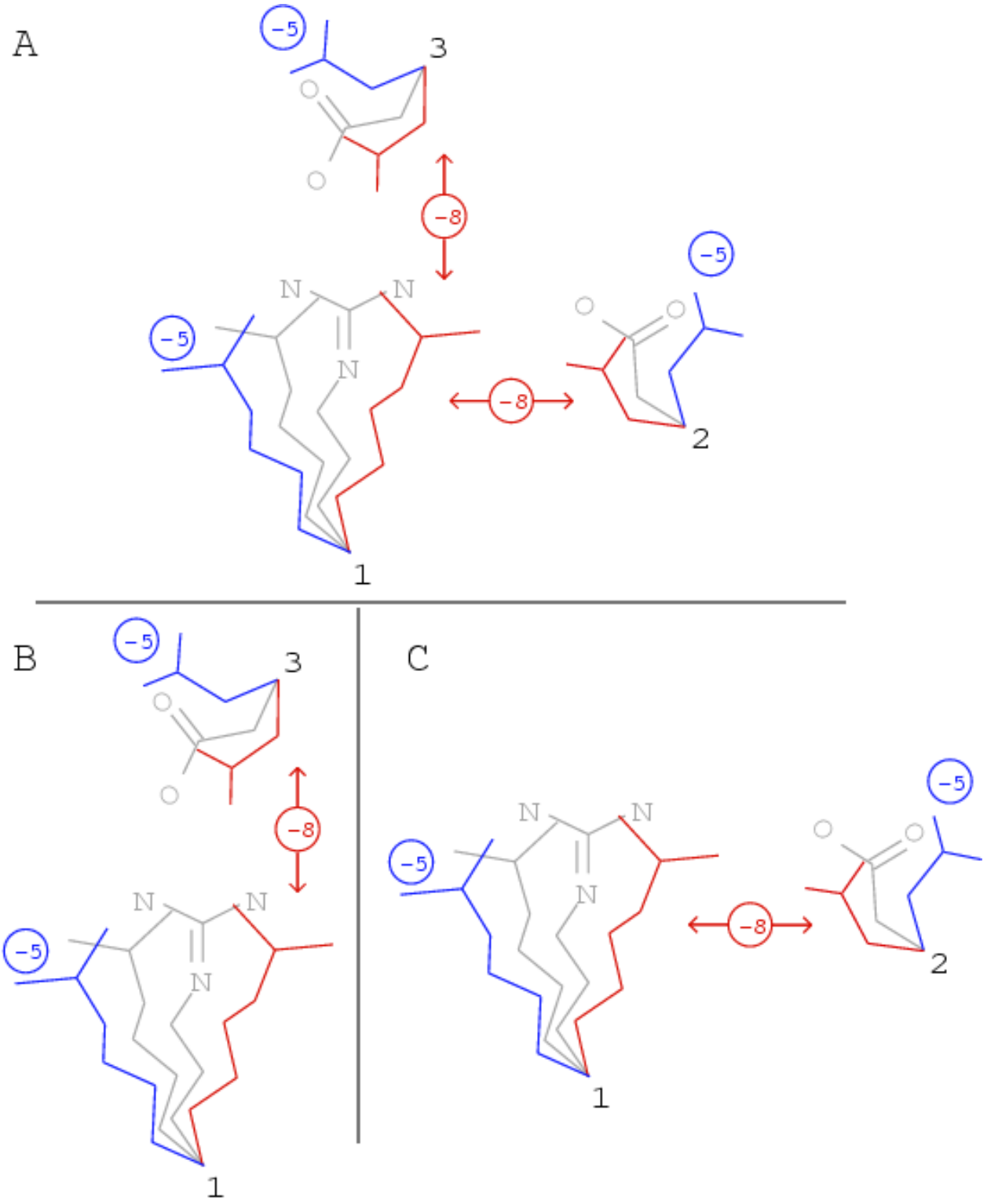


Figure 5.3: Illustration of how F_{total} can overestimate E of a true conformation. The blue subrotamers at each fleximer have a self E of -5. The red subrotamers shown have interactions worth -8. All other self and pair interactions are 0. The best conformation of all three residues (A) is therefore when all occupy the red subrotamers. However, when only residue 1 and 2 are considered (B), their best conformation is the blue subrotamers. Likewise, when residue 1 and 3 are considered (C), their best conformation is also the blue subrotamers.

Chapter 6

Design of Hydrophobic Core Packing in P22 Arc Repressor

6.1 Introduction

Substantial progress has recently been made toward the design of stably folded proteins. Computational design algorithms have produced novel sequences that adopt a known fold [29, 86, 128], as well as novel structures [54]. Despite some measure of success in these efforts, our understanding of the determinants of protein folding is not complete. The use of design algorithms in new systems will help test the generality of these design methods, and aid in our understanding of protein folding.

In this work, we examine the core of P22 Arc repressor. Arc is a 53-residue protein that folds as a homodimer, with each monomer containing a β-strand and two α-helices. The structure of Arc has been determined by NMR and x-ray crystallography [13, 17, 115]. The subunits of the Arc dimer are intertwined so that they form a single hydrophobic core, hence the folding and dimerization of Arc are essentially the same process [15]. We sought to change the sequence of Arc so that its preference for forming homodimers is changed to a heterodimer preference. This protein has been modified to form heterodimers previously by changing electrostatic properties of the monomers [58, 99], but here we attempt to redesign Arc as a heterodimer by altering the hydrophobic core. Changes in the core packing have the potential to provide a

higher degree of specificity than changes in the electrostatic residues at the surface.

Arc has been observed to form a helix in place of each strand when the sequence Asn-11, Leu-12, is switched to Leu-11, Asn-12 [27]. The resulting structure, known as "switch" Arc, has been found to exist in equilibrium with its wild-type structure when both residues 11 and 12 are leucine. We have also attempted to design sequences that favor switch Arc over the wild-type structure of Arc.

Previous design efforts [28, 29, 86, 128] have focused primarily on stabilizing a desired fold by placing side chains in such a way that the fold is predicted to be particularly stable. Other possible folds have been essentially ignored, because these other folds would be difficult to predict, or because a sequence that stabilizes a particular fold is probably unlikely to stabilize other folds. Both of the design problems considered here involve choosing a sequence that distinguishes between two similar structures (i.e. sequences that prefer heterodimers to homodimers or switch Arc structure to wild-type Arc structure). These problems will therefore provide a stringent test of how specific the designed sequences are for a desired structure. In fact they seem to require explicit consideration of the structure to be disfavored as well as the favored structure, a consideration which complicates the design process. These systems may thus help us gain insight into the determinants of specificity in protein folding.

6.2 Methods

Design of Heterodimeric Arc. The starting coordinate set for the calculations described here was the first of two dimers in the asymmetric unit of a 1.8 Å resolution structure of Arc repressor [103]. Polar hydrogens were added to the structure using the HBUILD facility in the program CHARMM [18, 19]. Only hydrophobic amino acids were used, since each of the positions considered was buried. The Dunbrack and Karplus [34] rotamer library was expanded by adding rotamers for the aromatic side chains at ± 1 SD about χ_1 and χ_2 angles. Trial calculations showed that the crystal structure was predicted to be unstable with this rotamer library because of the default

covalent geometry of a Trp side chain differed from the crystal structure at its C_α–C_β–Cγ angle. In the crystal structure, this angle is 116.4° in Trp A14 and 117.4° in Trp B14. The default value of this angle as tryptophan is built is 112.5° using CHARMM PARAM19 parameters. The rotamer sets for each aromatic side chain were expanded to include 115.0° and 117.5° for this angle for aromatic residues. The total number of rotamers per amino acid was: Ala, 1; Ile, 9; Leu, 9; Met, 27; Phe, 162; Trp, 243; Val, 3.

The CHARMM PARAM19 parameter set [18] was used to compute van der Waals interactions and torsional angle energies for each rotamer at each position in the protein. All rotamers were built in their default geometry in this force field. Covalent strain (of bonds, bond angles) was not considered, except when explicitly specified in the rotamer library. Thus, the appropriate penalty for distorting the C_α–C_β–Cγ angle was included. The fixed atoms of the system consisted of the protein backbone and all the residue positions that were not optimized. Rotamers were eliminated prior to DEE/A* if their interaction with the template was greater than 25 kcal/mol. Each rotamer's interactions with the fixed atoms and with all other rotamers were computed and stored prior to DEE/A*.

The search for low-energy conformations of the system used DEE of single residues [32, 48] followed by conformational splitting [84, 110] and elimination of pairs [32, 75]. The DEE was repeated until no further rotamers could be eliminated by application of the criteria, then an A* search [78, 145] was performed on the remaining space to find the optimal solution. Once the best energy solution was found, DEE was repeated with an energy cutoff so that all solutions within 10 kcal/mol of the optimum could be determined. A depth-first A* search was used so that all solutions within 10 kcal/mol of the minimum could be found.

For each sequence that was found, the lowest-energy structure corresponding to the sequence was minimized. The backbone and unoptimized side chains remained fixed at their crystal structure coordinates during the minimization. The minimization consisted of 50 steps of steepest-descent followed by up to 2000 steps of adopted basis Newton-Raphson [18]. The minimization terminated early if further steps would not

reduce the energy by more than machine precision. Any covalent strain which was added by minimization was included in the total energy of the structure.

A hydrophobic burial term of 25 cal mol^{-1} Å^{-2} [125] was added to the energy of each structure. Since the folding and dimerization of Arc are essentially one process, the surface area calculation requires a representation of the unfolded state. Each amino acid was built as an extended chain blocked "dipeptide," and its unfolded solvent-accessible surface area was computed using CHARMM. The folded surface area of each structure was also computed, and the burial upon folding of Arc was simply $\mathrm{SASA}_{folded} - \sum \mathrm{SASA}_{unfolded\ aa}$ where the sum is over the amino acids in the Arc structure.

The set of low energy Arc sequences was filtered as described in Results. Arc structures that had more than 45 atoms in the core, or had similar numbers of atoms in their heterodimers and homodimers, were not considered further. Each possible heterodimeric Arc was built with its A and B monomer sequences reversed. The process was repeated using DEE/A* to find the best rotamers given that sequence and minimizing the energy to give $E_{\mathrm{rev,min}}$.

Design of Switch Arc. The Dunbrack and Karplus [34] rotamer library was expanded by adding rotamers at $\pm 10°$ and $\pm 20°$ for χ_1 and χ_2. Met side chains were only expanded by $\pm 15°$ at each of their χ angles. Additional rotamers were added for the position of the hydroxyl H in the Tyr side chains. The total number of rotamers per amino acid was: Ala, 1; Ile, 225; Leu, 225; Met, 729; Phe, 300; Trp, 450; Tyr, 2700; Val, 15.

The rotamers that varied by only 10° or 20° were grouped into flexible rotamers [90], or fleximers. The DEE/A* search found the best fleximer states of the protein for each of the available 14 NMR structures [27]. The possible sequences that came out of the fleximer search were then optimized one at a time by allowing all possible individual rotamers at each variable position. This second set of DEE/A* searches gave structures with true low energies, and allowed us to discard the sequences that were only predicted to be low in energy because of the flexible rotamer approximation. Sequences were optimized in this way until all sequences that were within 10 kcal/mol

of the minimum energy were determined. (This often meant finding a set of structures with more than a 10 kcal/mol range in approximate energy.)

The energy function included a penalty for loss of side chain entropy upon folding [33]. Other details of the energy function, including hydrophobic surface area burial and van der Waals energies are the same as for Arc heterodimers, above.

6.3 Results and Discussion

6.3.1 Design of Heterodimeric Arc

The structure of Arc is shown in Figure 6.3. Residues Trp 14, Val 22, Ile 37 and Val 41 from each monomer form the hydrophobic core of Arc repressor. In wild-type, the majority of the inter-monomer contacts in the core are mediated through residue 41. We therefore describe the core as being divided into two half-sites. The first consists of residues Trp 14, Val 22, and Ile 37 from chain A, and Val 41 from chain B (left side of Figure 6.3). The first three residues contact one another and the have relatively little interaction with the second monomer. There is a second symmetry related subsite on the right-hand side of Figure 6.3.

The goal of this work was to redesign this hydrophobic core by placing side chains that would favor the formation of a heterodimeric Arc. The basic approach was to use a DEE/A* search that varied the sequence and rotamer states at all these positions and identify sequences that would most stabilize the structure. In the sequence results that follow, the first four letters refer to the amino acid side chains at positions 14, 22, 37, and 41 in the first Arc monomer, and the second four letters give the sequence in the second Arc monomer. Hence, the sequence of wild-type Arc is WVIV WVIV. A preliminary calculation in which the 100 lowest–energy sequences were determined using a DEE/A* search was performed. The designed sequences tended to have extra heavy atoms in the core. In wild-type Arc, the eight side chains considered in the calculation have a total of 40 heavy atoms. The number of atoms in each of the designed sequences is plotted as a histogram in Figure 6.1. The average number of

atoms in these sequences is 43.4. Ninety-two of these sequences have more atoms than wild-type, and only one has fewer atoms. Thus it appears that the design algorithm, as applied to this structure, is more likely to overpack the hydrophobic core of Arc than it is to underpack. The extent to which this problem occurs may depend on the structure of the protein of interest, and will almost certainly depend on the fineness of the rotamer library. When fewer rotamers are used, the number of atoms computed to be in the core tends to decrease because the highly packed sequences are not as likely to find a lower energy structure. The underpacking problem has been encountered in similar calculations by other groups. In fact, the underpacking caused by use of discrete rotamer libararies has prompted the use of reduced van der Waals radii in several design calculations [29, 30].

It is known from previous work that some conservative substitutions can be made in which the number of atoms in this part of the core increases from 40 to 42 (for example, Leu at position 22 or Ile/Val at position 41 [16, 92]). We refined our search algorithm so that any sequences were not considered further if they contained more than 45 atoms in these core positions. The best sequences that were found after applying this constraint presented another difficulty–they were not expected to form heterodimers selectively. Some of the low-energy sequences such as WLLV WVLV and WLII WVLI are nearly symmetric already. When their corresponding homodimers (e.g. WLLV WLLV and WVLV WVLV for the first sequence) are built, they are predicted to be nearly as stable as the desired heterodimer. In fact, WVLV WVLV is already known to be stable [92], and its stability diminishes the chance that one of its monomers can be used to select for heterodimers.

The basic DEE/A* algorithm simply attempts to find a structure with as low an energy as possible. There are cases where this minimization appears to be sufficient to prevent other structures from forming [29, 86]. However, in this system, the homodimer structure is similar enough to the desired heterodimer structure that we should consider the predicted stabilities of both. In fact, a simple search for the best energy can give homodimers.

An additional constraint to the design process was needed to increase the

heterodimer preference. The first half-site in the core of Arc is comprised of one side chain from segment B (residue 41) and three residues from segment A. If residue 41 is the same side chain in both segments, then the homodimer is likely to be approximately as stable as the heterodimer. For instance, if A41 is Phe, then the other residues in the first half-site would arrange to accomodate a Phe. If B41 is also Phe, then it is likely to be accomodated in a homodimer just as A41 is accomodated. However if B41 were Ala, then the homodimers are much less likely to be stable. (See Figure 6.2.) In the following calculations, the sequence was constrained so that residue 41 in segment A would have a different sized side chain than in segment B. Sequences were only considered if the size of the side chains A41 and B41 were different by 3 heavy atoms or more.

Since the energies that come directly from the DEE/A* search are not perfect descriptors of the protein's actual stability, several different measures were applied to each candidate structure in order to determine which structures have the best chance of folding as heterodimers. The energy E from the DEE/A* search consists of a van der Waals energy and a term for covalent strain. Each structure with a low E underwent minimization in which only the variable side chains (14, 22, 37, and 41) were allowed to move. The energy of the minimized structure $E_{\min}$ provided a second criteria to judge the stability of a structure. This measure is valuable since the discrete search can sometimes miss good conformations of a sequence. A simple validation of this procedure was performed on the wild-type structure. The side chain conformations predicted by the DEE/A* search are shown in red in Figure 6.4, along with the original side chains as shown in Figure 6.3. The side chains after minimization are shown in green. The predicted structure both before and after energy minimization is very closely superimposed on the actual crystal structure, and thus the procedure works in the case where the predicted structure can be verified.

After computing E and $E_{\min}$ Each heterdimeric sequence was then reversed so that the sequence ABCD WXYZ was placed in the crystal structure as WXYZ ABCD. The best conformation of this sequence was chosen via DEE/A* and the structure was minimized. In principle, this should give exactly the same protein and thus the

same stability. However, the available structures of Arc are asymmetric, and the energy of the reversed structure $E_{\mathrm{rev,min}}$ is often different than E_{min} in the original.

Taken together, the above considerations led to the following design procedure. A DEE/A* search was performed, allowing all eight side chains to be any hydrophobic residue (Ala, Val, Ile, Leu, Met, Phe, or Trp). All structures with a computed E better than the wild-type sequence were then filtered so that the remaining sequences contained 45 or fewer heavy atoms, and positions A41 and B41 had different sizes. Sequences were also filtered if neither of their corresponding homodimers were overpacked (45 or more atoms). The lowest-E structure for each remaining sequence was then minimized, giving its energy E_{min}. Each sequence also had its two corresponding homodimers built using DEE/A* and minimized, and its "reversed" sequence built and minimized. Sequences were considered further if all three energetic measures (E, E_{min}, and $E_{\mathrm{rev,min}}$) were better than the corresponding energies for wild-type Arc. There are 10 sequences that satisfy all the above criteria, all of which are predicted to have a heterodimer preference based on comparison of their E_{min} values. The sequences, shown in Table 6.1, fall into two categories. The first five shown have the sequence WFXX WXXA (X is a variable amino acid), and the other five have the sequence WAXX WXXF. For each sequence, the three energies E, E_{min}, and $E_{\mathrm{rev,min}}$ are shown, and the total number of heavy atoms in the core is shown. Below each sequence, the two competing homodimers are listed, along with their respective values of E, E_{min} (for a homodimer, E_{min} must equal $E_{\mathrm{rev,min}}$), and number of core atoms. Many of the homodimers shown in Table 6.1 have a very unfavorable value of E because the discrete rotamer states did not allow efficient packing of the side chains. However, minimization often relieved the strain in these homodimers, and hence it is not entirely certain that they will be unstable.

Sequence (1) has the most favorable E of the first group of sequences. The predicted structure of this protein is shown in Figure 6.5, along with the side chains of the original Arc repressor just as in Figure 6.3. In the heterodimer, Phe A22 occupies some of the space vacated by side chain B41 as it mutated from Val to Ala. Some of this space is also filled by Leu A37, which is branched at its C_γ, whereas Ile A37 in

wild-type is not. In filling the space at the left of Figure 6.5, however, Leu A37 leaves a gap in the center of the core. Ile A41 partially fills the space left by mutation of both Ile B37 and Val B41, but the net effect of these mutations is to enlarge the gap at the center. The gap can be clearly seen in bottom panel of Figure 6.5, where the side chains are shown with space-filling spherical atoms. The gap in the center is not present in wild-type Arc (see Figure 6.4, bottom panel), and raises doubt about the stability of the heterodimer structure.

Other sequences in the first group of Table 6.1 have a similar gap in the center. The sequence with a different packing is sequence (3)—WFLM WMIA. The predicted structure is shown in Figure 6.6. The packing at the left hand side of the figure is similar to sequence (1), which is unsurprising since residues A14, A22, A37, and B41 all have the same identity as in sequence (1). The difference in this structure is that Ile B37 (conserved from wild-type) is β-branched and partially fills in the center of the core. Also, Met A41 extends toward this central region, and fills a part of it. Met B22 in turn fills space opened by the mutation to Met A41. Although sequence (3) does not have the large hole seen in sequence (1), there are still small gaps. Residues Phe A22 and Met B22 leave a small space since the Val's in wild-type are β-branched. Perhaps more importantly, sequence (3) contains two methionines, which have greater torsional freedom in the unfolded state and therefore carry a greater penalty when folding into the structure shown. The entropic penalty is not explicited counted in Table 6.1, but has been estimated at 0.7–0.8 kcal/mol per residue greater than in leucine or isoleucine [33]. This entropy cost may make sequence (3) slightly less stable than wild-type.

Sequence (7) is predicted to be the most stable of the WAXX WXXF sequences based on its computed energy. The predicted structure with this sequence is shown in Figure 6.7. Note that the main difference between the first group of sequences (WFXX WXXA) and the second group (WAXX WXXF) is the interchange of Phe and Ala on positions A22 and B41 in the lower left part of Figures 6.5–6.7. Sequence (7) results in a structure that appears to be better packed than sequence (1). Phe B41 occupies some of the space taken up by Val B41 in wild-type, and Ile A37 extends

toward the central part of the core to space filled by B37 and B41 in wild-type. Ile A41 has an extra methyl group, which fills the gap left by the mutation of Val A22 to Ala (left-hand side, rear of figure).

It is worth looking at the criteria that were used to select these structures in order to understand which possibile structures were eliminated by each type of criterion. One energetic measure, E_{rev}, the energy of the reversed sequence before energy minimization, was not included in the selection process. If we require that this is less than E for wild-type Arc in addition to the other three energetic measures, then the sequences that result are all either homodimers or heterodimers with a small predicted heterodimer specificity (less than 2 kcal/mol). These predicted heterodimers leave a thin margin for error, and it is difficult to predict with any confidence that they will preferentially form heterodimers.

Since the heterodimers with E_{rev} better than wild-type do not appear to form heterodimers, it was necessary to relax this restriction and stipulate that $E_{\mathrm{rev,min}}$ is less than $E_{\mathrm{rev,min}}$ for wild-type. The results already presented in Table 6.1 show sequences with a substantially larger heterodimer preference. These sequences were constrained to be heterodimers by requiring that (1) there was a difference in the size of side chains A41 and B41, and (2) that one of the homodimers was overpacked (more than 45 heavy atoms in the core). These two criteria have essentially the same effect on the set of sequences that is selected. Removing restriction (1) would add one more sequence to Table 6.1, WFLV WVLA, which is quite similar to sequence (1) in the table. Lifting restriction (2) would also add one more sequence to the set, WMLI WVLA, which is also similar to sequence (1) in the table. However, lifting both heterodimer criteria results in the selection of an additional 40 sequences. All 40 of the new sequences had a smaller predicted heterodimer preference than all 11 of the sequences shown in Table 6.1. Thus the "qualitative" criteria (1) and (2) are equivalent to an energy cutoff for heterdimer preference in the system studied here.

A curious feature of the results on sequence (1) and its homologues is that the energy of the core was predicted to be lower than that of wild-type, but the packing left a substantial hole in the center of the protein. The van der Waals energy associated

with packing the side chains on either side of the core compensates for the loss of van der Waals interactions with atoms in the core. Other sequences also appear to have slight flaws in their core packing. Determining the extent to which these flaws affect the stability would require experimental tests.

In the end, there were no sequences that were unambiguously predicted to form stable heterodimers specifically. There may be a variety of reasons for this result. The crystal structure is asymmetric in such a way that when the sequences were built in reverse, the energies differed significantly. There were a few sequences that were stable without minimization both forward and reversed, but none of these were predicted to have a significant heterodimer preference. The results shown have a low energy reversed sequence only after minimization.

A second reason that no sequence stands out unambiguously is the energy function that was used to evaluate the structures. It is known that some structures with a low predicted energy are not as stable as wild-type Arc, as shown previously with the overpacked structure. This led us to compensate for the energy function with number-of-atom based cutoffs, but the ideal solution would be to improve the energy function. In this system, one might consider increasing the van der Waals radii in order to prevent overpacking. Molecular mechanics parameters are designed for simulation at $\approx$ 300 K, but the dead-end elimination calculation done here was essentially at 0 K, since we seek the lowest energy structure. The calculation does not account for the range of motion atoms would undergo at 300 K, and as a result the structure can be overpacked.

Finally, there is no guarantee that a stably folded heterodimeric Arc exists. If it could, it may require changing more residues than were permitted here. There may have been unique problems for this core, as Trp 14 was highly conserved in the solutions found, and the remaining 6 residues were all β-branched in wild-type Arc. Changing a β-branched side chain to another β-branched side chain is always a conservative substitution, and changing to a non-β-branched side chain leaves a gap where the branch occurred. This type of gap was shown at the center of the core in the designed sequences. It may be more likely that such gaps can be filled if a

repacking of the core involves more residues or a wider variety of (non-natural) amino acid side chains.

6.3.2 Design of Switch Arc

Switch Arc adopts a helical structure in residues 9–14 [27], which form a β-strand in wild-type Arc. The helical structure of switch Arc allows Leu 11 to be buried in the core, whereas Asn 11 in wild-type is exposed to solvent. The structures of the two proteins are shown in Figure 6.8. Switch is shown with residues Phe 10, Trp 14, and Leu 19, the three side chains that contact Leu 11. Note that in wild-type Arc, Leu 11 is solvent exposed, but the other three side chains are buried.

A variant of Arc in which both residues 11 and 12 are Leu (Arc-N11L) has been found to exist in equilibrium between an Arc and a switch Arc structure [26]. This bridge between the two structures provides a means of studying the determinants of stability of the two structures. Here we attempted to repack the eight residues shown in Figure 6.8 so that the sequence would favor the switch Arc structure over the wild-type structure.

Mutation of residue 11 in switch Arc. As a test of the repacking methodology, a DEE/A* search was performed in which residue 11 could mutate to any hydrophobic residue, and residues 10, 14, and 19 were constrained to their switch Arc sequence but were free to adopt any rotamer conformation. All other protein atoms were fixed. The energy of each sequence was computed as

$$E = E_{vdw} + E_{mm} + E_{np} + E_{sc}$$

where E_{vdw} is simply the van der Waals energy of the protein computed in the CHARMM PARAM19 force field, and E_{mm} is the molecular mechanics potential from dihedral angles and covalent geometry using that force field. E_{np} is the energy of burying non-polar surface area, which is computed after the DEE/A* search as 0.025 kcal mol^{-1} Å^{-2}. E_{sc} is the contribution of side chain entropy to folding. The loss of entropy for each side chain was calculated from the values reported by Doig and

Sternberg [33], who surveyed different techniques of computing the entropy associated with each type of side chain.

The energies were evaluated for each of 13 NMR structures of switch Arc and for the minimized average NMR structure [27]. The results are shown in Figure 6.9. Each NMR structure is represented as a column, and each amino acid substitution is represented by a row in the figure. For each structure, the minimum energy sequence is defined as having an energy of 0 kcal/mol, and the other sequences' energies are measured relative to this best sequence. The colors in Figure 6.9 illustrate the energies, from white representing the minimum, to yellow, orange, and red representing higher energies, to black, the color of all sequences over 10 kcal/mol. The wild type sequence FLWLFLWL (where the letters represent the one-letter amino acid codes for the mobile residues 10, 11, 14, and 19 in chains A and B) is most often predicted to be best, although an L11M mutation, and in one case a L11A mutation, are predicted to be slightly more stable on a few of the backbones.

Experiments in which residue 11 was mutated [3] show that each of these mutations results in a stable protein. Near UV spectra and NMR spectra show that the position 11 mutants bear more structural similarity to switch than to Arc. These results suggest that switch Arc can adapt its structure more than the calculations allowed, either by changing the position of the backbone or by changing the rotamer states of additional side chains.

Design of switch Arc-specific packing. Although the DEE/A* search does not consider the full range of motion available to the protein, it still may be possible to design a sequence that accomodates one or more of the NMR structures. A DEE/A* search was performed in which the amino acids at positions 10, 11, 14, and 19 were allowed to vary in the background of all 14 switch NMR structures. The energy function was the same as in the calculation on the position 11 mutants. All sequences that gave a structure within 10 kcal/mol of the best structure were identified.

The minimum energy sequences from the computational results are shown in Figure 6.10. As in Figure 6.9, the color represents the energy of each sequence in each structure, but note that the reference energy is different in this figure, since

more residues were allowed to change identity. Each of the sequences that was a minimum in one or more structures is shown. One interesting feature of the results is that none of the lowest energy sequences appear among the best 10 kcal/mol in more than 6 out of the 14 possible NMR structures. The wild-type sequence is among those chosen most often, and its rebuilt structure is shown in Figure 6.8. The side chains all adopt nearly the same conformation as observed in the NMR structures.

With some backbones, there are other sequences that give more favorable energies than the wild-type. Two of the sequences seen in Figure 6.10 are simply mutations of Leu 19. The structures corresponding to the sequences FLWIFLWI and FLWWFLWW are shown in Figure 6.12. The L19I mutation in switch is relatively conservative, and the structure does not change a great deal, as seen in the figure. It does appear to leave a gap where one of the methyl groups of Leu 11 is, and this is one reason that the sequence is predicted to be less stable in the context of all the other NMR structures.

The structure predicted from an L19W mutation in switch is also shown in Figure 6.12 (bottom panel). Here, the Trp 19 residues extend away from the core and pack against the N-terminal portion of the backbone and the side chain of Met 7, as shown in the figure. The primary reason that this sequence is stable in some NMR structures and not in others is that the region of residues 1–7 has relatively few NOESY restraints, and is therefore flexible. A Trp 19 side chain does not fit into several NMR structures due to the conformation of this portion of the protein. In those that Trp does fit, it is predicted to have a fairly favorable folding energy. Note that the considerations of overpacking that applied to wild-type Arc do not apply to this position, since there is an opening (at least in some structures) for the Trp side chain to grow into space unoccupied by Leu. Averaging over the five FLWWFLWW structures highlighted in Figure 6.10, Trp 19 is 81% buried on average, and its ND atom is 4.0 Å from the carbonyl oxygen of Lys 6. Given the uncertainty in the structure of the N-terminal portion of the protein, it is at least possible that Trp forms a hydrogen bond here (or that it is solvent accessible).

Experiments have probed the stability of wild-type Arc to mutation at several

positions, including Leu 19. An alanine scanning mutagenesis [91] showed that L19A is less stable than wild-type by 1.9 kcal/mol. A selection for stable mutants gave Val and Gln residues at position 19 [16]. The L19I mutant in wild-type is predicted by a DEE/A* calculation to destabilize the wild-type structure relative to the wild-type sequence. The calculation does not tolerate Trp at position 19 at all, since this position is 98% buried in the wild-type structure. Therefore, if the L19I or L19W mutants of switch prove stable, they are predicted to increase specificity for the switch fold.

The minimum energy sequences also include AYWLAYWL and AYWWAYWW. These two structures are similar, and are shown in Figure 6.13. The placement of side chains does not vary significantly over the structures highlighted as low energy in Figure 6.10. Both include a similar arrangement of residues Ala 10, Tyr 11, and Trp 14. The stability of the structure with the L19W mutation depends on the particular NMR structure being used, as we have seen. The AYWL structure appears to pack the space well. The one potential problem is that Tyr 11 is buried without making a good hydrogen bond. It gets within 4.0 Å of the carbonyl O of Val 33, but the penalty for desolvating this side chain is not explicitly accounted for in the energy function. The L11Y single mutant of switch Arc has been found to have nearly the same stability as switch Arc [3], so evidently the desolvation of this residue does not completely preclude its burial.

There is experimental data that indicates how stable the F10A L11Y is in the wild-type structure. The F10A mutant was found to destabilize wild-type Arc by 2.7 kcal/mol [91]. Mutating position 11 was found to have little effect on stability either in the Ala scan or in the selection for stable mutants, where several mutants at this position were found [16]. Since residues 10 and 11 are on opposite sides of the β-strand in switch Arc, their cumulative effect is likely to be the same as the effect of the mutations individually. Thus we expect these mutations to destabilize wild-type.

The AYWWAYWW is simply the combination of the F10A L11Y with the L19W mutant described above. As before, the predicted stability of the protein with Trp at position 19 in switch depends on the structure of the N-terminal region, for which

there is little information. The region's flexibility among the NMR structures is the principal reason that aromatic residues are seen in some structures of the NMR family, and smaller side chains are seen in others. This result is even more readily apparent when we consider the sequences that were found most often. All sequences that were found to be low in energy (within 10 kcal/mol of the best sequence) in at least 5 of the 14 NMR structures are shown in Figure 6.11. The list of sequences is readily divided into two groups: one in which a non-aromatic residue is present at position 19 in both monomers (first column and top of second column in Figure 6.11), and a second in which Trp is present at position 19 in both monomers. NMR structures 5, 6, 7, 11, and 12 always give sequences in the second group, and most of the other structures give sequences in the first group.

Another feature that stands out in Figure 6.11 is that the vast majority of sequences are conservative substitutions from switch Arc with one or more Met residues (this is also true of some of the minimum energy sequences in Figure 6.10). This side chain is commonly observed apparently because it is unbranched and flexible (it has 3 rotatable bonds), and therefore can fit more easily into different spaces. The side chain entropy penalty associated with Met is greater than the other residues [33] precisely because of this flexibility, but the penalty is apparently not large enough to prevent its appearance at many positions in this system. In some design calculations, Met has simply not been considered as a possible core residue [128], but allowing it without having it appear in so many solutions may remain a challenge.

Conclusion. The cores of two different structural variants of Arc repressor have been redesigned using a DEE/A* search to generate a number of candidate structures, then using additional criteria to evaluate the structures and introduce other desired properties—particularly the specificity of one fold over another. The process resulted in some sequences that appear to satisfy the design conditions, but there is some ambiguity in the predicted stability of the new sequences. In both cases, relatively small variations in the structure of the backbone influenced the stability of predicted sequences (in one case, the normal and "reversed" sequences gave different results; in the other, different NMR structures gave different results). Experimental tests of the

stability of these proteins may shed some light on this problem. The DEE/A* search is an efficient method for placing side chains to stabilize a particular backbone, but ideally, in problems such as these, one could take into account the fluctuations in a protein backbone as well as design against an undesired fold. DEE/A* alone does not readily address the latter two concerns. Future computation and experiment may help us understand how to carry out the process more efficiently.

Table 6.1: Sequences with a Predicted Heterodimer Preference.

	Sequence	E	E_{min}	$E_{\text{rev,min}}$ [a]	n_{atoms} [b]
wild-type	WVIV WVIV	-86.0	-103.8		40
(1)	WFLI WVLA	-91.9	-109.3	-109.2	43
homodimer 1	WFLI WFLI	337.1	-113.0		50
homodimer 2	WVLA WVLA	-80.6	-93.4		36
(2)	WFII WVLA	-89.2	-108.5	-104.7	43
homodimer 1	WFII WFII	342.3	-103.8		50
homodimer 2	WVLA WVLA	-80.6	-93.4		36
(3)	WFLM WMIA	-88.5	-110.1	-106.5	44
homodimer 1	WFLM WFLM	6.3	-98.4		50
homodimer 2	WMIA WMIA	-82.2	-96.5		38
(4)	WFLI WALA	-88.1	-105.9	-103.9	41
homodimer 1	WFLI WFLI	337.1	-113.0		50
homodimer 2	WALA WALA	-69.1	-81.1		32
(5)	WFLM WVLA	-87.9	-107.5	-108.8	43
homodimer 1	WFLM WFLM	6.3	-98.4		50
homodimer 2	WVLA WVLA	-80.6	-93.4		36
(6)	WFLI WILA	-86.1	-106.4	-111.5	44
homodimer 1	WFLI WFLI	337.1	-113.0		50
homodimer 2	WILA WILA	-76.8	-93.6		38
(7)	WAII WVLF	-92.5	-110.0	-104.1	43
homodimer 1	WAII WAII	-73.2	-92.7		38
homodimer 2	WVLF WVLF	-32.8	-109.4		48
(8)	WALM WMIF	-90.2	-111.1	-105.2	44
homodimer 1	WALM WALM	-79.7	-101.3		38
homodimer 2	WMIF WMIF	48.5	-95.1		50
(9)	WAIM WMIF	-88.8	-109.3	-104.5	44
homodimer 1	WAIM WAIM	-75.4	-95.9		38
homodimer 2	WMIF WMIF	48.5	-95.1		50
(10)	WAII WILF	-86.7	-106.7	-106.4	44
homodimer 1	WAII WAII	-73.2	-92.7		38
homodimer 2	WILF WILF	-34.2	-106.7		50
(11)	WALI WILF	-87.8	-107.4	-105.2	44
homodimer 1	WALI WALI	-79.6	-99.0		38
homodimer 2	WILF WILF	-34.2	-106.7		50

All free energy values are in kcal/mol.

[a] Values of $E_{\text{rev,min}}$ for homodimers are equal to the values of E_{min}.

[b] n_{atoms} is the number of heavy atoms in the hydrophobic core consisting of the side chains of residues 14, 22, 37, and 41.

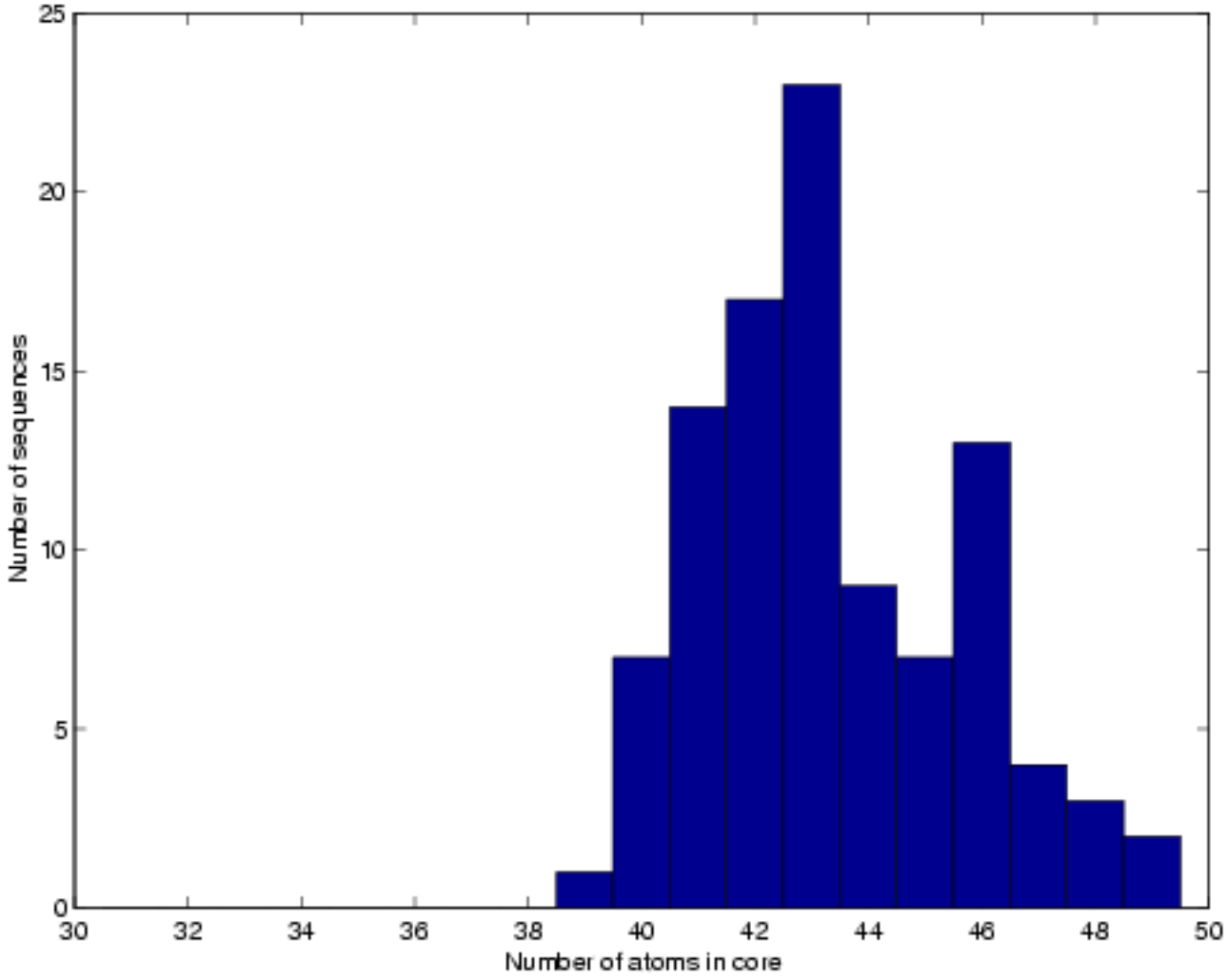


Figure 6.1: The number of heavy atoms in the hydrophobic core of various repacked Arc structures. The histogram shows the frequency of each number of atoms observed in a repacking calculation. Wild-type Arc has 40 heavy atoms in the space being repacked. Figure produced with the program MOLSCRIPT [73].

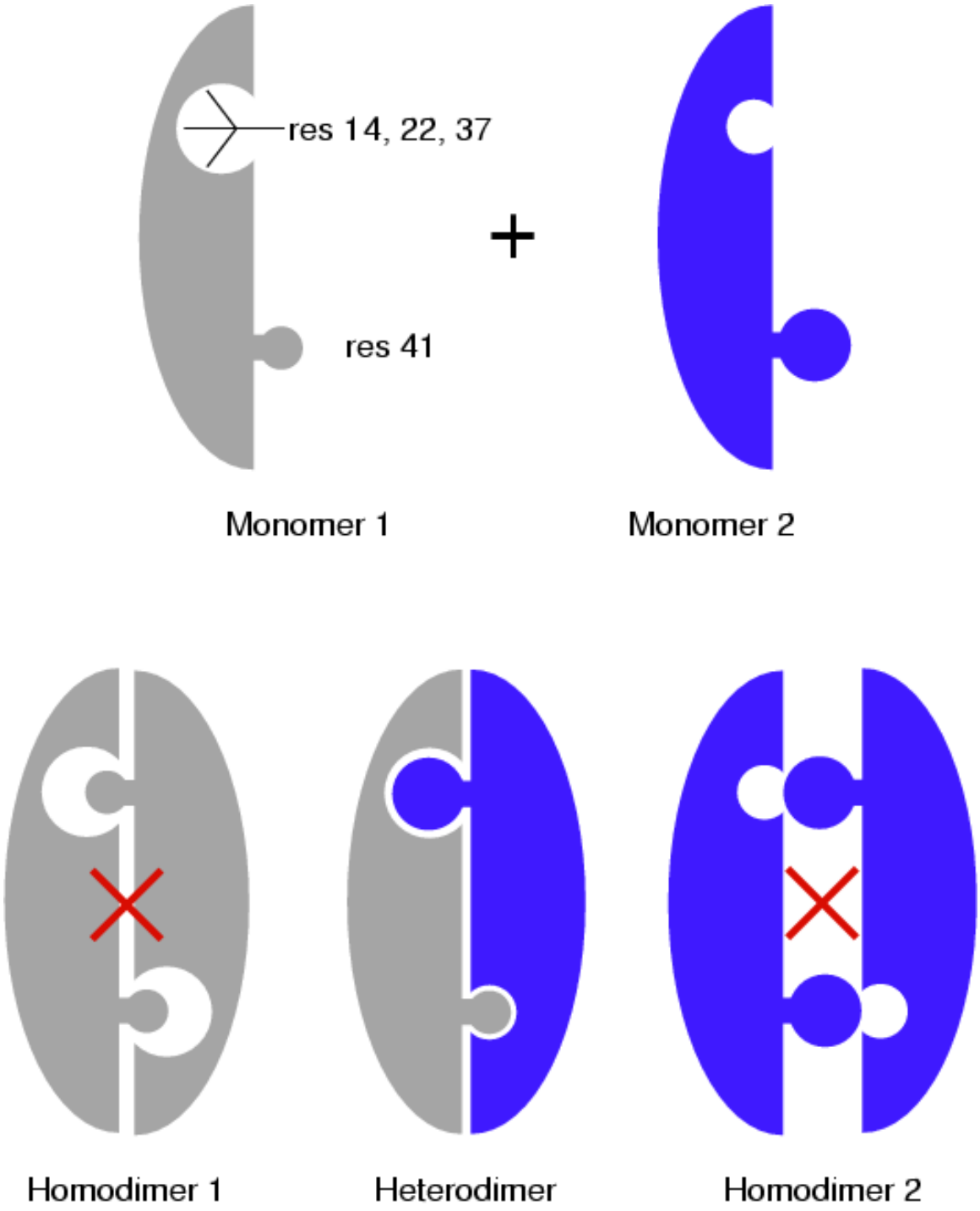


Figure 6.2: The strategy for selecting sequences with a strong heterodimer preference. In one monomer (gray), residue 41 has a small side chain and residues 14, 22, and 37 are arranged to accomodate a large side chain. In the other monomer (blue), residue 41 has a large side chain, and the other three residues accomodate a small one. As a result, the blue and gray monomers prefer forming heterodimers to homodimers.

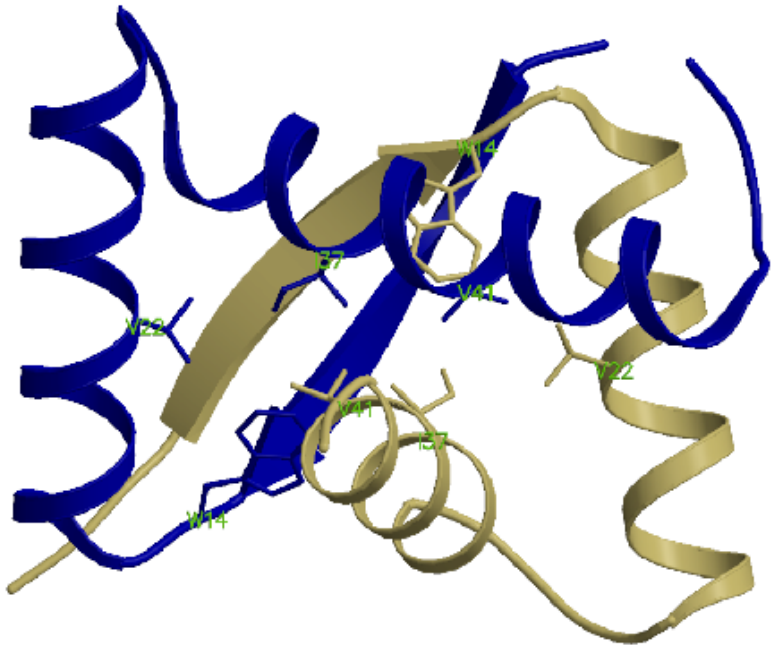


Figure 6.3: The structure of Arc repressor. Monomer A is shown in blue and monomer B is shown in yellow. The hydrophobic core residue side chains are shown in the same color as the backbone. The sequence as denoted in the text, WVIV WVIV, can be read from this picture by starting with the blue W14 residue in the lower left and following the labels clockwise around the figure.

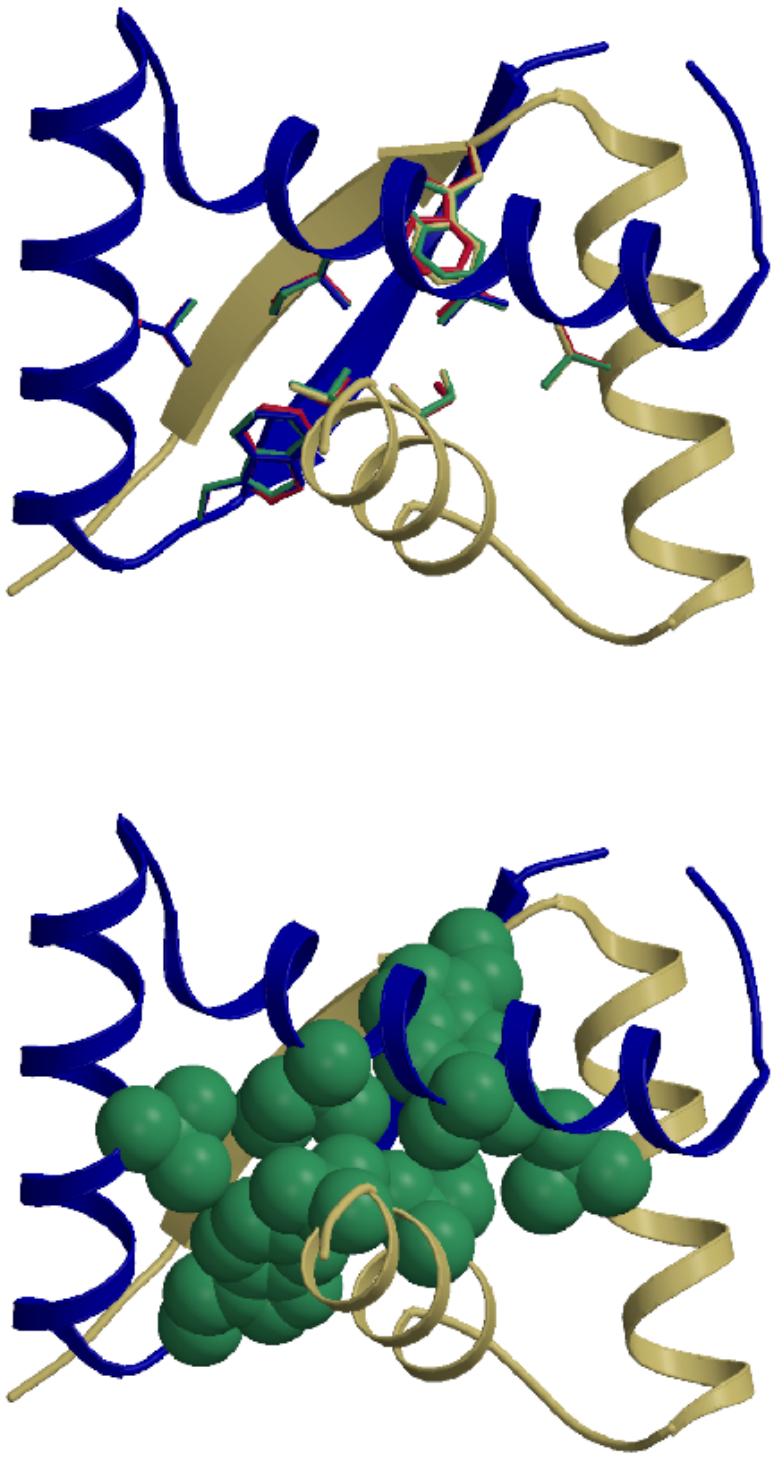


Figure 6.4: The structure of Arc repressor rebuilt. The backbone and crystal structure side chain positions are shown as in Figure 6.3. In the top panel, the rotamers selected from the DEE/A* search are shown in red, and the energy minimized positions of those side chains are shown in green. In the bottom panel, the energy minimized side chains (including C_α) are shown as space-filling spheres.

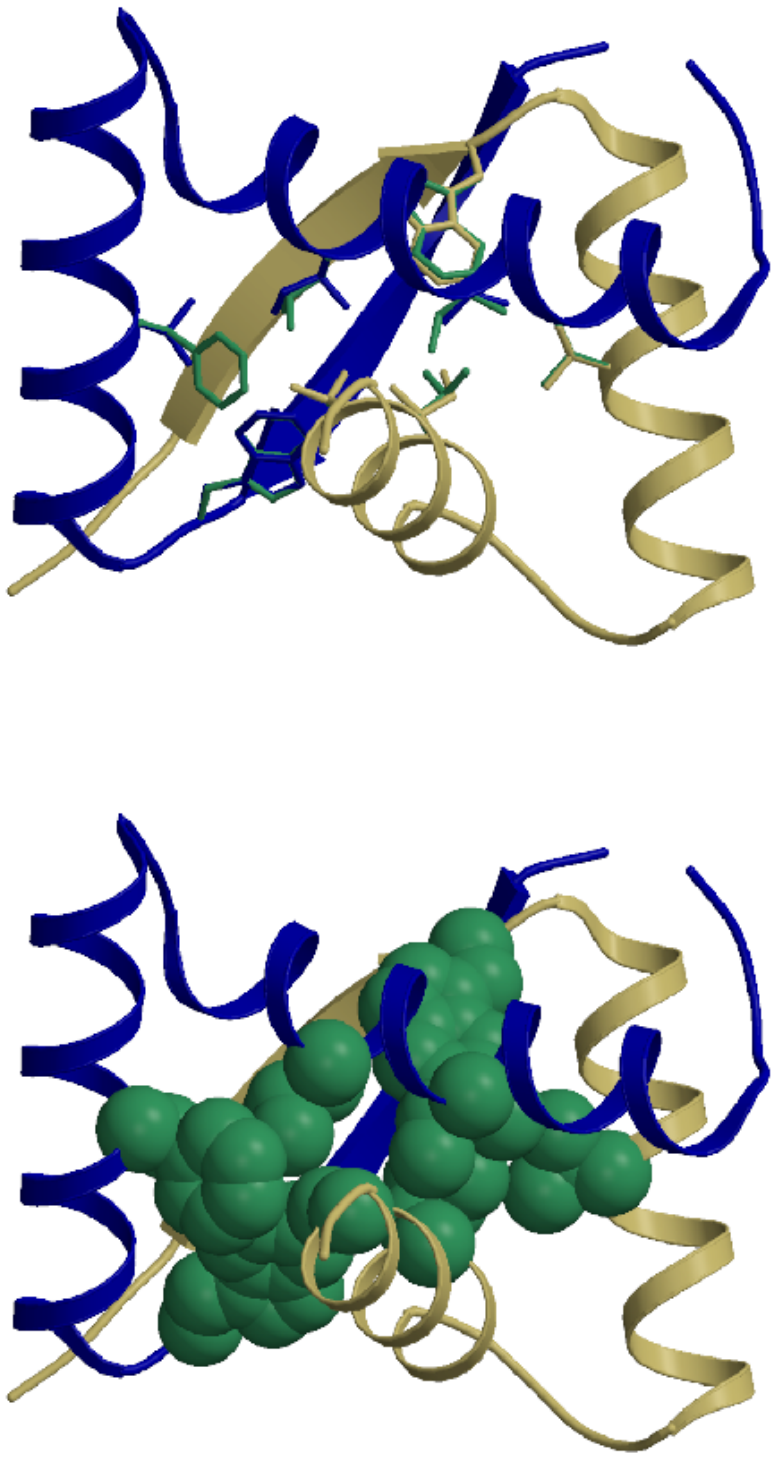


Figure 6.5: The structure of Arc rebuilt with the sequence WFLIWVLA. The crystal structure coordinates are represented in blue and yellow as in Figure 6.3. The top panel shows the energy minimized coordinates of the side chains, and the bottom panel shows the same side chains in space-filling spheres.

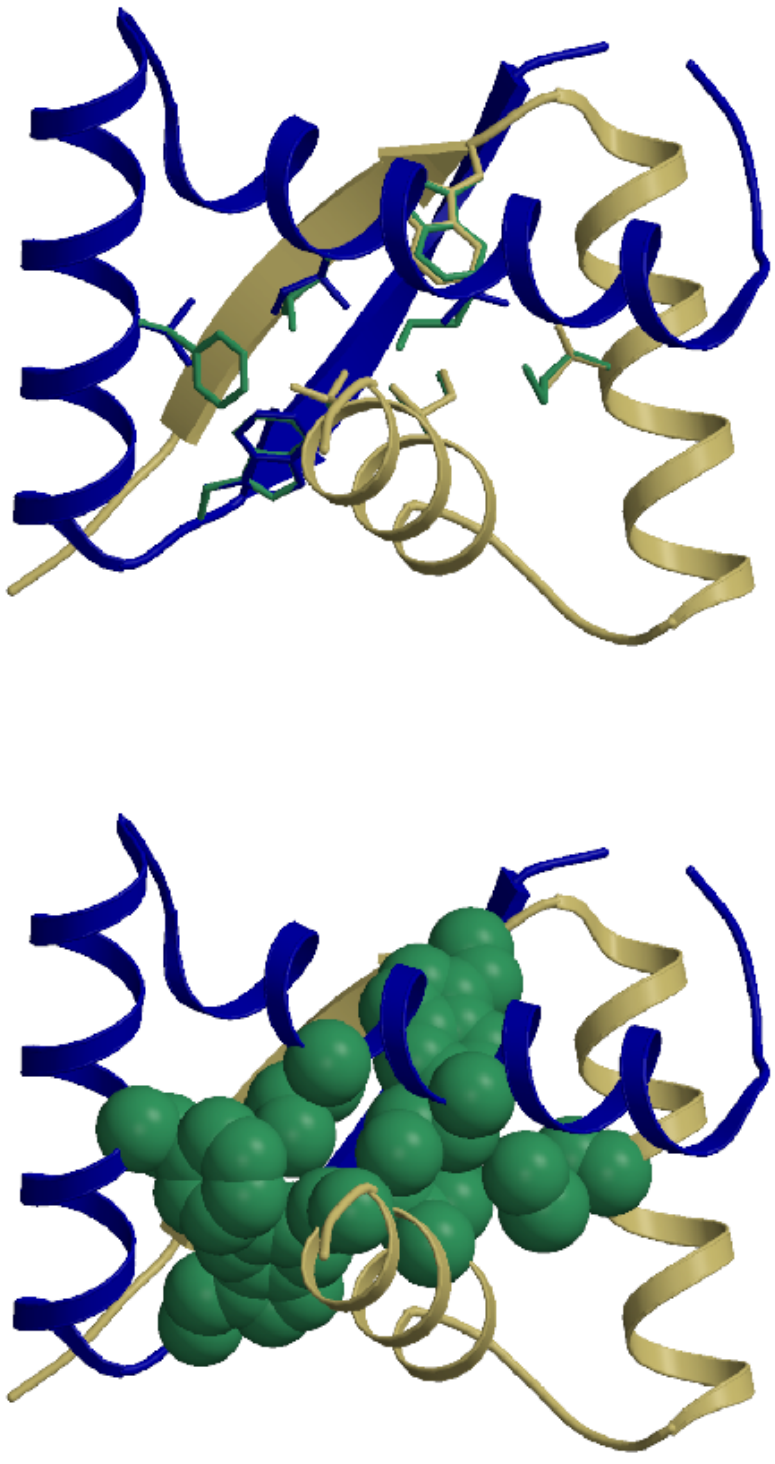


Figure 6.6: The structure of Arc rebuilt with the sequence WFLMWMIA. The crystal structure coordinates are represented in blue and yellow as in Figure 6.3. The top panel shows the energy minimized coordinates of the side chains, and the bottom panel shows the same side chains in space-filling spheres.

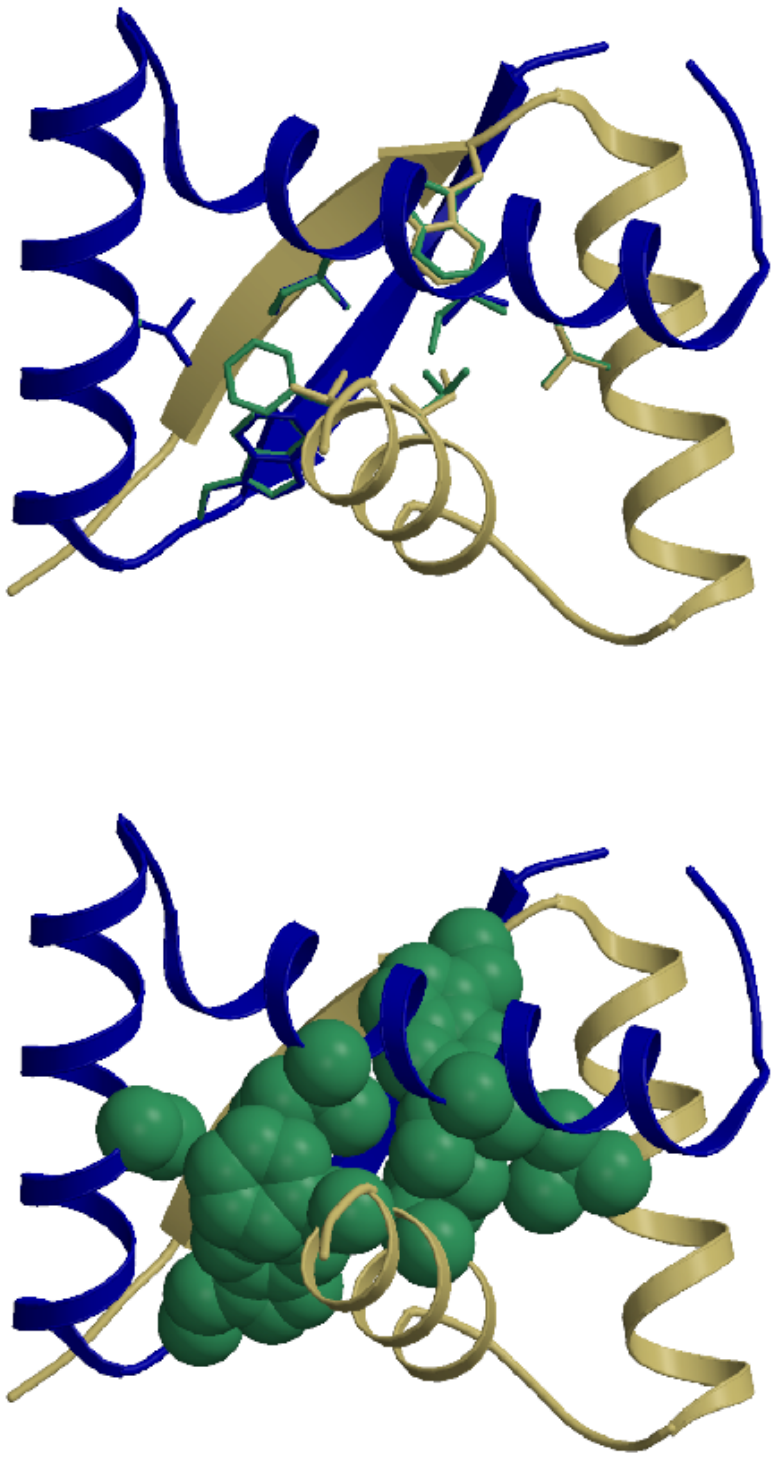


Figure 6.7: The structure of Arc rebuilt with the sequence WAIIWVLF. The crystal structure coordinates are represented in blue and yellow as in Figure 6.3. The top panel shows the energy minimized coordinates of the side chains, and the bottom panel shows the same side chains in space-filling spheres.

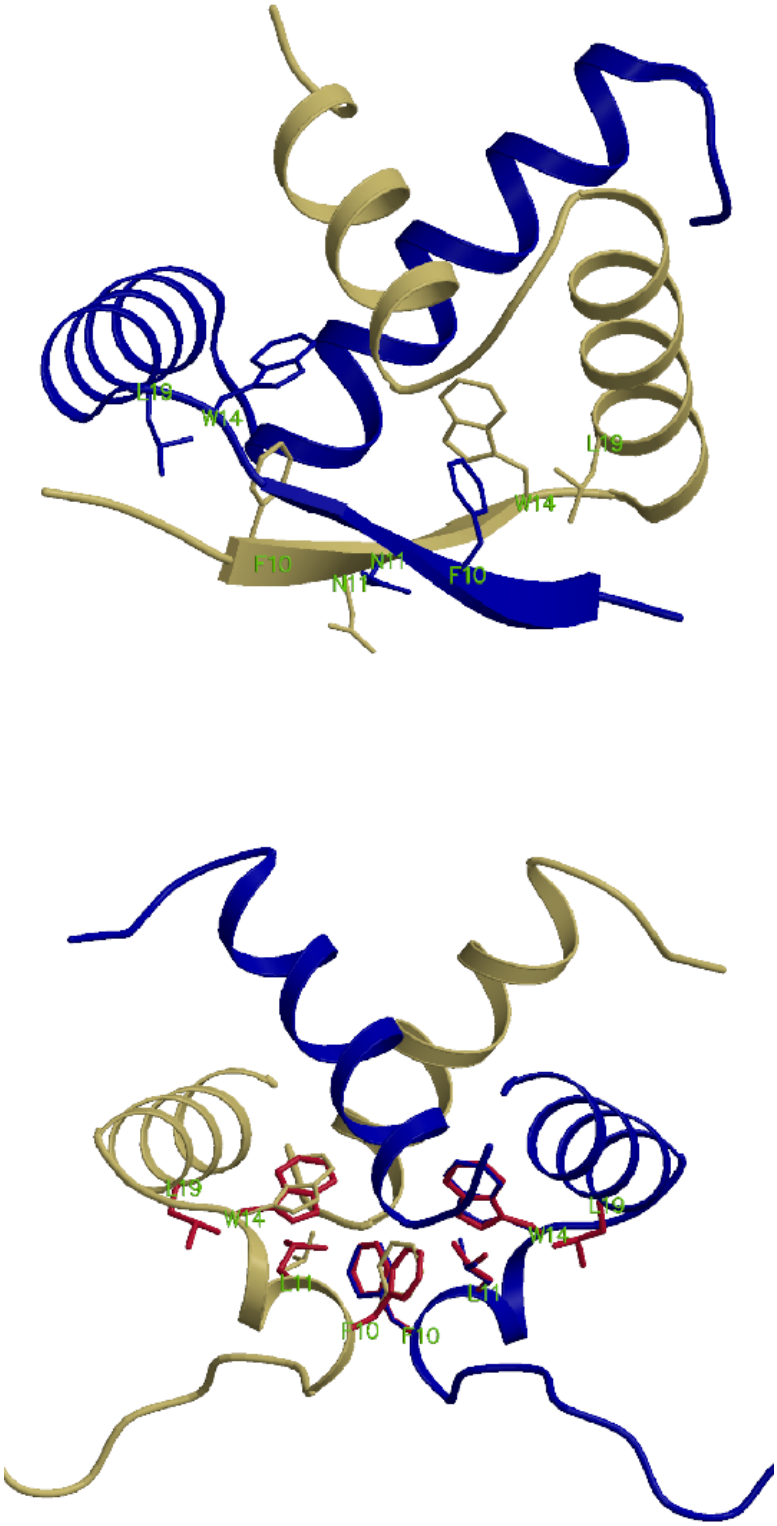


Figure 6.8: (top) The structure of wild-type Arc with residues 10, 11, 14, and 19 shown. (bottom) The structure of switch Arc with the same side chains shown. The NMR structure coordinates are represented in blue and yellow. The results of a DEE/A* search with the sequence constrained to regular switch Arc is shown in red. The rebuilt Leu 19 side chains are close enough to the NMR coordinates that they block the view of the NMR side chains.

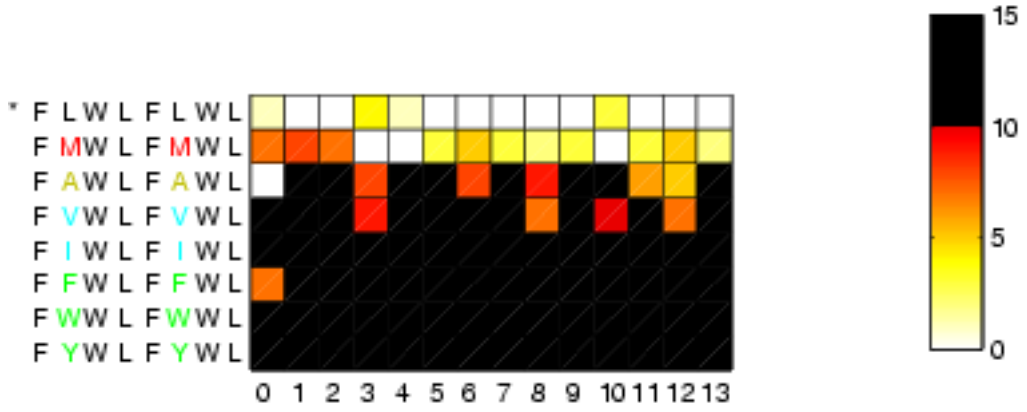


Figure 6.9: The relative stabilities of switch Arc predicted with various amino acid substitutions at position 11. Side chains at positions 10, 11, 14 and 19 are listed, and were allowed to move. The wild type amino acids are shown in black letters, Met is shown in red; Ala is shown in yellow; aliphatic side chains are shown in light blue; and aromatic side chains are shown in green. An asterisk (*) denotes the wild-type sequence. The energies were calculated for the 14 NMR starting structures, which correspond to the columns of the figure. Structures 1-13 are the NMR structures, and structure 0 is the minimized average structure backbone. The white squares correspond to the best sequence for a given structure, and the other colors (yellow–orange–red), show higher energy structures according to the scale at the right. Sequences higher than 10 kcal/mol in energy are shown in black.

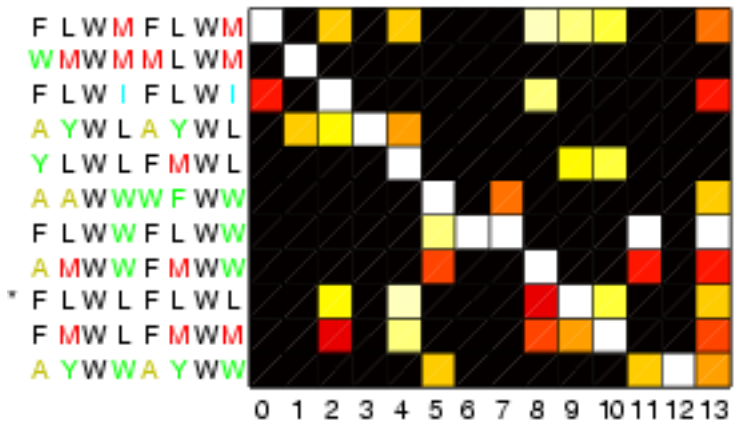


Figure 6.10: The minimum energy sequences of switch Arc predicted with amino acid substitutions at positions 10, 11, 14, and 19. The energies are computed relative to the minimum energy sequence for that structure. (Note the difference between this and Figure 6.9, in which only residue 11 was allowed to mutate.) An asterisk (*) denotes the wild-type sequence. The energy scale, the amino acid sequence colors, and the 14 starting structures are the same as in Figure 6.9.

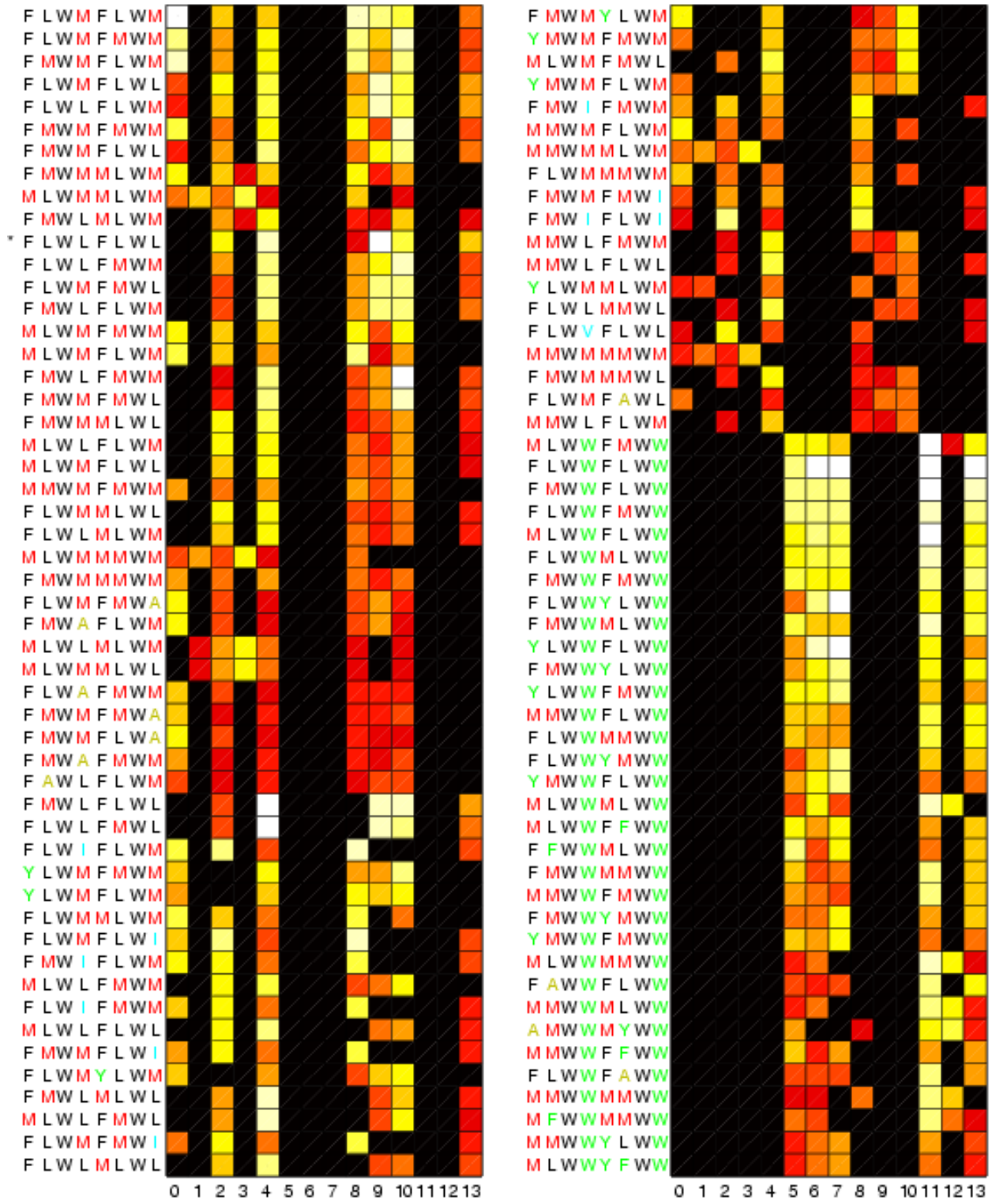


Figure 6.11: The most commonly occuring sequences placed in switch Arc. All sequences that occured in 5 or more NMR structures are shown. The energy scale, the amino acid sequence colors, and the 14 starting structures are the same as in Figure 6.9.

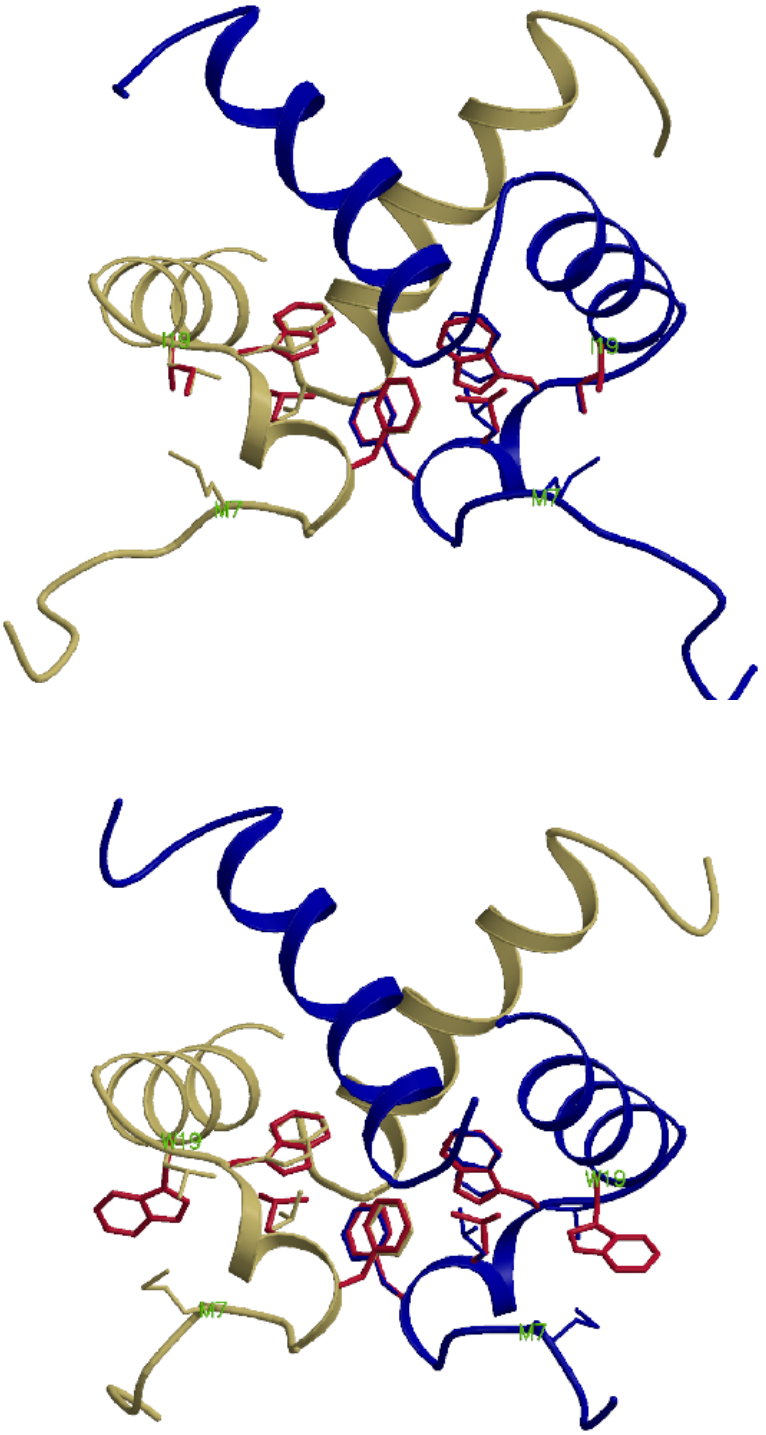


Figure 6.12: (top) The predicted structure of switch Arc with the sequence FLWI at positions 10, 11, 14, and 19. The backbone of NMR structure #2 of switch is shown in blue and yellow. The predicted side chain placements of the new sequence are shown in red. (bottom) The predicted structure of switch Arc with the sequence FLWW in the background of NMR structure #6.

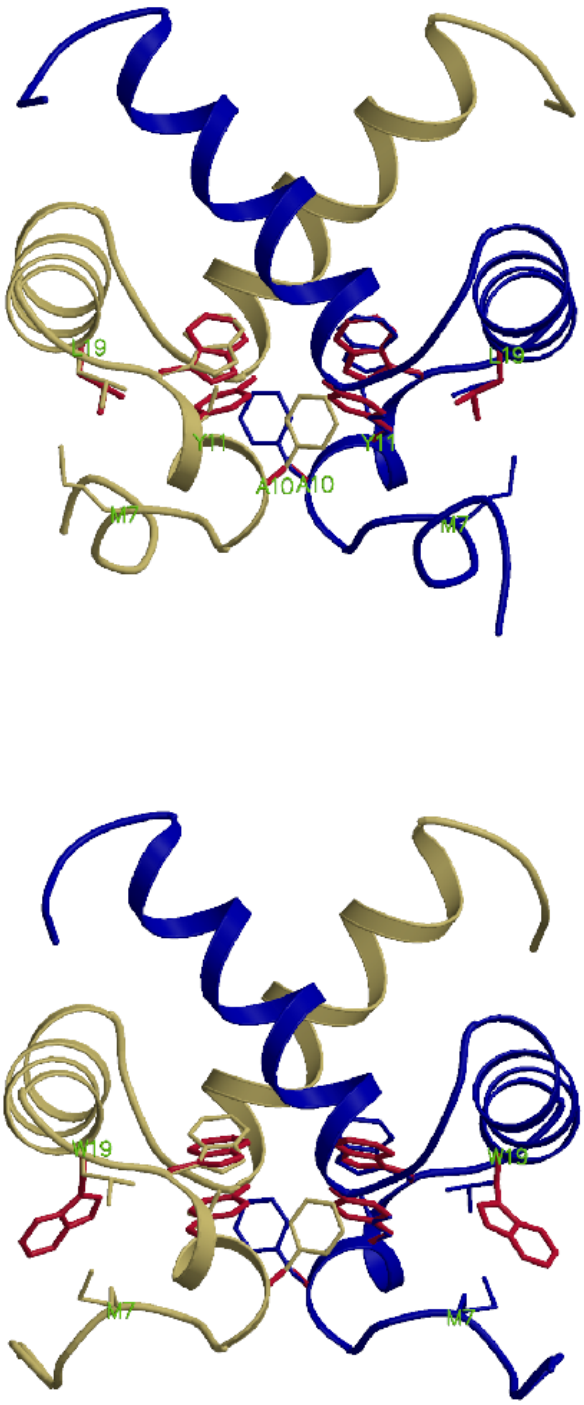


Figure 6.13: (top) The predicted structure of switch Arc with the sequence AYWL at positions 10, 11, 14, and 19. The backbone of NMR structure #3 of switch is shown in blue and yellow. The predicted side chain placements of the new sequence are shown in red. (bottom) The predicted structure of switch Arc with the sequence AYWW in the background of NMR structure #12.

Chapter 7

Repacking a Protein Interface: Application to Zif268 Zinc Finger

7.1 Introduction

Accurate prediction of the atomic packing in proteins is an important goal in understanding protein function. Several computational methods including dead-end elimination (DEE) [32, 48] and A* (or branch-and-bound) [52, 78] search algorithms have been used recently to repack the side chains of proteins. In these methods, the conformational search space is represented as a discrete set of side chain rotamers [34, 111] which allow the searches to proceed rapidly over a large number of conformations. These search methods have been applied to several problems including the prediction of packing in a known structure [55], the design of novel packing in a known structure [29, 86, 128], and the design of a novel structure [54]. These methods have mainly been applied to hydrophobic cores of proteins, where the effects of solvent and electrostatics appear to be unimportant.

We would like to extend the use of repacking methods to protein binding interfaces, where electrostatic effects typically play a substantial role. Continuum models of aqueous solvent have proved useful in understanding these effects [61]. Continuum models have been applied to understanding electrostatic effects in protein folding [59, 148, 149] and protein binding [60, 121], and as a basis for designing tight-binding

ligands [69, 79]. Although the continuum solvent is a way of accounting for different arrangements of the solvent, it is too computationally intensive to permit evaluations of many different protein conformations. Analytical methods such as the generalized Born model [114, 134] and the ACE model [119] have been developed to approximate continuum electrostatic calculations at a much lower cost.

A complete understanding of protein systems would include a treatment of both side chain packing and electrostatics. In this work, we attempt to combine the search capabilities of the DEE and A* algorithms with electrostatic effects treated by a continuum solvent model. The use of DEE/A* is limited to certain types of energy functions. For DEE, the energy must be pairwise additive, and although A* does not explicitly require the energy to be pairwise, the search can be made more efficient if the energy is pairwise. Continuum electrostatic energy functions do not give energies that are pairwise additive. The basic approach will be to use a low-resolution energy function to identify many possible candidate structures, which will then be evaluated by a higher-resolution energy function that fully accounts for solvation effects.

We will use the approach to dock the Zif268 zinc finger to DNA. Zif268 is a transcription factor whose structure has been determined [38, 107] at high resolution. Several sequence variants of this protein that were selected for high binding affinity [117] have also been studied crystallographically [36]. Computational methods will be used to attempt reproduce the docked structure of the zinc finger–DNA complex and a complex with a zinc finger variant. The results give some insight on the reason that the zinc finger side chains pack the way they do in the crystal structure.

7.2 Methods

Structure Preparation. The calcluations were carried out using the crystal structure of Zif268 bound to DNA from the Protein Data Bank [38]. (pdb code 1aay) and the structure of the RADR mutant bound to the same DNA sequence [36] (pdb code 1a1j). Hydrogens were added using the HBUILD facility of CHARMM [19].

Analytical Continuum Electrostatics. The algorithm for analytical contin-

uum electrostatics (ACE) [119] was used as implemented in CHARMM version 27. There are no parameters for the atomic volumes of DNA atoms, so the parameters were fit by means of comparison to a continuum electrostatic calculation using the program Delphi [46, 47, 126]. First, the protein was divided into groups as follows. Each amino acid was divided into a side chain and neutral amino (consisting of N, H, and C_α) and carbonyl groups. The DNA was divided into phosphate (consisting of each P atom and the four oxygens bonded to it) and neutral base and ribose groups. The contribution of each group to binding in the wild-type Zif268–DNA complex was computed.

Each group pair has an interaction energy $\Delta G_{\mathrm{i,j;int}}$ defined to be the difference in interaction between two groups in the bound and unbound states of the complex. Each group has a solvation $\Delta\Delta G_{\mathrm{solv}}$ for the difference in self energy of a group between bound and unbound state, and a total interaction with all other groups $\Delta G_{\mathrm{i;int}}$. The contribution $\Delta\Delta G_{\mathrm{contrib}}$ is defined as a group's solvation plus one-half of its interactions so that the sum of all the contributions is the total binding energy of the complex. The mutation energy of a group $\Delta\Delta G_{\mathrm{mut}}$ is equal to the sum of a group's full interactions and its solvation. When ACE is used, we can find the RMS deviation between each of these terms and the more accurate values computed by solving the Poisson–Boltzmann equation. The ACE parameters were varied such that the product of the RMS deviations of all five of the above terms [rmsd(pair interaction), rmsd(solvation), rmsd(total interaction), rmsd(contribution), rmsd(mutation)] was minimized.

We used a procedure to search a many-dimensional parameter space for parameter sets that minimize any desired error function. Minimization of a nonlinear function with a many-dimensional parameter space is an inherently difficult problem. The downhill simplex method of Nelder and Mead with simulated annealing added as by Press [112] was used to minimize the above objective function. The parameters (α and 42 effective atom volumes $\tilde{V}$ as defined in ACE [119]) are restricted to non-negative values. The optimized values of the RMS deviations were: rmsd($\Delta G_{\mathrm{i,j;int}}$) = 0.06 kcal/mol, rmsd($\Delta\Delta G_{\mathrm{solv}}$) = 0.30 kcal/mol, rmsd($\Delta G_{\mathrm{i;int}}$) = 0.45 kcal/mol,

rmsd($\Delta\Delta G_{contrib}$) = 0.31 kcal/mol, rmsd($\Delta\Delta G_{mut}$) = 0.44 kcal/mol.

DNA–protein orientation. The double-helical axis of the DNA was determined using the program Curves [76, 77, 116] so that it was linear. The linear axis of the alpha helix of zinc finger 1 was determined using the program P-curves [132]. The helical axes of the intact complex were rotated and translated so that the reference point on the α-helix axis (point on the axis closest to His 25 C_α) was at the origin. The intact complex was then rotated about the origin so that the DNA axis crossed the y-axis and was parallel to the z-axis. This procedure uniquely defines the orientation of the intact protein–DNA complex.

The DNA and protein could be rotated and translated individually from their reference positions to define the parameters of orientation of the complex. The reference point on the DNA axis was defined as the point closest to the midpoint of the C1' atoms of the first base pair recognized by the zinc finger (base pair 8 for finger 1). The DNA reference position was defined as placing the reference point on the at the origin so that its axis (from 5' to 3') aligns with the $+z$-axis. The C1' connecting line was oriented so that it pointed in the $+x$ direction and the major groove pointed in the $+y$ direction. The DNA can be moved from its reference orientation to its orientation in the complex by rotation about the z axis by and angle α followed by translations in y and z. The required rotation and translations define the three parameters listed in Table 7.2. The α-helix reference position is defined as placing the helix reference point at the origin, pointing the helical axis so that it is aligned with the z-axis of a coordinate system, and pointing Cα of His 25 along the $+y$ axis. Rotation by Euler angles ϕ, θ, and ψ places the helix as it is oriented in the complex. These six parameters therefore uniquely describe the orientations of the α-helix and the DNA.

In order to perform principal component analysis, each of the parameters had to be normalized to the same dimensions. The Zif268–DNA complex crystal structure was modified by changing each of the angles by increments of 0.5 degrees, then observing the effect on the Cα RMS deviation of the backbone atoms in zinc finger 1. The following results were obtained for each of the four angles: ϕ, 0.128 Å/degree; θ,

0.121 Å/degree; ψ, 0.155 Å/degree; α, 0.222 Å/degree. The angles for each of the α-helices were multiplied by these factors before principal component analysis was performed on the data.

Dead-End Elimination and A*. The Dunbrack and Karplus [34] rotamer library was expanded as described in Results. The total number of rigid rotamers available for each residue type was: Arg, 6561; Glu, 729; Asp, 81; Ser, 81; Thr, 81; Ala, 1. For each conformational search, the protein was kept fixed and the DNA was moved to account for the available protein–DNA orientations. Van der Waals and electrostatic energies were computed using the CHARMM PARAM19 parameters [18] for protein and an experimental set for DNA [146].

For purposes of the conformational search, the DNA orientations were treated just as another side chain—i.e. DNA positions could be eliminated by DEE just as rotamers are. First the Goldstein singles DEE criterion [32, 48] was applied at every variable position repeatedly until no more rotamers could be eliminated. Then split DEE [84, 110] was applied to all positions with all other positions serving as splitting positions. (See Chapter 5 for additional explanation of all these methods.) Again, the criterion was applied iteratively until no further rotamers could be eliminated. Finally, elimination of rotamer pairs was performed using the "magic bullet" pair as defined by Gordon and Mayo [51]. The system was checked to determine if elimination of pairs resulted in elimination of single rotamers [75]. This set of DEE criteria was repeatedly applied from the beginning until no further rotamers could be eliminated.

The remaining rotamer states were searched using the A* algorithm [78] with a modified bound. The A* search found the lowest energy state. To find all states within 15 kcal/mol of the minimum, the search was modified to traverse the tree without keeping all nodes in memory. This depth-first A* search is just as efficient as the original version when the best energy of the system is known.

Continuum Electrostatic Calculations. All Poisson–Boltzmann electrostatic calculations were carried out using a locally modified version of the program DELPHI [46, 47, 126]. The interior dielectric constant was set to 4 and the exterior dielectric to 80. The Poisson equation was solved with an ionic strength of 0.0 (to match the

ACE model). Each calculation of the potential in the bound and unbound states was repeated with ten different translations of the grid with respect to the molecule (i.e., the molecular geometry was unchanged). Contributions from individual groups were calculated on a 129x129x129 grid with three levels of focussing: 23%, 92%, and overfocussing at 184%. All complexes that were compared were placed on the same grid, where the final grid spacing was 4.2 grids/Å. Short to medium range interactions between groups were computed from the overfocussed grid. Longer range interactions were computed using the finest grid upon which both groups fit. The electrostatic energy of the entire complex was determined by calculation on a 257x257x257 grid.

$E_{\mathrm{h}\phi}$ was determined using the solvent accessible surface area computed in CHARMM. A coefficient of 25 cal/mol/Å^2 was used [125].

7.3 Results and Discussion

7.3.1 Rotamer search

The structure of zinc finger 1 bound to DNA is shown in Figure 7.1. In this work, we will repack the side chains of finger 1 that have any atom within 5.0 Å of the DNA. There are 10 such side chains, shown in the figure. The four labelled side chains contact the DNA bases, and they are Arg 18, Asp 20, Glu 21, and Arg 24. These side chains contact bases 8, 9, and 10 of the unprimed strand of the DNA. The six other side chains shown in the figure are Arg 3, Arg 14, Ser 17, which are near the phosphate backbone of the unprimed strand of DNA, and Ser 19, Thr 23, and Arg 27, which contact the primed strand of DNA.

These side chains were placed in the Zif268–DNA complex structure using the discrete set of rotamers given in the Dunbrack and Karplus [34] rotamer library. The conformations are evaluated with a simple energy function

$$E_{\mathrm{low}} = E_{\mathrm{vdw}} + E_{\mathrm{elec,4r}} \tag{7.1}$$

The energy function consists of a van der Waals term and an electrostatic interaction

term in which coulombic interactions between atoms are scaled by $4r$, where r is the interatomic distance in Å. This low-resolution energy function is pairwise as required for the conformational search. The lowest energy structures are identified by means of a search using dead-end elimination (DEE) and A*. The best structures are then reevaluated with a higher resolution energy function that better accounts for solvation effects:

$$E_{\text{high}} = E_{\text{vdw}} + E_{\text{h}\phi} + E_{\text{elec,ace}} \tag{7.2}$$

The van der Waals term in this equation is exactly the same as in the expression for E_{low}. $E_{\text{h}\phi}$ is a term favoring burial of hydrophobic surface, and $E_{\text{elec,ace}}$ is an approximate analytical continuum electrostatic function [119] that accounts for electrostatic interactions of solute and solvent.

When the standard rotamer library is used, one can see immediately that the set of rotamers is not sufficient to represent the possible arrangement of side chains in the protein. One example of such a problem is illustrated in Figure 7.2. The crystal structure conformation of Arg 18 and Asp 20 is shown with gray carbons. The rotamer of Arg 18 that is closest to the crystal structure rotamer is shown in red. This rotamer is turned so that the guanidinium group is not in the same plane as the Gua 10 base, which forms two hydrogen bonds with Arg 18. This red library rotamer has a steric clash with Cyt 9 (further into the figure, but not shown for clarity) and is therefore not accessible. The rotamer chosen in the minimum energy structure is shown in green. It makes only one hydrogen bond with Gua 10, and its new position forces Asp 20 to change conformation as well.

The rotamer library must be augmented in order for the correct structure to be a possibility. Small variations about side chain dihedral angles were added to the Dunbrack & Karplus library. The different sets of variations and their affect on the stability of the structure are shown in Table 7.1. The structure was forced to be a conformational state where the Dunbrack & Karplus rotamers were closest to the rotamers in the actual crystal structure. The value of E_{low} was minimized with each side chain allowed to occupy either the library rotamer or one of its subrotamers with

the indicated deviations from the library. The first set of subrotamers (second line of Table 7.1) consisted of variations of both χ_1 and χ_2 by $\pm 10^\circ$. This degree of flexibility reduces the van der Waals strain in the molecule substantially, from a structure that is over 3000 kcal/mol higher in E_{vdw} than the crystal structure to a structure that is 24.96 kcal/mol above the crystal structure. The key residues and residue pairs that create strain in the repacked structures are shown in the four middle columns of the table. Arg 18, pictured in Figure 7.2, contributes to van der Waals clashes with both the DNA and Asp 20. The energies of Arg 27 interacting with Arg 24 and the protein backbone also benefit from allowing extra flexibility in the rotamer library.

In subsequent rows of the table, additional degrees of freedom of the side chains are varied. The third row shows results for introducing a variation in covalent geometry, which for this complex is less successful at reducing strain than additional variation of side chain dihedral angles. When χ_3 is varied by $\pm$ 10, 20, 30, or 40 degrees, the strain is reduced substantially more than when these angles have only their library values. Allowing all these variations at once improves the energy slightly over simply varying χ_3 by $\pm$ 20° or 30°. Variation at χ_4 also reduces the strain in the molecule, particularly in Arg 18. Allowing 81 subrotamers for each side chain rotamer creates a large number of states to search. In particular, there are 81 library rotamers for Arg side chains, so 81 fine variations of each rotamer gives $81 \times 81 = 6561$ total rotamers of Arg. Any additional rotamers would make repacking calculation too computationally expensive. From the table, it appears that the best reduction in strain comes from varying each of the four χ angles separately.

The minimum energy structure with extra flexibility at all 4 variable χ angles is shown in Figure 7.3. Here, because of the additional subrotamers, Arg 18 can be positioned to make two hydrogen bonds with Gua 10, just as in the crystal structure. Once the conformational search allows the possiblity of this rotamer, the energy of the correct conformation is found to be minimum. Note that once Arg 18 is able to find its correct conformation, Asp 20 also goes to the correct rotamer, as do the other side chains in finger 1.

7.3.2 Docking of the DNA to the protein

In order to allow a full range of conformations of the protein–DNA complex to be explored, it will be important to allow the DNA to move relative to the protein. Motion of the DNA relative to the protein along with variation of side chain conformations can create a large number of conformations. In order to keep the size of the conformational space as small as possible, we sought to reduce the possible DNA–protein orientations to a set that is known to be feasible from the solved structures of zinc fingers bound to DNA. There are eight structures of finger 1 variants bound to DNA, each with different protein and/or DNA sequences [36]. The eight structures are shown in Figure 7.4 with their DNA backbones aligned so that the variation in the orientation of finger 1 can be seen. We wish to reproduce the variation in the repacking calculations.

First, a system of parameters to describe these different orientations was established. For rigid translations and rotations of the DNA relative to the protein, there are six degrees of freedom. We determine the α-helical axis and define a reference point on the axis that is closest to the C_α of His 25, the conserved His that coordinates a zinc ion in the zinc finger. The reference position of the α-helix will be defined as placing the reference point at the origin, pointing the helical axis so that it is aligned with the z-axis of a coordinate system, and pointing Cα of His 25 along the $+y$ axis. Similarly, the reference position for the DNA will have it placed so that its axis (from 5' to 3') aligns with the $+z$-axis. A line connecting the C1' atoms of base pair 8 (the first base pair contacted by the zinc finger) points in the $+x$ direction so that the major groove is approximately in the $+y$ direction. The complex is then produced from these reference orientations as follows. The α-helix is rotated in place to its proper orientation as follows. It is turned about its axis an angle ϕ, then about the x-axis by an angle θ, then turned about the z-axis again by an angle ψ. Then the DNA is rotated about its axis an angle α, and translated in the y and z directions to produce the final zinc finger complex.

The definition of such a system of parameters is somewhat arbitrary, but it gives

six parameters that uniquely describe the orientation of a zinc finger relative to a DNA double helix. The values of these parameters were computed for the eight different fingers in the variant Zif268 complexes. The results are shown in Table 7.2. There is significant variation in some parameters; for instance, the angle ϕ has a range of nearly 20°, but the range of orientations seen here is much smaller than the full range of possible orientations.

Given the parameters for the structures that we have, it may be possible to come up with a system of fewer than 6 parameters to describe the motion available to the zinc finger and DNA backbones. Here we make use of principal component analysis (PCA) to reduce the dimensionality of the data shown in Table 7.2. The sets of parameters are treated as six-dimensional vectors in a space where the origin represents the average value of each parameter. PCA yields a set of vectors in this space such that the first vector (or principal component) represents a combination of rotations and translations that best accounts for the variance in the data. The second principal component is orthogonal to the first, and its direction is selected to account for the most remaining variance in the data. The remaining principal components account for the remainder of the data.

When applying PCA to the data in Table 7.2, we must be aware that some parameters have units of degrees, and others have units of Ångstroms. These parameters were compared by determining the RMS deviation of the zinc finger 1 C_α atoms produced by changing each parameter. For example, changing the value of ϕ by 1° resulted in a C_α RMS deviation of 0.128 Å, so the variations in ϕ were weighted by 0.128 to make each of the dimensions have comparable units. PCA yielded a set of vectors that described the data, and the representation of the data in terms of the principal components is shown as blue X's in Figure 7.5. In the first plot the data is shown as it is projected on each of the first two principal components. The variance of the data in these two dimensions is clearly greater than in the third and fourth dimension, which are illustrated in the middle plot of Figure 7.5. Taken together, these four dimensions account for more than 99% of the variance of the eight structures. As seen in the bottom plot, all eight structures are clustered fairly

tightly together in the fifth and sixth dimension.

The data show that the known orientations of the DNA and protein backbone are described well by a four-dimensional space. We use the four-dimensional space to produce a grid of orientations, which will be searched during repacking of the protein–DNA interface. The set of orientations is illustrated by the red dots in Figure 7.5. A grid was chosen so that it surrounded all of the known data points for finger 1. The fineness of the grid was adjusted to make the search computationally feasible. The set of orientations used for the repacking of Zif268 was $15 \times 8 \times 6 \times 4$, giving a total of 2880 DNA–protein orientations.

7.3.3 Docked structure of Zif268–DNA complex

Given the discrete set of DNA–protein orientations, we may use a DEE/A* search as before to find the lowest energy conformations of the system. For purposes of the search, the DNA is treated just as if it were another side chain—some orientations can be removed by dead-end elimination, and the rest may be searched efficiently by the A* algorithm. With the augmented set of rotamers and the ability of the DNA and protein to change their relative orientation, the repacking calculation becomes prohibitively expensive. However, we may take advantage of the fact that the protein side chain rotamers consist of groups of rotamers that differ by relatively fine adjustments. A group of closely spaced rotamers, or subrotamers, can be considered as a single flexible rotamer in a method described by Mendes et al. [90]. (The flexible rotamer method is described more completely in Chapter 5.) The system is described in terms of interactions of flexible rotamers (or fleximers), in which each side chain can occupy one of several states in order to make fine adjustments in its position. In these calculations, an arginine fleximer consists of 81 subrotamers. The arginine side chain can adopt whichever of these subrotamers is best when interacting with another side chain. The interaction between one of its fleximers and another fleximer consists of the energetic benefit of the interaction of two subrotamers plus the cost of moving the arginine from its best subrotamer in isolation.

The fleximer model is an approximation that reduces the search space dramati-

cally. Instead of 6561 arginine rotamers at each position, there are 81 fleximers. In this complex, the fleximer approximation reduces the size of the search space from 2.8×10^{33} total rigid rotamer conformations to 1.1×10^{19} fleximer conformations of the system. The drawback of the fleximer approximation is that the total energy of interaction of all the fleximers (the fleximer energy F) can be artificially low due to a side chain's ability to adopt two different subrotamer states when interacting with two different positions. To correct for this approximation, once several fleximer states of the system with a low F are identified, each of the fleximers corresponding to the state is frozen into rigid subrotamer conformation that will minimize the energy of the entire molecule. Some fleximer conformations will give a high energy when frozen, because the low fleximer energy depends on some side chains being able to occupy two subrotamer states at the same time. Others will have an energy E close to or even less than the approximate fleximer energy.

The DEE/A* search was applied to the Zif268–DNA complex using the fleximer library and set of DNA–protein orientations described above. The energy function E_{low} in equation 7.1 was used to evalulate the energies. Conformations with a value of F within 10 kcal/mol of the best fleximer energy were identified. This procedure gave over 2.2×10^5 fleximer states. Each of these fleximer states was frozen to give a rigid rotamer state of the system as described. A plot of energy of the frozen conformation vs. the fleximer energy is shown in Figure 7.6. In this figure, the fleximer energy of the best fleximer state is defined as having an energy of 0. The plot illustrates that in many cases, E of the frozen conformation is substantially higher than F. The black line in the plot represents $F = E$. We see that in relatively rare cases, the frozen energy is lower than the fleximer energy. The greatest margin by which F exceeds E is 1.4 kcal/mol in this system. We can therefore be reasonably confident that no other conformational states exist with a lower value of E than those discovered here.

Of the points plotted in Figure 7.6, only 25692 have a value of E less than 10 kcal/mol. (Many have a high enough value of E that they do not appear in the plot.) There are relatively few distinguishing features in the plot—the energetic cost of freezing the fleximers have just about any value depending which fleximers are

involved. The few features that can be seen are sets of points that fall approximately on a diagonal line near the edges of the plot. Two such clusters are highlighted in green and red. All of the green points in the plot have the same fleximer states of all Arg residues and the DNA position. The only differences between them are the positions of Ser 17, Ser 19, and Thr 23. Because these three residues do not interact very strongly with other side chains, they can occupy one subrotamer throughout the fleximer approximation. The energy cost of freezing the fleximers is determined by the states of all the other side chains. Since all green points have the same fleximers at these other positions, their cost of freezing the conformation ($F-E$) is approximately constant, and all the green points fall near a line parallel to $F = E$. The red points in Figure 7.6 are a different example of this; all of the red points again have the same fleximer states of the arginine side chains and the DNA, but the states are different than those of the green points. For the red points also, changing the fleximer state of a serine or threonine may change the total energy, but it does not significantly affect the cost of freezing the fleximers, so these points also fall on a diagonal line. These particular sets of points do not end up being among those with the lowest values of E_{high}. They are highlighted to illustrate the reason for the feature in the plot—not because they necessarily represent the best structures.

The structures with a favorable low-resolution energy (E_{low}) were reevaluated using the high-resolution function E_{high} which accounts for solvation effects. The plot of E_{high} vs. E_{low} for the docking of this complex is shown in Figure 7.7. All values of E_{high} and E_{low} have the lowest values of E_{high} and E_{low} subtracted from them, so the lowest value of each is 0 in the plot. There is not an especially strong correlation between E_{high} and E_{low}, which is unsurprising. The important feature of Figure 7.7 for the purpose here is that there are no points in the lower right corner, as this would indicate that E_{low} could be high for a conformation whose true energy is favorable. When E_{low} is substantially higher than its minimum value it is usually due in part to less favorable van der Waals packing, which is part of the E_{high} function as well as E_{low}.

The structure with the most favorable value of E_{high} is shown in Figure 7.8. The

structure shows good agreement with the crystal structure. Arg 18 and 24 each occupy the fleximer closest to the crystal structure rotamer, and they make hydrogen bonds with Gua 8 and 10, respectively. Asp 20 and Glu 21 also occupy rotamers closest to the correct rotamers, and make contacts similar to those in the crystal structure. Arg 27, in the lower right of the figure, comes within 4.2 Å of phosphate 11', as opposed to 4.6 Å in the crystal structure. Although Arg 27 is not in the correct rotamer conformation, it appears to be flipped in a way that it occupies nearly the same region of space. The two other side chains in the figure, Ser 19 and Thr 23, have a different conformation than in the crystal structure. The DNA moves closer to the zinc finger, and the methyl group of Thy 12' (not shown in figure) forces Ser 19 to move to a different conformation. Three other side chains that were mobile in the repacking calculation are shown with a view from the opposite side of the complex in Figure 7.9. Ser 17, Arg 3 and Arg 14 each occupy the correct rotamer conformation. The two arginines make contacts with phosphates 7 and 8 as in the crystal structure, and the O_γ of Ser 17 is 4.3 Å from phosphate 8. Thr 23 makes a hydrogen bond with O3' of ribose 12', which is closer in the repacked structure than in the crystal structure (3.3 Å vs. 4.1 Å).

Other structures are seen to be similar in energy to the minimum energy structure. There were 311 structures with an energy within 2 kcal/mol of the best structure. These structures include 9 different DNA orientations, 2 different fleximers of Arg 14, 5 fleximers of Ser 17, 3 fleximers of Ser 19, 3 fleximers of Thr 23, 2 fleximers of Arg 24, and 3 fleximers of Arg 27. Most of these variations are relatively minor and do not affect which groups contact one another across the interface. For example, Ser 19 and Thr 23 adopt different conformations by simply rotating their hydroxyl protons to three different rotamer states. The most significant change is the variation of Arg 24, which adopts a conformation where it still makes one hydrogen bond with Gua 8 (it makes two in the correct structure) and makes a second hydrogen bond with the carbonyl group of Gua 9' in the other strand of DNA. In the lowest energy structure with this rotamer, all of the other side chains and DNA stay in exactly the same conformation. This structure, which we label conformation (2), is shown in

Figure 7.10. The reason for the difference in energy between this conformation and the minimum energy conformation (conformation (1)) is primarily electrostatic, as shown in Table 7.3. (Conformation (3) will be discussed below.) For a few conformations of the complex, the total electrostatic energy was recomputed by solving the Poisson–Boltzmann (PB) equation numerically, an approach which is more accurate than ACE. Since solving the Poisson–Boltzmann equation is more computationally demanding, it could not be used on nearly as many conformations. The total energy with this term is

$$E_{\mathrm{tot,PB}} = E_{\mathrm{vdw}} + E_{\mathrm{h}\phi} + E_{\mathrm{elec,PB}} \tag{7.3}$$

This is the same energy as E_{high}, only with ACE energy replaced by PB electrostatic energy.

When the PB electrostatic energy is calculated, we find that conformation (1) is favored by 1.1 kcal/mol over conformation (2). The reason for the difference in energy is summarized in the second portion of the table, where the contributions of a few chemical groups to $E_{\mathrm{elec,PB}}$ are listed. The complex was divided into groups as follows. The protein backbone consists of amino and carbonyl groups, and each protein side chain is treated as a separate group. The DNA was divided into phosphate (including the O5' and O3'), ribose, and base groups. All groups have an integer charge. The $\Delta\Delta G_{\mathrm{solv}}$ for a group is the cost of moving the group from solvent by itself into the position it occupies in the complex. $\Delta\Delta G_{\mathrm{int}}$ is the solvent-screened interaction between two groups in the complex. The second part of Table 7.3 shows these terms for the local environment of Arg 24, the residue that occupies different rotamers in conformations (1) and (2). In conformation (2), Arg 14 is slightly more solvent exposed, and thus has a slightly smaller desolvation penalty. However, the interactions it makes with DNA (primarily bases Gua 10 and Gua 9') are weaker in conformation (2) because the hydrogen-bond geometry is not as good. In addition, conformation (2) brings Arg 24 closer to Arg 27, which decreases the stability of (2) by a little over 1 kcal/mol. There are other small differences in screening of electrostatics caused by the conformational difference, but the net result is that conformation (1)

is still favored by 1.4 kcal/mol.

7.3.4 Repacking of RADR mutant Zif268–DNA complex

The docking/repacking procedure was applied to the RADR mutant of Zif268 in complex with the same DNA. The sequence RADR refers to the amino acids at positions 18, 20, 21, and 24 of Zif268 (the wild type is RDER). The RADR mutant was isolated from a selection for DNA binding using phage display [117], and its structure in complex with the Zif268 consensus DNA binding site has been solved. In this calculation, the backbone and DNA were taken from the wild-type Zif268–DNA complex. The structures of the protein and DNA are fairly well conserved throughout the known zinc finger structures [36], so this calcuation will test whether the structure is consistent enough that one backbone can be used to reproduce another complex.

The search for states with low fleximer energy yielded 2.1×10^5 states within 10 kcal/mol of the minimum. After E_{high} was computed for these states, there were 248 that were within 2 kcal/mol of the structure with the best E_{high}. These structures included 7 different DNA orientations, 3 different Arg 3 fleximers, 2 fleximers of Arg 14, 4 fleximers of Ser 17, 2 fleximers of Arg 18, 3 fleximers of Ser 19, 3 fleximers of Thr 23, and 3 fleximers of Arg 27. As in the repacked wild-type complex, most of these variations cause relatively little change in the contacts between residues. The serine and threonine rotamers are primarily simple changes in the position of the hydroxy hydrogen, and the arginine rotamers contact the same phosphates with different nitrogens of their guanidinium groups. The most substantial difference is the change in conformation of Arg 18, which can contact either the base of Gua 10 or the phosphate group of residue 9. The change in the conformation of this residue is accompanied by changes in the orientation of the DNA.

The lowest energy state, which we label conformation (1), is shown in Figure 7.11. This structure has a significantly different DNA–protein orientation than in the crystal structure, but the amino acid base contacts are still conserved. Arg 18 and Asp 20 both have two conformations in the crystal structure. Figure 7.11 shows the conformations labelled “B” because they are closer to the side chain positions in

conformation (1). Arg 18 makes two hydrogen bonds with Gua 10 in both the crystal structure and the repacked structure. Arg 18 occupies a different rotamer state in the two crystal structures because of the reorientation of the protein, but the guanidinium group is still positioned to make two hydrogen bonds. Arg 24 also occupies a different rotamer state, but is positioned to make two hydrogen bonds with Gua 8.

The conformation that is significantly different from (1) is labelled conformation (2) and is shown in Figure 7.12. This structure also has a significant difference in protein–DNA orientation from the crystal structure. Arg 18 occupies a rotamer in which it makes contact with phosphate 9 of the DNA. This conformation is similar to the alternate conformation "A" of Arg 18 in the crystal structure. This is the conformation shown with the purple backbone and gray carbons in the figure. After the electrostatic energy is corrected using PB electrostatics, conformation (2) is predicted to be only 0.4 kcal/mol higher in energy than conformation (1). Although the backbones of these conformations are placed differently than in the crystal structure, the small difference in energy between the two arrangements of Arg 18 may help explain why the two alternate conformations are observed.

The contributions of residues interacting with Arg 18 are shown in Table 7.4. These are the electrostatic effects in the local vicinity of Arg 18. Other contributions to the energy are different between the two conformations as well, but these terms illustrate the tradeoff faced by Arg 18. In conformation (1), it makes stronger interactions via hydrogen bonds to Gua 10, but conformation (2) includes stronger electrostatic interactions with phosphate 9 and Asp 21. Conformation (2) also brings Arg 18 closer to Arg 3, which disfavors this conformation slightly. The movement of Arg 18 to conformation (2) exposes it more to solvent, but desolvates phosphate 9. The net effect of the components listed in Table 7.4 is to favor conformation (1) only slightly. Thus, the result that Arg 18 can occupy two conformations is consistent with the experimental results.

To further understand the reason for the alternate conformation of Arg 18, we reexamined the results for the wild-type zinc finger. The lowest energy structure of this complex with Arg 18 contacting phosphate 9 was determined from the repacking

calculation described above. This is conformation (3) in Table 7.3, and its total energy is predicted to be 3.8 kcal/mol higher than conformation (1) of this complex. The electrostatic energy of (3) is 2.6 kcal/mol higher than conformation (1). By looking at the components of the electrostatic energy, we can begin to understand why Arg 18 prefers to recognize the base in this complex, while it is capable of contacting either the base or the phosphate in the RADR complex. The most striking difference between the two sequences is the interaction between Arg 18 and residue 20. In the RADR complex, residue 20 is alanine, and thus there is no electrostatic interaction in either conformation. In the wild-type, the R18–D20 interaction favors conformation (1) by 5.4 kcal/mol. This preference is reduced because the desolvation penalty of Asp 20 is 2.5 kcal/mol larger in conformation (1) of wild-type. However, in the RADR mutant, there are no electrostatic interactions with Ala 20. Thus conformation (1) is not as strongly favored in this mutant. Residue 21 also contributes to the Arg 18–base contact preference. In the wild-type, Glu 21 pays 3.3 kcal/mol in solvation when conformation (3) is adopted and gets only -2.5 in additional interaction with Arg 18. In the RADR sequence, Asp 21 has a better tradeoff with Arg 18 when it approaches the phosphate. Asp 21 gets -2.2 kcal/mol in interaction and only pays 0.3 kcal/mol in solvation. The Asp at position 21 does not extend as far toward solvent, and is therefore more desolvated regardless of the position of Arg 18. There are additional contributions to the relative preference of Arg 18 conformation in the table. The total preference depends on subtle changes in the DNA–protein orientation as well as the direct interactions of residues 20 and 21. The repacking calculation helps us understand the reason for these conformational preferences and suggests that Asp 20 and Glu 21 may both contribute to the base specificity of the zinc finger, as suggested by Elrod-Erickson et al. [36]. When Arg 18 is held in contact with the base by these residues, it would be expected to recognize guanine at position 10. However, when residues 20 or 21 mutate, Arg 18 may be more likely to contact the phosphate, where it has weaker interactions with the base pair and presumably less specificity.

Conclusion. A method for docking a protein–DNA complex with side chain flexibility has been presented. The approach uses a discrete search of a library of side

chain rotamers and docking orientations to identify a number of candidates with a low-resolution energy function, then narrows the search down to a few candidates using a higher-resolution energy function that better accounts for solvation. The placement of the side chains in the repacked structures shows fairly good agreement with crystal structures, although the exact orientation of the protein backbone with respect to the DNA is not always predicted perfectly. The repacking calculation does appear to demonstrate the energetic basis for a known side chain conformational preference. In the future, the approach presented here may be extended for use in building and designing new protein sequences that fold well or bind tightly to a desired target.

Table 7.1: Difference in van der Waals energy between repacked structure nearest crystal structure and actual crystal structure

Subrotamers added	# rot[a]	R27–bb	DNA–R18	R18–D20	R24–R27	Total
Standard Library	1	70.5	64.9	8.0	2991.2	3144.0
$\chi_1, \chi_2 \pm 10$	9	3.1	9.9	2.4	10.1	25.0
$\chi_1, \chi_2 \pm 10$, angle[b]	27	3.5	7.9	2.1	9.0	22.7
$\chi_1, \chi_2, \chi_3 \pm 10$	27	2.6	9.9	2.4	1.7	15.0
$\chi_1, \chi_2 \pm 10$; $\chi_3 \pm 40$	27	1.2	4.0	3.6	−0.2	11.2
$\chi_1, \chi_2 \pm 10$; $\chi_3 \pm 30$	27	1.6	2.4	3.0	−0.2	6.2
$\chi_1, \chi_2 \pm 10$; $\chi_3 \pm 20$	27	1.9	2.3	2.3	−0.2	5.5
$\chi_1, \chi_2 \pm 10$; $\chi_3 \pm$ 10–40[c]	81	1.2	2.3	2.3	−0.3	3.2
$\chi_1, \chi_2 \pm 10$; $\chi_3, \chi_4 \pm 20$	81	1.5	0.7	1.2	−0.3	2.1
$\chi_1, \chi_2 \pm 10$; $\chi_3 \pm$ 10–40[c]; $\chi_4 \pm 20, 40$	405	1.2	0.7	1.2	−0.4	1.7

All free energy values are in kcal/mol.

[a] This column indicates the total number of rotamer variations allowed for Arg, a side chain with 4 variable dihedral angles. Other side chains may have correspondingly fewer rotamer variations.

[b] The improper dihedral angle placing C_β relative to the backbone was varied by 2.5° and 5.0°.

[c] A range of 10–40 indicates that the χ angle variations were 10, 20, 30, and 40 degrees.

Table 7.2: Parameters for zinc finger–DNA orientation in Zif268 variants.

Sequence[a]	ϕ	θ	ψ	α	y	z
GAC DSNR	152.69	−124.14	29.13	−144.22	10.71	−3.34
GCG DSNR	159.62	−132.13	34.60	−141.06	11.83	−3.21
GCA QGSR	158.05	−129.67	30.22	−142.31	11.84	−2.83
GCA RADR	170.99	−129.81	44.06	−137.10	12.05	−2.83
GCG RADR	168.60	−129.31	43.31	−137.30	11.72	−3.58
GAC RADR	165.25	−126.57	46.02	−133.99	11.83	−2.65
GCA RDER	171.36	−129.99	42.65	−136.93	12.05	−2.91
GCG RDER	165.57	−133.55	35.66	−139.69	12.04	−2.11
Average	164.02	−129.39	38.21	−139.07	11.76	−2.93
Std Dev.	6.66	2.95	6.62	3.35	0.44	0.45

Definitions of parameters are given in the text. Angles are in degrees and lengths are in Å.

[a] The variable parts of the sequences of the zinc finger. The first three letters represent the DNA sequence at bases 8, 9, and 10. The next four letters represent the one-letter amino acid codes for residues 18, 20, 21, and 24. The last entry in the table is the wild-type Zif268 complex.

Table 7.3: Contributions to stability of repacked Zif268–DNA complex.

Conformation	min energy	(1)	(2)	(3)
E_{vdw} [a]		0.0	−0.2	0.9
E_{surf} [a]		0.0	−0.1	0.3
E_{ace} [a]		0.0	0.7	2.0
E_{high} [a]		0.0	0.4	3.2
E_{PB} [a]		0.0	1.4	2.6
$E_{\mathrm{tot,PB}}$ [a]		0.0	1.1	3.8
Electrostatic components				
$\Delta\Delta G_{\mathrm{int}}$(R24–DNA)		−13.1	−11.7	−13.0
$\Delta\Delta G_{\mathrm{solv}}$(R24)		11.9	10.5	9.7
$\Delta\Delta G_{\mathrm{solv}}$(R24–R27)		2.3	3.4	2.1
$\Delta\Delta G_{\mathrm{int}}$(R18–R3)		0.6	0.6	1.6
$\Delta\Delta G_{\mathrm{int}}$(R18–D20)		−7.5	−7.5	−2.1
$\Delta\Delta G_{\mathrm{int}}$(R18–E21)		−3.3	−3.2	−5.8
$\Delta\Delta G_{\mathrm{int}}$(R18–Gua 10)		−5.4	−5.4	−0.5
$\Delta\Delta G_{\mathrm{int}}$(R18–phos 9)		−0.9	−0.9	−3.4
$\Delta\Delta G_{\mathrm{solv}}$(R18)		9.0	9.0	6.8
$\Delta\Delta G_{\mathrm{solv}}$(D20)		9.6	9.6	7.1
$\Delta\Delta G_{\mathrm{solv}}$(E21)		7.6	7.5	10.9
$\Delta\Delta G_{\mathrm{solv}}$(phos 9)		1.7	1.7	2.8

[a] The energy terms are shifted so that the minimum energy structure has a value of zero.

Table 7.4: Contributions to stability of repacked RADR mutant Zif268–DNA complex.

Conformation	min energy (1)	(2)
E_{vdw} [a]	0.0	0.9
E_{surf} [a]	0.0	−1.1
E_{ace} [a]	0.0	1.3
E_{high} [a]	0.0	1.1
E_{PB} [a]	0.0	0.6
$E_{\mathrm{tot,PB}}$ [a]	0.0	0.4
Electrostatic components		
$\Delta\Delta G_{\mathrm{int}}$(R18–R3)	0.7	2.0
$\Delta\Delta G_{\mathrm{int}}$(R18–A20)	0.0	0.0
$\Delta\Delta G_{\mathrm{int}}$(R18–D21)	−3.0	−5.2
$\Delta\Delta G_{\mathrm{int}}$(R18–Gua 10)	−4.8	−0.5
$\Delta\Delta G_{\mathrm{int}}$(R18–phos 9)	−0.9	−4.9
$\Delta\Delta G_{\mathrm{solv}}$(R3)	7.4	7.3
$\Delta\Delta G_{\mathrm{solv}}$(R18)	8.5	7.5
$\Delta\Delta G_{\mathrm{solv}}$(A20)	0.0	0.0
$\Delta\Delta G_{\mathrm{solv}}$(D21)	11.1	11.4
$\Delta\Delta G_{\mathrm{solv}}$(R24)	11.9	12.4
$\Delta\Delta G_{\mathrm{solv}}$(phos 9)	1.9	3.0

[a] The energy terms are shifted so that the minimum energy structure has a value of zero.

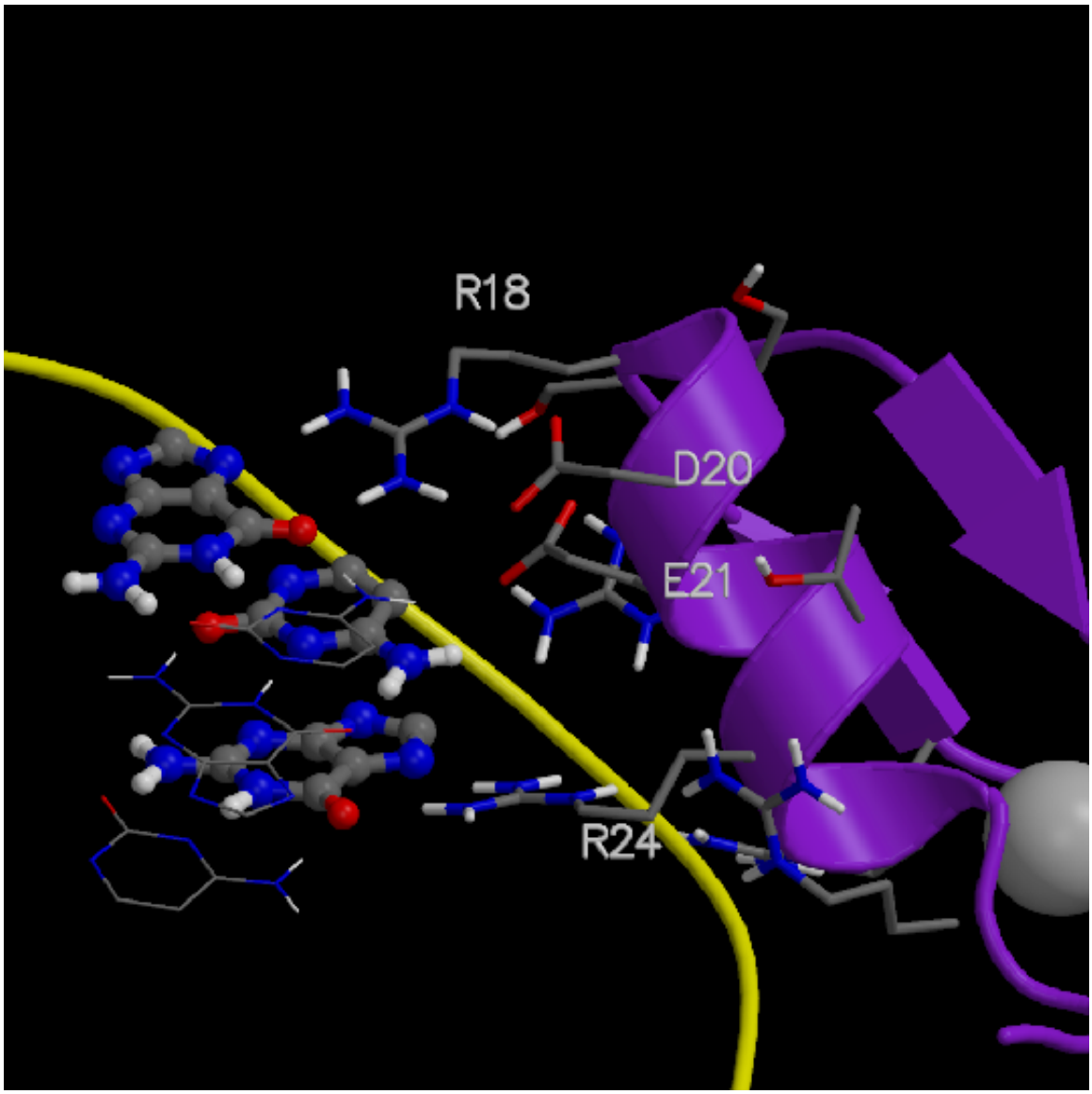


Figure 7.1: Finger 1 of the Zif268–DNA complex. The side chains that contact DNA bases and phosphates are shown. The three bases contacted by these side chains are shown with balls and sticks. The bases complementary to the contacted bases are shown with narrow bonds.

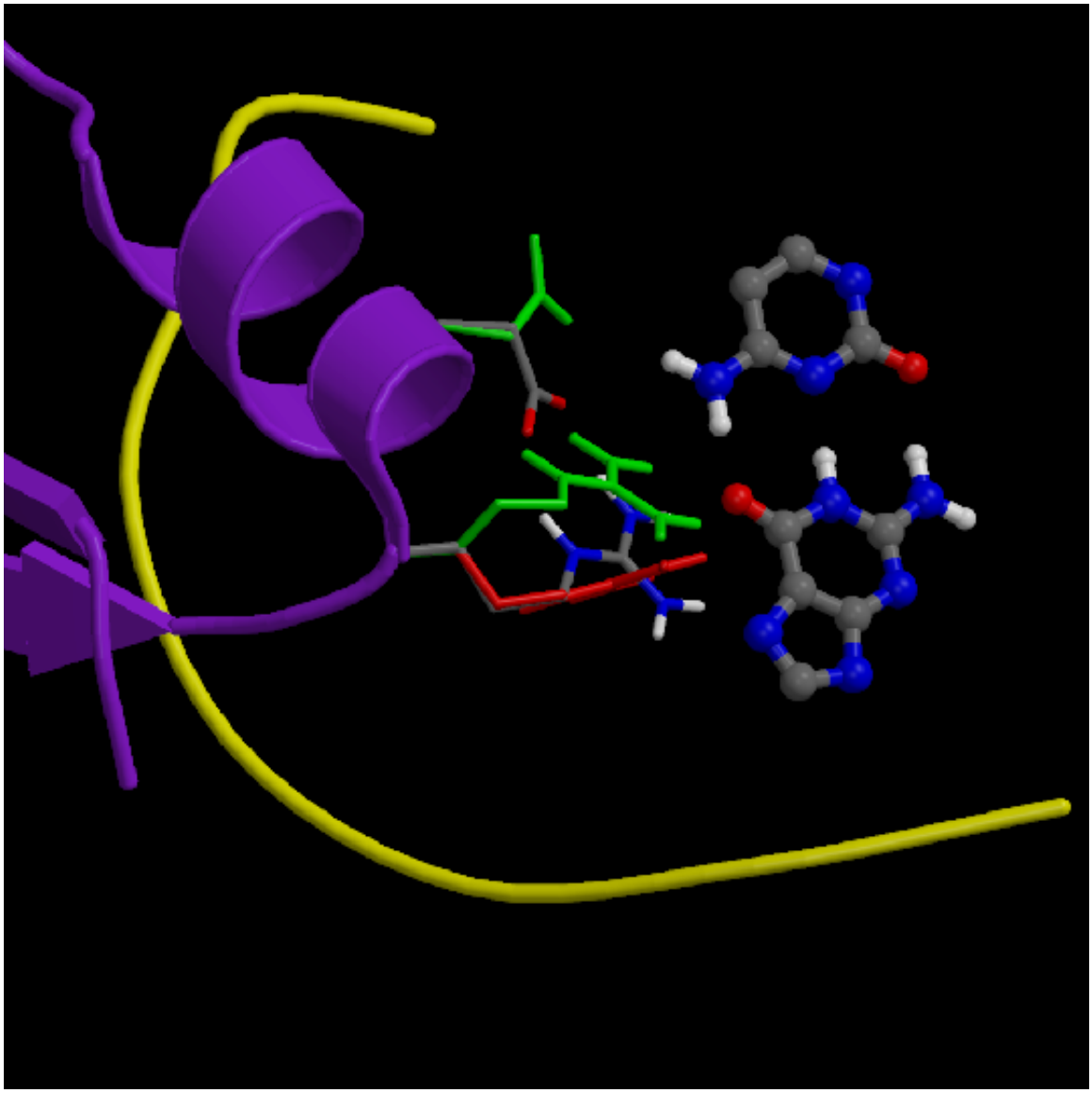


Figure 7.2: Side chains Arg 18 and Asp 20 across from base pair 10 in the Zif268–DNA complex. The crystal structure is shown with gray carbons. The rotamer in a standard library closest to the x-ray structure is shown in red. The minimum energy conformation of the two side chains is shown in green.

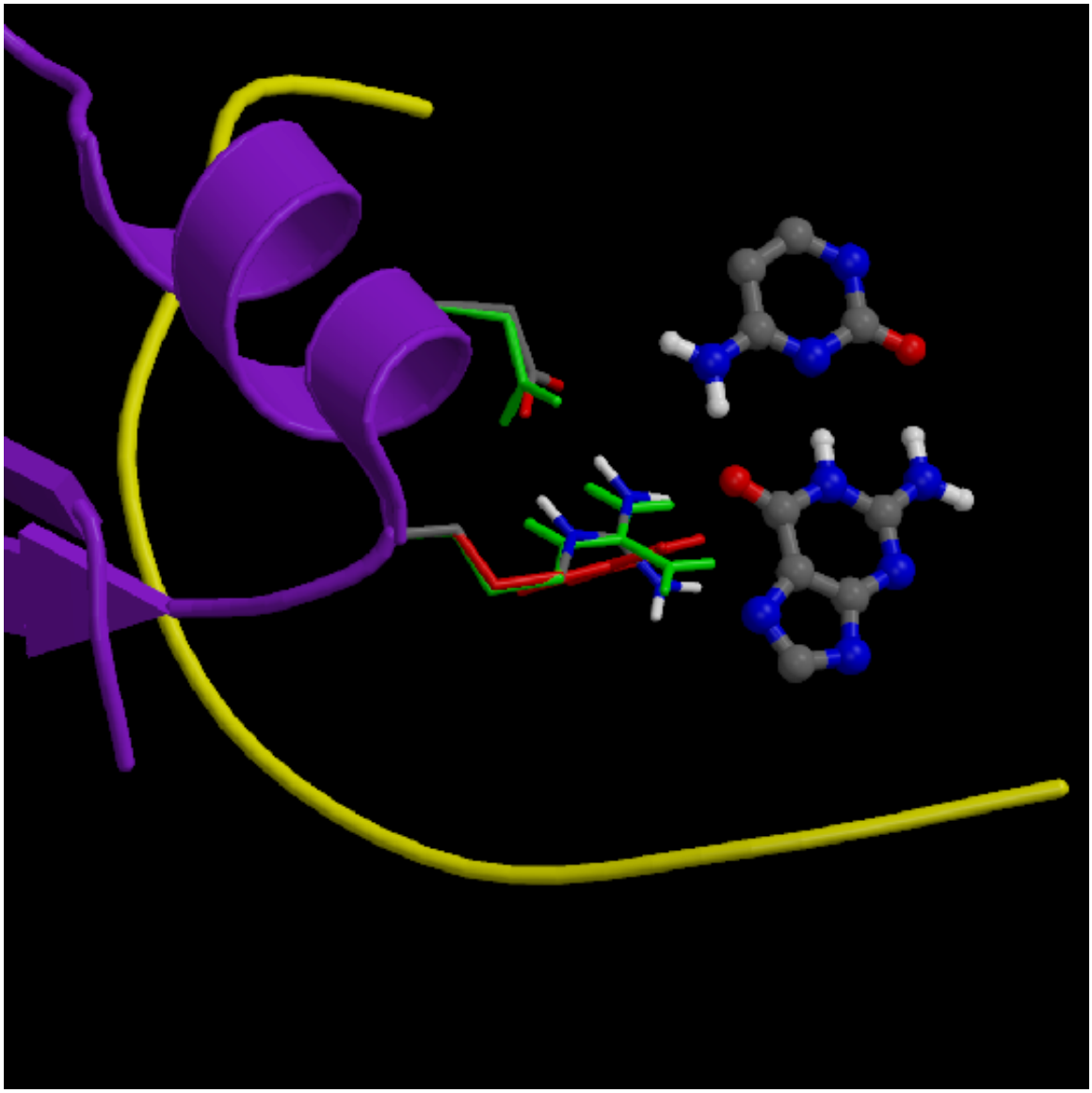


Figure 7.3: Side chains Arg 18 and Asp 20 across from base pair 10 in the Zif268–DNA complex after each rotamer is allowed multiple subrotamers. The minimum energy side chains are again shown in green.

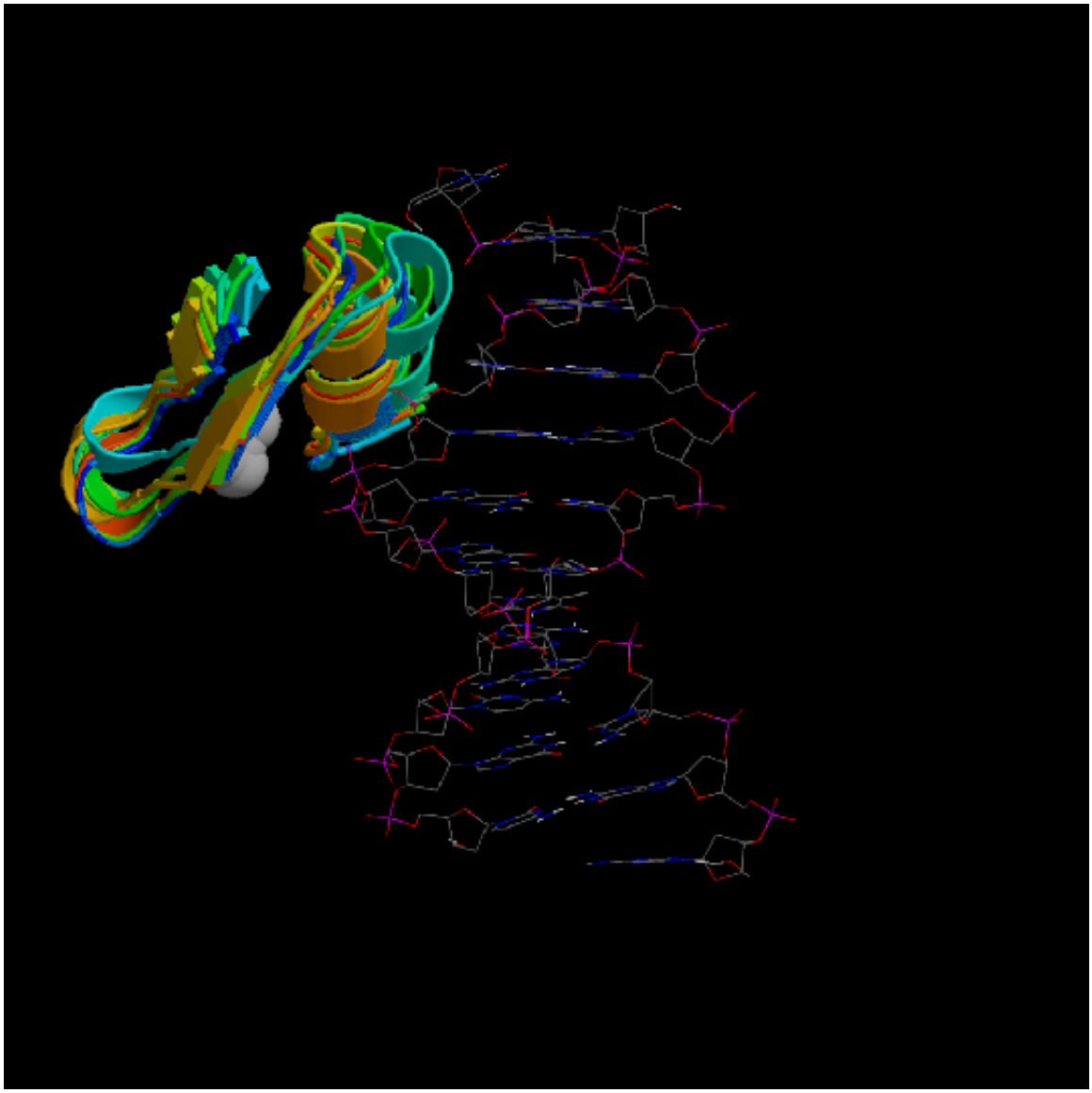


Figure 7.4: Side chains Arg 18 and Asp 20 across from base pair 10 in the Zif268–DNA complex after each rotamer is allowed multiple subrotamers. The minimum energy side chains are again shown in green.

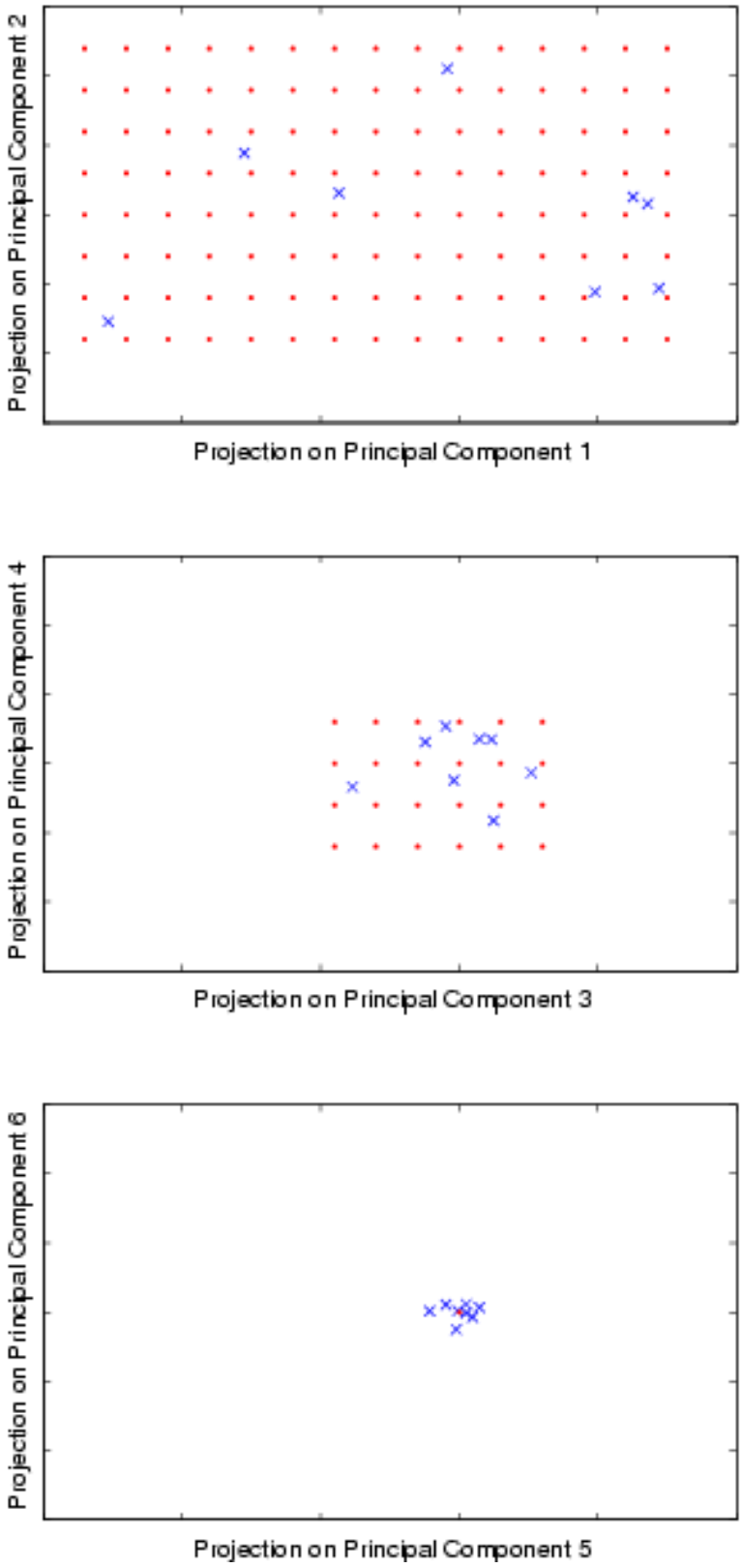


Figure 7.5: A plot of the helical parameters for eight zinc finger structures projected into principal component space. Each known structure is represented by a blue X. The grid used to repack the Zif268–DNA complex is represented by red dots.

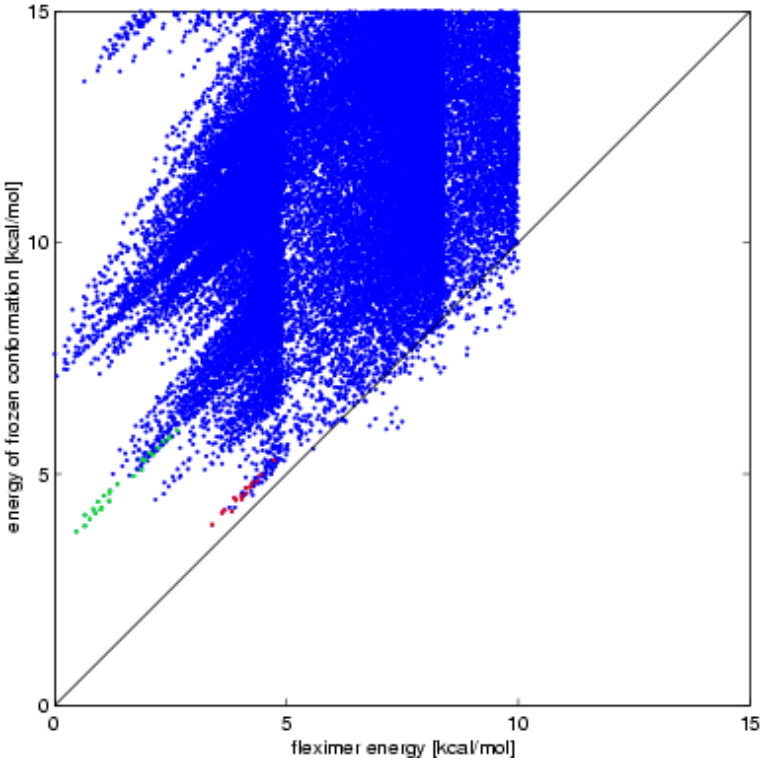


Figure 7.6: A plot of the fleximer energy for each conformational state (horizontal axis) vs. the energy of the system when the fleximers are frozen into one conformation.

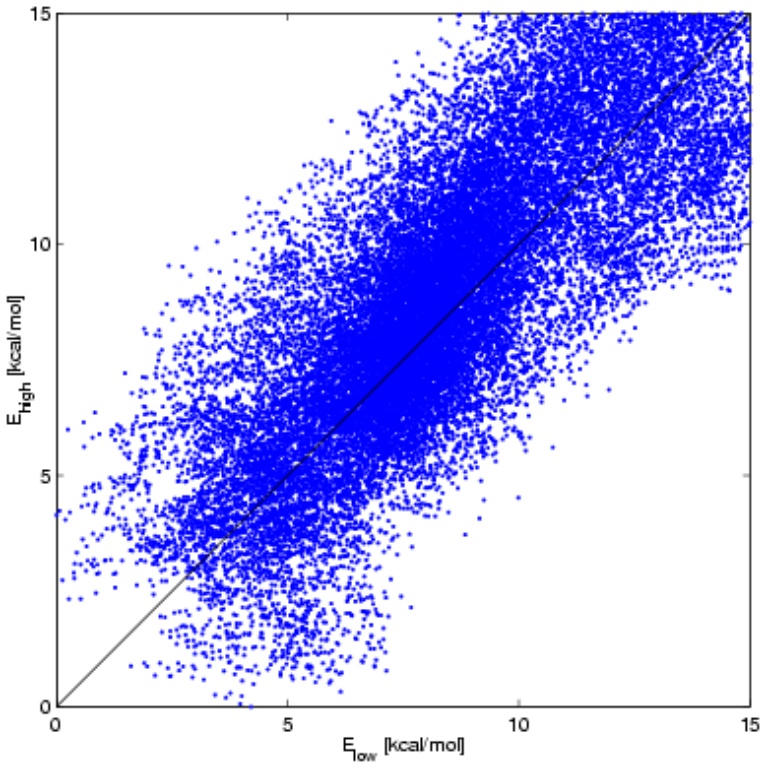


Figure 7.7: A plot of E_{high} vs. E_{low} for conformations of Zif268 complexed to DNA.

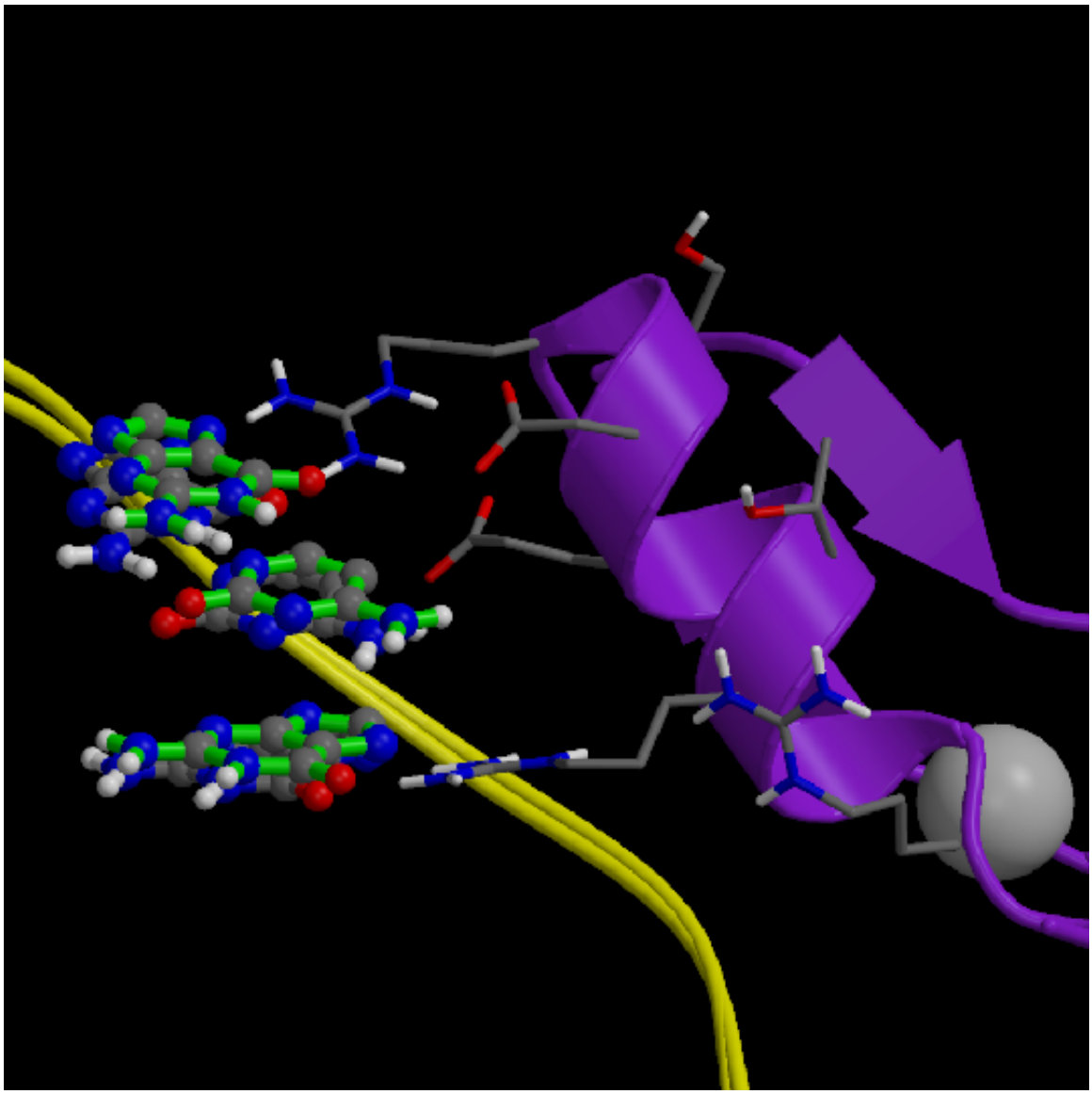


Figure 7.8: The lowest energy conformation of the wild type Zif268–DNA complex. The wild type side chains and bases are shown with gray carbons. The built side chains are shown in green, and the DNA of the repacked structure is shown with green bonds. Side chains of Arg 18, Ser 19 (top of figure), Asp 20, Glu 21, Thr 23, Arg 24, and Arg 27 are shown. The protein backbone (identical for the crystal structure and repacked structure) is shown in purple.

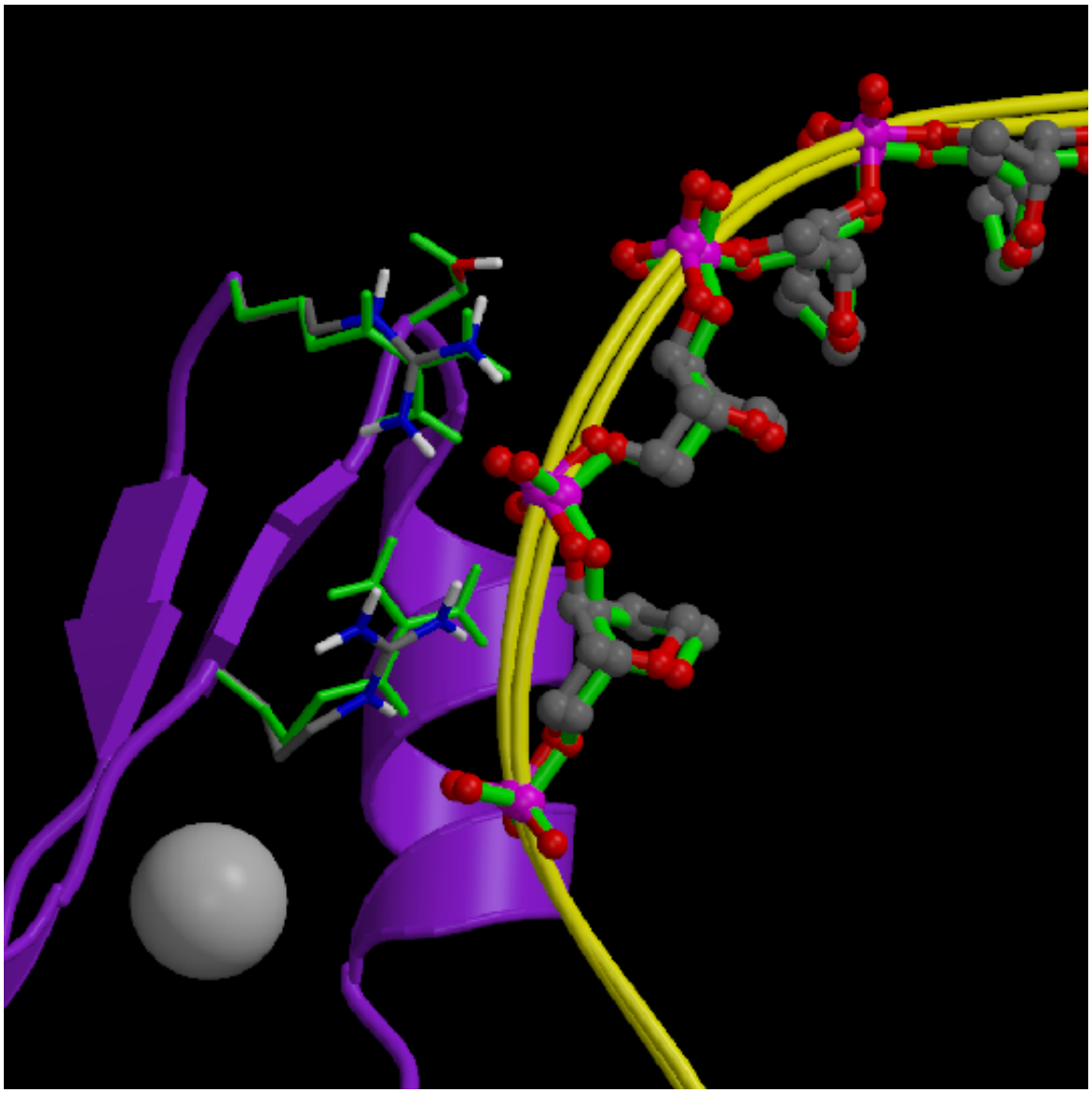


Figure 7.9: The lowest energy conformation of the wild type Zif268–DNA complex, with a view from the opposite side from Figure 7.8. Again, the wild type side chains and bases are shown with gray carbons. The built side chains are shown in green, and the DNA of the repacked structure is shown with green bonds. Side chains of Ser 17, Arg 3, and Arg 14 are shown along with the phosphate backbone of the main contacted strand of DNA.

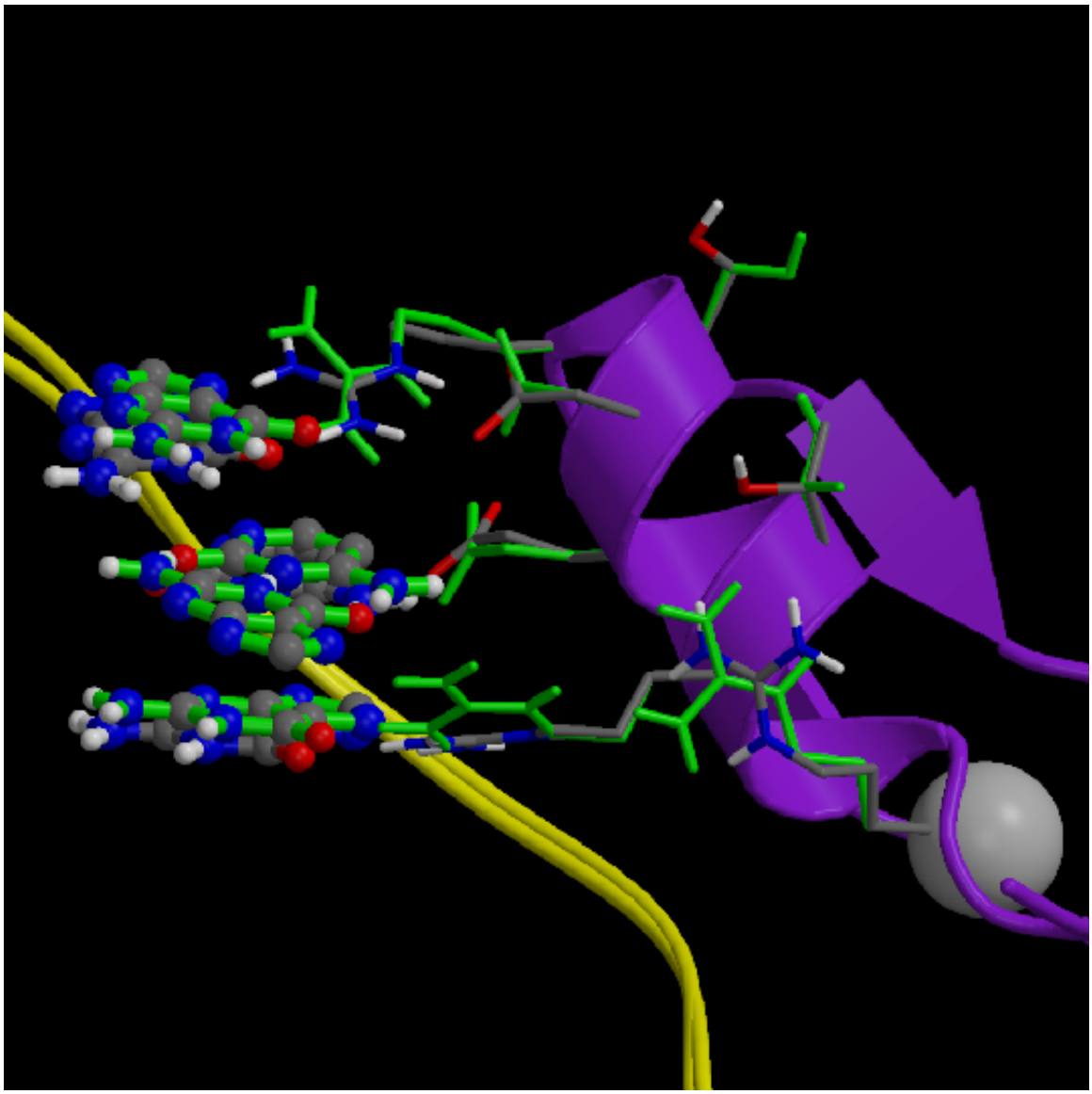


Figure 7.10: Conformation (2) of the wild type Zif268–DNA complex. Representations of molecules are as in Figure 7.8.

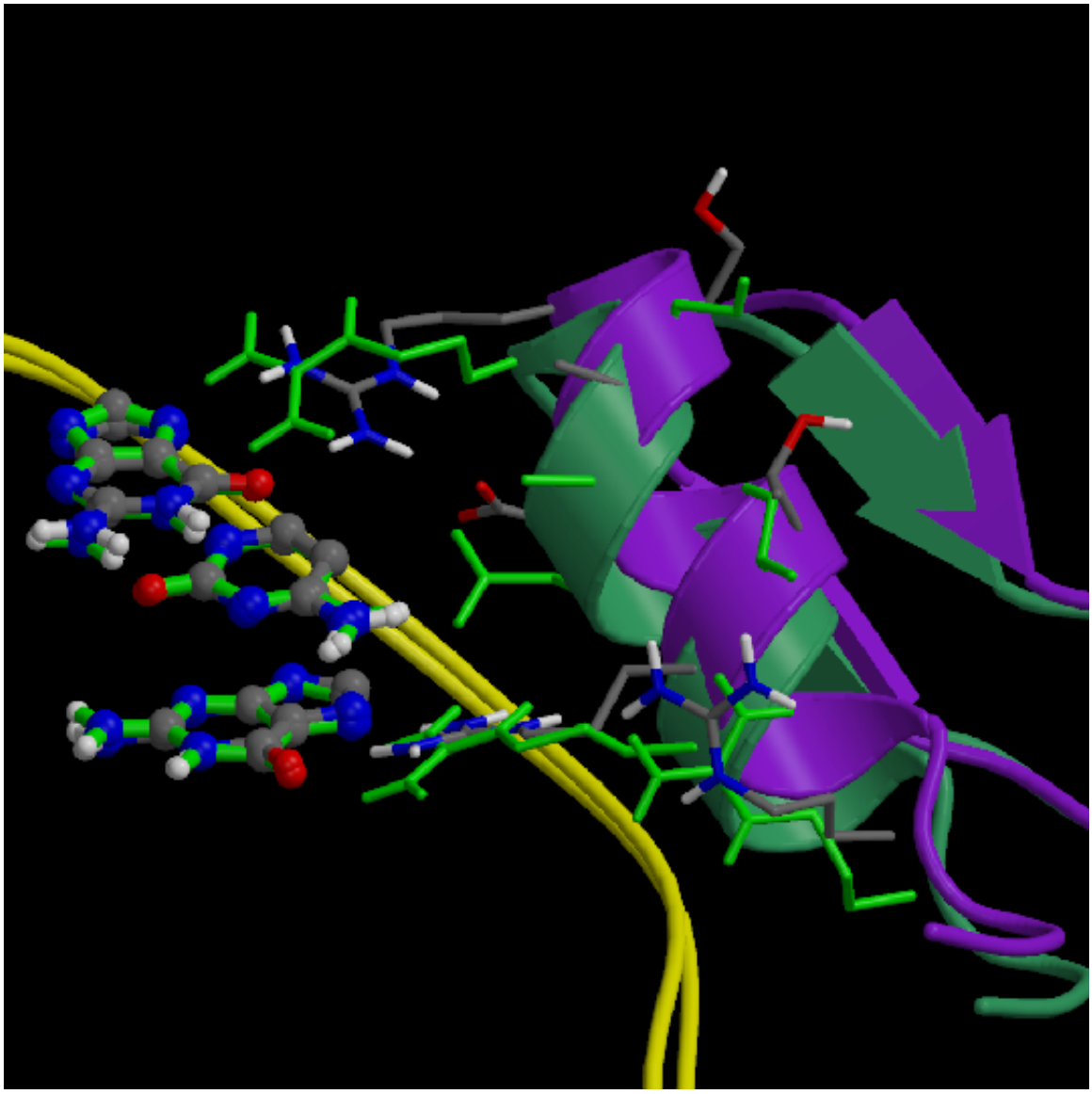


Figure 7.11: Conformation (1) [lowest in energy] of the repacked RADR mutant Zif268–DNA complex. The crystal structure is of the RADR mutant complexed to DNA. The protein backbone and DNA in the repacking calculation come from the wild-type structure. The structures are aligned by DNA bases 8, 9, and 10 in the DNA. The protein backbone of the crystal structure is shown in purple, and that of the repacked structure is shown in green.

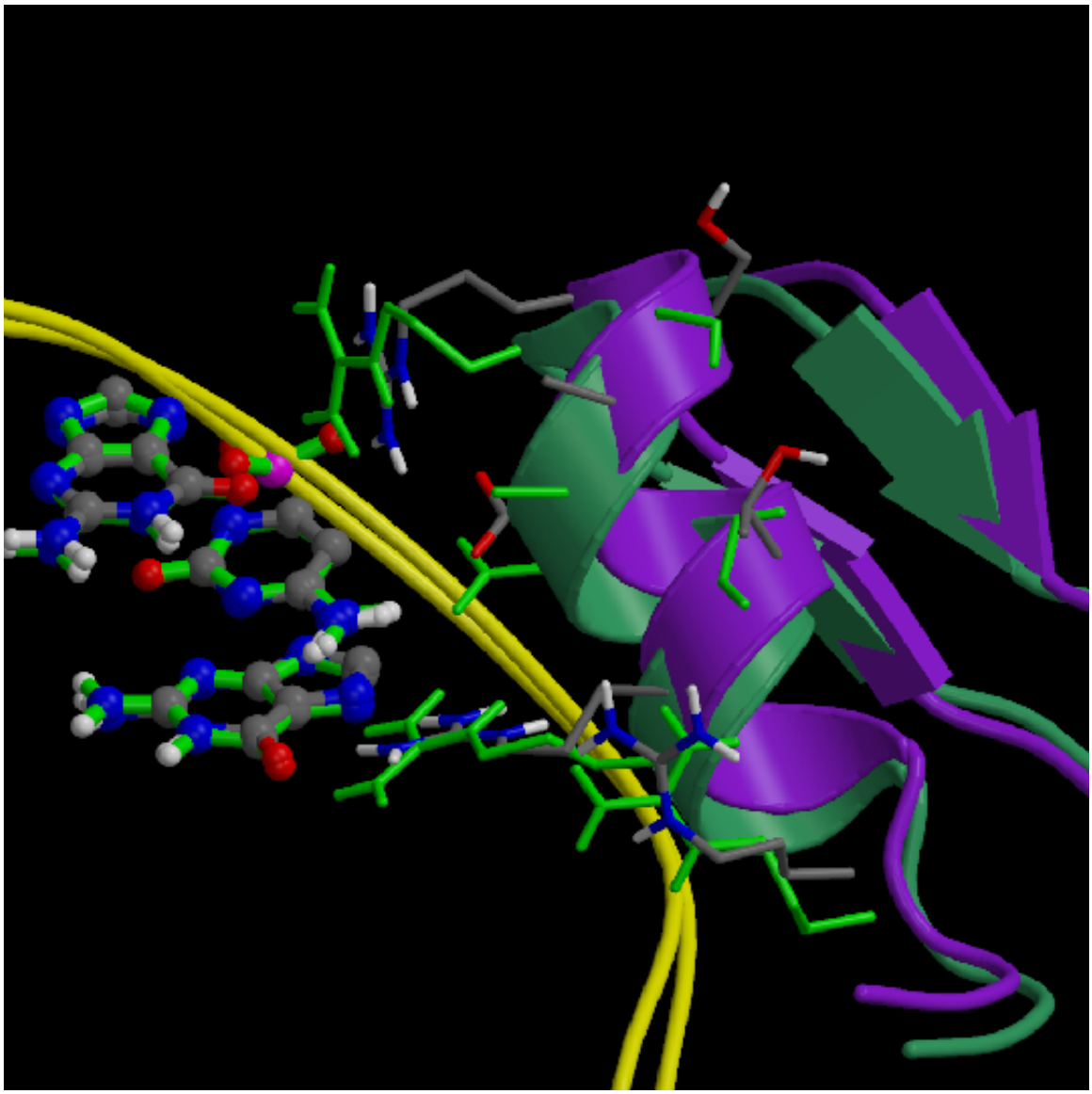


Figure 7.12: Conformation (2) of the repacked RADR mutant Zif268–DNA complex. Phosphate 9 is shown as ball-and-stick, as are the DNA bases contacted by the protein.

Chapter 8

General Conclusions

The role of electrostatics and packing in protein folding and binding was investigated in this thesis. The charge ladder experiments involving carbonic anhydrase II binding were a useful test of continuum electrostatic theory because they were interactions at long-range, and hence the observed variation in binding affinity was expected to be entirely because of electrostatic effects. The results showed that positively charged lysines that were 10-20 Å from the charged inhibitor could have a contribution to binding on the order of 0.1 kcal/mol. Charged residues that were farther from the active site contributed less. The average predicted effect of the lysines agreed well with experiment, but there was no experimental information directly measuring the effect of individual lysines. The electrostatic analysis of the Zif268–DNA complex also revealed strong electrostatic effects from residues away from the interface that are conserved as Arg/Lys. The overall electrostatic effect in this complex is unfavorable to binding, as it is in a number of other complexes. However, the electrostatic interactions are networked in such a way that there are relatively few individual groups whose charge can be deleted to give a much more favorable binding energy. Two of the groups (Glu side chains) that are unfavorable are known to be important for specificity.

A measure of electrostatic complementarity was also developed using the continuum model. When applied to myoglobin structures, the structures of known myoglobins were found to have better electrostatic complementarity than hypothetical

chimeric myoglobin structures. The potentials used in the complementarity measure require less time to compute than a full group-by-group electrostatic analysis, but still give an indication of where a protein's electrostatic interactions could be modified to improve binding or folding. The electrostatic complementarity measure appeared to have better predictive value than other readily available measures of complementarity.

Packing of protein side chains was predicted and optimized using two efficient search algortihms, dead-end elimination and A*. These algorithms to design of an Arc repressor that preferentially forms heterodimers, and an Arc repressor that forms a switch-Arc type helical structure. These calculations showed that, when designing a preference between two closely related structures, DEE/A* may need to generate many possible candidates. DEE/A* itself only finds stable structures, so the screening for structural specificity requires a separate step. The best way to accomplish this screening will require additional experimental data.

Finally, the methods of accounting for electrostatic and shape effects were combined in a calculation that allowed limited docking and repacking of the Zif268 zinc finger–DNA complex. Use of a flexible rotamer approximation allowed the side chain flexibility necessary for the charged and polar side chains to find their correct conformations. The approach involved using a low resolution energy function to identify many candidate structures followed by the use of a more expensive energy function to model solvation and electrostatic effects. The calculated structure showed good agreement with the crystal structure, and the calculations provided an explanation for the difference in conformation between two zinc finger variants.

Bibliography

[1] S. Albeck, R. Unger, and G. Schreiber. Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. *J. Mol. Biol.*, 298:503–520, 2000.

[2] S. A. Allison, G. Ganti, and J. A. McCammon. Simulation of the diffusion-controlled reaction between superoxide and superoxide dismutase. I. Simple models. *Biopolymers*, 24:1323–1336, 1985.

[3] T. A. Anderson. *Mutagenic Effects on Protein Folding and Stability.* PhD thesis, Massachusetts Institute of Technology, 2002.

[4] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. The determinants of pK_a's in proteins. *Biochemistry*, 35:7819–7833, 1996.

[5] L. Z. Avila, Y.-H. Chu, E. C. Blossey, and G. M. Whitesides. Use of affinity capillary electrophoresis to determine kinetic and equilibrium constants for binding of arylsulfonamides to bovine carbonic anhydrase. *J. Med. Chem.*, 36:126–133, 1993.

[6] D. Bashford and M. Karplus. pK_a's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry*, 29:10219–10225, 1990.

[7] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for determining

atom-centered charges: the RESP model. *J. Phys. Chem.*, 97:10269–10280, 1993.

[8] P. Beroza and D. A. Case. Including side chain flexibility in continuum electrostatic calculations of protein titration. *J. Phys. Chem.*, 100:20156–20163, 1996.

[9] P. Beroza, D. R. Fredkin, M. Y. Okamura, and G. Feher. Electrostatic calculations of amino acid titration and electron transfer, $Q_A^- Q_B \rightarrow Q_A Q_B^-$ in the reaction center. *Biophys. J.*, 68:2233–2250, 1995.

[10] G. I. Birnbaum, S. V. Evans, M. Przybylska, and D. R. Rose. 1.70 Ångstrom resolution structure of myoglobin from yellowfin tuna—an example of a myoglobin lacking the D-helix. *Acta Crystallogr. D*, 50:283–289, 1994.

[11] D. A. Bisig, E. E. DiIorio, K. Diederichs, K. H. Winterhalter, and K. Piontek. Crystal-structure of asian elephant (*Elephas-maximus*) cyano-metmyoglobin at 1.78-Å resolution—Phe(29)(B10) accounts for its unusual ligand-binding properties. *J. Biol. Chem.*, 270:20754–20762, 1995.

[12] M. Bolognesi, S. Onesti, G. Gatti, A. Coda, P. Ascenzi, and M. Brunori. *Aplysia-Limacina* myoglobin—crystallographic analysis at 1.6-Å resolution. *J. Mol. Biol.*, 205:529–544, 1989.

[13] A. M. J. J. Bonvin, H. Vis, J. N. Breg, M. J. M. Burgering, R. Boelens, and R. Kaptein. NMR solution structure of the Arc repressor using relaxation matrix calculations. *J. Mol. Biol.*, 236:328–341, 1994.

[14] P. A. Boriack, D. W. Christianson, J. Kingery-Wood, and G. M. Whitesides. Secondary interactions significantly removed from the sulfonamide binding pocket of carbonic anhydrase II influence inhibitor binding constants. *J. Med. Chem.*, 38:2286–2291, 1995.

[15] J. U. Bowie and R. T. Sauer. Equilibrium dissociation and unfolding of the Arc repressor dimer. *Biochemistry*, 28:7139–7143, 1989.

[16] J. U. Bowie and R. T. Sauer. Identifying determinants of folding and activity for a protein of unknown structure. *Proc. Natl. Acad. Sci. U.S.A.*, 86:2152–2156, 1989.

[17] J. N. Breg, J. H. J. Opheusden, M. J. M. Burgering, R. Boelens, and R. Kaptein. Structure of Arc repressor in solution: Evidence for a family of β-sheet DNA-binding proteins. *Nature (London)*, 346:586–589, 1990.

[18] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[19] A. T. Brünger and M. Karplus. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins: Struct., Funct., Genet.*, 4:148–156, 1988.

[20] J. A. Caravella, J. D. Carbeck, D. C. Duffy, G. M. Whitesides, and B. Tidor. Long-range electrostatic contributions to protein–ligand binding estimated using protein charge ladders, affinity capillary electrophoresis, and continuum electrostatic theory. *J. Am. Chem. Soc.*, 121:4340–4347, 1999.

[21] P. J. Carter, G. Winter, A. J. Wilkinson, and A. R. Fersht. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (*Bacillus stearothermophilus)*. *Cell*, 38:835–840, 1984.

[22] L. T. Chong, S. E. Dempster, Z. S. Hendsch, L.-P. Lee, and B. Tidor. Computation of electrostatic complements to proteins: A case of charge stabilized binding. *Protein Sci.*, 7:206–210, 1998.

[23] Y. Choo and A. Klug. A role in DNA binding for the linker sequences of the first three zinc fingers of TFIIIA. *Nuc. Acids Res.*, 21:3341–3346, 1993.

[24] Y.-H. Chu, L. Z. Avila, J. Gao, and G. M. Whitesides. Affinity capillary electrophoresis. *Acc. Chem. Res.*, 28:461–468, 1995.

[25] M. L. Connolly. Shape complmentarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. *Biopolymers*, 25:1229–1247, 1986.

[26] M. H. Cordes, , R. E. Burton, N. P. Walsh, C. J. McKnight, and R. T. Sauer. An evolutionary bridge to a new protein fold. *Nature Struct. Biol.*, 7:1129–1132, 2000.

[27] M. H. Cordes, N. P. Walsh, C. J. McKnight, and R. T. Sauer. Evolution of a protein fold *in vitro*. *Science*, 284:325–327, 1999.

[28] B. I. Dahiyat and S. L. Mayo. Protein design automation. *Protein Sci.*, 5: 895–903, 1996.

[29] B. I. Dahiyat and S. L. Mayo. De novo protein design: Fully automated sequence selection. *Science (Washington, D.C.)*, 278:82–87, 1997.

[30] B. I. Dahiyat and S. L. Mayo. Probing the role of packing specificity in protein design. *Proc. Nat. Acad. Sci. USA*, 94:10172–10177, 1997.

[31] R. L. DesJarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.*, 31:722–729, 1988.

[32] J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature (London)*, 356: 539–542, 1992.

[33] A. J. Doig and M. J. E. Sternberg. Side-chain conformational entropy in protein-folding. *Protein Sci.*, 4:2247–2251, 1995.

[34] R. L. Dunbrack, Jr. and M. Karplus. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.*, 230:543–574, 1993.

[35] A. H. Elcock and J. A. McCammon. The low dielectric interior of proteins is sufficient to cause major structural changes in DNA on association. *J. Am. Chem. Soc.*, 118:3787–3788, 1996.

[36] M. Elrod-Erickson, T. E. Benson, and C. O. Pabo. High-resolution structures of variant Zif268–DNA complexes: Implications for understanding zinc finger–DNA recognition. *Structure*, 6:451–464, 1998.

[37] M. Elrod-Erickson and C. O. Pabo. Binding studies with mutants of Zif268. *J. Biol. Chem.*, 274:19281–19285, 1999.

[38] M. Elrod-Erickson, M. A. Rould, L. Nekludova, and C. O. Pabo. Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. *Structure*, 4:1171–1180, 1996.

[39] A. E. Eriksson, T. A. Jones, and A. Liljas. Refined structure of human carbonic anhydrase II at 2.0 Å resolution. *Proteins: Struct., Funct., Genet.*, 4:274–282, 1988.

[40] M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. A. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, J. Cioslowski, B. B. Stefanov, A. Nanayakkara, M. Challacombe, C. Y. Peng, P. Y. Ayala, W. Chen, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople. *Gaussian 94 (Revision C.2)*. Gaussian, Inc,, Pittsburgh, PA, 1995.

[41] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui,

K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Goll, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle, and J. A. Pople. *Gaussian 98 (Revision A.1).* Gaussian, Inc., Pittsburgh, PA, 1998.

[42] H. A. Gabb, R. M. Jackson, and M. J. E. Sternberg. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, 272:106–120, 1997.

[43] J. Gao, M. Mammen, and G. M. Whitesides. Evaluating electrostatic contributions to binding with the use of charge ladders. *Science (Washington, D.C.)*, 272:535–537, 1996.

[44] R. Gilmanshin, R. B. Dyer, and R. H. Callender. Structural heterogeneity of the various forms of apomyoglobin: Implications for protein folding. *Protein Science*, 6:2134–2142, 1997.

[45] M. K. Gilson. Multiple-site titration and molecular modeling: Two rapid methods for computing energies and forces for ionizable groups in proteins. *Proteins: Struct., Funct., Genet.*, 15:266–282, 1993.

[46] M. K. Gilson and B. H. Honig. Calculation of electrostatic potentials in an enzyme active site. *Nature (London)*, 330:84–86, 1987.

[47] M. K. Gilson, K. A. Sharp, and B. H. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comput. Chem.*, 9:327–335, 1988.

[48] R. F. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, 66:1335–1340, 1994.

[49] F. A. Gomez, L. Z. Avila, Y.-H. Chu, and G. M. Whitesides. Determination of binding constants of ligands to proteins by affinity capillary electrophoresis: Compensation for electroosmotic flow. *Anal. Chem.*, 66:1785–1791, 1994.

[50] A. C. Good, T. J. A. Ewing, D. A. Gschwend, and I. D. Kuntz. New molecular shape descriptors: Application in database screening. *J. Comput.-Aided Mol. Design*, 9:1–12, 1995.

[51] D. B. Gordon and S. L. Mayo. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comp. Chem.*, 19:1505–1514, 1998.

[52] D. B. Gordon and S. L. Mayo. Branch-and-terminate: A combinatorial optimization algorithm for protein design. *Structure*, 7:1089–1098, 1999.

[53] H. A. Greisman and C. O. Pabo. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science (Washington, D.C.)*, 275:657–661, 1997.

[54] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science (Washington, D.C.)*, 282:1462–1467, 1998.

[55] P. B. Harbury, B. Tidor, and P. S. Kim. Repacking protein cores with backbone freedom: Structure prediction for coiled coils. *Proc. Natl. Acad. Sci. U.S.A.*, 92:8408–8412, 1995.

[56] M. Helmer-Critterich and A. Tramontano. Puzzle: A new method for automated protein docking based on surface shape complementarity. *J. Mol. Biol.*, 235:1021–1031, 1994.

[57] Z. S. Hendsch. *Continuum Electrostatic Calculations of Biological Macromolecules*. PhD thesis, Massachusetts Institute of Technology, 2001.

[58] Z. S. Hendsch, M. J. Nohaile, R. T. Sauer, and B. Tidor. Preferential heterodimer formation via undercompensated electrostatics. *J. Am. Chem. Soc.*, 123:1264–1265, 2001.

[59] Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.*, 3:211–226, 1994.

[60] Z. S. Hendsch and B. Tidor. Electrostatic interactions in the GCN4 leucine zipper: Substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci.*, 8:1381–1392, 1999.

[61] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science (Washington, D.C.)*, 268:1144–1149, 1995.

[62] B. Honig, K. Sharp, and A.-S. Yang. Macroscopic models of aqueous solutions: Biological and chemical applications. *J. Phys. Chem.*, 97:1101–1109, 1993.

[63] A. Horovitz. Non-additivity in protein–protein interactions. *J. Mol. Biol.*, 196: 733–735, 1987.

[64] S. R. Hubbard, W. A. Hendrickson, D. G. Lambright, and S. G. Boxer. X-ray crystal structure of a recombinant human myoglobin mutant at 2.8 Ångstroms resolution. *J. Mol. Biol.*, 213:215–218, 1990.

[65] G. H. Jacobs. Determination of the base recognition positions of zinc fingers from sequence analysis. *EMBO J.*, 11:4507–4517, 1992.

[66] A. C. Jamieson, S.-H. Kim, and J. A. Wells. *In Vitro* selection of zinc fingers with altered DNA-binding specificity. *Biochemistry*, 33:5689–5695, 1994.

[67] A. C. Jamieson, H. Wang, and S.-H. Kim. A zinc finger directory for high-affinity DNA recognition. *Proc. Nat. Acad. Sci. USA*, 93:12834–12839, 1996.

[68] E. Kangas. *Optimizing Molecular Electrostatic Interactions: Binding Affinity and Specificity*. PhD thesis, Massachusetts Institute of Technology, 2000.

[69] E. Kangas and B. Tidor. Optimizing electrostatic affinity in ligand–receptor binding: Theory, computation, and ligand properties. *J. Chem. Phys.*, 109: 7522–7545, 1998.

[70] A. E. Keating, V. N. Malashkevich, B. Tidor, and P. S. Kim. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. *Proc. Nat. Acad. Sci. USA*, 98:14825–14830, 2001.

[71] T. K. Kerppola and T. Curran. Fos-Jun heterodimers and Jun homodimers bend DNA in opposite orientations: Implications for transcription factor cooperativity. *Cell*, 66:317–326, 1991.

[72] I. Klapper, R. Hagstrom, R. Fine, K. Sharp, and B. Honig. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *Proteins: Struct., Funct., Genet.*, 1:47–59, 1986.

[73] P. J. Kraulis. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, 24:946–950, 1991.

[74] J. H. Laity, H. J. Dyson, and P. E. Wright. DNA-induced α-helix capping in conserved linker sequences is a determinant of binding affinity in Cys_2-His_2 zinc fingers. *J. Mol. Biol.*, 295:719–727, 2000.

[75] I. Lasters and J. Desmet. The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Eng.*, 6:717–722, 1993.

[76] R. Lavery and H. Sklenar. The definition of generalised helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dynamics*, 6:63–91, 1988.

[77] R. Lavery and H. Sklenar. Defining the structure of irregular nucleic acids: Conventions and principles. *J. Biomol. Struct. Dynamics*, 6:655–667, 1989.

[78] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins: Struct., Func., Genet.*, 33:227–239, 1998.

[79] L.-P. Lee and B. Tidor. Optimization of electrostatic binding free energy. *J. Chem. Phys.*, 106:8681–8690, 1997.

[80] L.-P. Lee and B. Tidor. Barstar is electrostatically optimized for tight-binding to barnase. *Nature Struct. Biol.*, 8:73–76, 2000.

[81] L.-P. Lee and B. Tidor. Optimization of binding electrostatics: Charge complementarity in the barnase–barstar protein complex. *Protein Sci.*, 10: 362–377, 2001.

[82] S. Lindskog, P. Engberg, C. Forsman, S. A. Ibrahim, B.-H. Jonsson, I. Simonsson, and L. Tibell. Kinetics and mechanism of carbonic anhydrase isozymes. *Ann. NY Acad. Sci.*, 429:61–75, 1984.

[83] R. Loewenthal, J. Sancho, T. Reinikainen, and A. R. Fersht. Long-range surface charge–charge interactions in proteins: Comparison of experimental results with calculations from a theoretical method. *J. Mol. Biol.*, 232:574–583, 1993.

[84] L. L. Looger and H. W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. *J. Mol. Biol.*, 307:429–445, 2001.

[85] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[86] S. M. Malakauskas and S. L. Mayo. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.*, 5:470–475, 1998.

[87] R. Maurus, C. M. Overall, R. Bogumil, Y. Luo, A. G. Mauk, M. Smith, and G. D. Brayer. A myoglobin variant with a polar substitution in a conserved hydrophobic cluster in the heme binding pocket. *Biochim. Biophys. Acta*, 1341: 1–13, 1997.

[88] A. J. McCoy, V. C. Epa, and P. M. Colman. Electrostatic complementarity and protein/protein interfaces. *J. Mol. Biol.*, 268:570–584, 1997.

[89] A. D. McLachlan. Rapid comparison of protein structures. *Acta Crystallogr.*, A38:871–873, 1982.

[90] J. Mendes, A. M. Baptista, M. Arménia Carrondo, and C. M. Soares. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins: Struct., Funct., Genet.*, 37:530–543, 1999.

[91] M. E. Milla, B. M. Brown, and R. T. Sauer. Protein stability effects of a complete set of alanine substitutions in Arc repressor. *Nature Struct. Biol.*, 1: 518–523, 1994.

[92] M. E. Milla and R. T. Sauer. Critical side-chain interactions at a subunit interface in the Arc repressor dimer. *Biochemistry*, 34:3344–3351, 1995.

[93] V. K. Misra, J. L. Hecht, K. A. Sharp, R. A. Friedman, and B. Honig. Salt effects on protein–DNA interactions: The λcI repressor and EcoRI endonuclease. *J. Mol. Biol.*, 238:264–280, 1994.

[94] V. K. Misra, J. L. Hecht, A.-S. Yang, and B. Honig. Electrostatic contributions to the binding free energy of the λ cI repressor to DNA. *Biophys. J.*, 75:2262–2273, 1998.

[95] V. K. Misra, K. A. Sharp, R. A. Friedman, and B. Honig. Salt effects on ligand–DNA binding: Minor groove binding antibiotics. *J. Mol. Biol.*, 238:245–263, 1994.

[96] R. L. Murry. Continuum electrostatic analysis of DNA bending. Master's thesis, Massachusetts Institute of Technology, 1996.

[97] J. Nardelli, T. Gibson, and P. Charnay. Zinc finger–DNA recognition: analysis of base specificity by site-directed mutagenesis. *Nucleic Acids Res.*, 20:4137–4144, 1992.

[98] A. Nicholls, K. A. Sharp, and B. Honig. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct., Funct., Genet.*, 11:281–296, 1991.

[99] M. J. Nohaile, Z. S. Hendsch, B. Tidor, and R. T. Sauer. Altering dimerization specificity by changes in surface electrostatics. *Proc. Nat. Acad. Sci. USA*, 98: 3109–3114, 2001.

[100] J. Novotny, R. E. Bruccoleri, M. Davis, and K. A. Sharp. Empirical free energy calculations: A blind test and further improvements to the method. *J. Mol. Biol.*, 268:401–411, 1997.

[101] J. Novotny and K. Sharp. Electrostatic fields in antibodies and antibody/antigen complexes. *Prog. Biophys. Molec. Biol.*, 58:203–224, 1992.

[102] H. Oberoi, J. Trikha, X. Yuan, and N. M. Allewell. Identification and analysis of long-range electrostatic effects in proteins by computer modeling: Aspartate transcarbamylase. *Proteins: Struct., Func., Genet.*, 25:300–314, 1996.

[103] U. Obeysekare, C. Kissinger, L. J. Keefe, B. E. Raumann, R. T. Sauer, and C. O. Pabo. Unpublished results.

[104] T. J. Oldfield, S. J. Smerdon, Z. Dauter, K. Petratos, K. S. Wilson, and A. J. Wilkinson. High-resolution x-ray structures of pig metmyoglobin and 2 CD3 mutants. *Biochemistry*, 31:8732–8739, 1992.

[105] D. N. Paolella, Y. Liu, M. A. Fabian, and A. Schepartz. Electrostatic mechanism for DNA bending by bZIP proteins. *Biochemistry*, 36:10033–10038, 1997.

[106] R. V. Pappu and D. L. Weaver. The early folding kinetics of apomyoglobin. *Protein Science*, 7:480–490, 1998.

[107] N. P. Pavletich and C. O. Pabo. Zinc finger–DNA recognition: Crystal structure of a Zif268–DNA complex at 2.1 Å. *Science (Washington, D.C.)*, 252:809–817, 1991.

[108] K. Phillips and S. E. V. Phillips. Electrostatic activation of the *Escherichia coli* methionine repressor. *Structure*, 2:309–316, 1994.

[109] S. E. V. Phillips and B. P. Schoenborn. Neutron-diffraction reveals oxygen-histidine hydrogen-bond in oxymyoglobin. *Nature*, 292:81–82, 1981.

[110] N. A. Pierce, J. A. Spriet, J. Desmet, and S. L. Mayo. Confomational splitting: A more powerful criterion for dead-end elimination. *J. Comp. Chem.*, 21:999–1009, 2000.

[111] J. W. Ponder and F. M. Richards. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193:775–791, 1987.

[112] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, 1986.

[113] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, second edition, 1992.

[114] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A*, 101:3005–3014, 1997.

[115] B. E. Raumann, M. A. Rould, C. O. Pabo, and R. T. Sauer. DNA recognition by β-sheets in the Arc repressor–operator crystal structure. *Nature (London)*, 367:754–757, 1994.

[116] G. Ravishanker, S. Swaminathan, D. L. Beveridge, R. Lavery, and H. Sklenar. Conformation and helicoidal analysis of 30 psec of molecular dynamics on the d(cgcgaattcgcg)) double helix: Curves, dials and windows. *J. Biomolec. Str. and Dyn.*, 6:669–699, 1989.

[117] E. J. Rebar and C. O. Pabo. Zinc finger phage: Affinity selection of fingers with new DNA-binding specificities. *Science (Washington, D.C.)*, 263:671–673, 1994.

[118] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234:779–815, 1993.

[119] M. Schaefer and M. Karplus. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.*, 100:1578–1599, 1996.

[120] G. Schreiber, C. Frisch, and A. R. Fersht. The role of Glu73 of barnase in catalysis and the binding of barstar. *J. Mol. Biol.*, 270:111–112, 1997.

[121] T. Selzer and G. Schreiber. Predicting the rate enhancement of protein complex formation from the electrostatic energy of interaction. *J. Mol. Biol.*, 287:409–419, 1999.

[122] L. Serrano, A. Horovitz, B. Avron, M. Bycroft, and A. R. Fersht. Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry*, 29:9343–9352, 1990.

[123] K. Sharp. Electrostatic interactions in hirudin–thrombin binding. *Biophys. Chem.*, 61:37–49, 1996.

[124] K. Sharp, R. Fine, and B. Honig. Computer simulations of the diffusion of a substrate to an active site of and enzyme. *Science (Washington, D.C.)*, 236: 1460–1463, 1987.

[125] K. A. Sharp. Calculation of hyhel10-lysosyme binding free energy changes: Effect of ten point mutations. *Proteins: Struct., Func., Genet.*, 33:39–48, 1998.

[126] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, 19:301–332, 1990.

[127] J. Shen and J. Wendoloski. Electrostatic binding energy calculation using the finite difference solution to the linearized Poisson–Boltzmann equation: Assessment of its accuracy. *J. Comput. Chem.*, 17:350–357, 1996.

[128] M. Shimaoka, J. M. Shifman, H. Jing, L. Takagi, S. L. Mayo, and T. A. Springer. Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.*, 7:674–678, 2000.

[129] T. Simonson. Dielectric constant of cytochrome c from simulations in a water droplet including all electrostatic interactions. *J. Am. Chem. Soc.*, 120:4875–4876, 1998.

[130] T. Simonson and C. L. Brooks, III. Charge screening and the dielectric constant of proteins: Insights from molecular dynamics. *J. Am. Chem. Soc.*, 118:8452–8458, 1996.

[131] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.

[132] H. Sklenar, C. Etchebest, and R. Lavery. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins: Struct., Func., Genet.*, 6:46–60, 1989.

[133] W. S. Somers, J. B. Rafferty, K. Phillips, S. Strathdee, Y-Y. He, T. McNally, I. Manfield, O. Navratil, I. G. Old, I. Saint-Girons, P. G. Stockley, and S. E. V. Phillips. The Met repressor–operator complex: DNA recognition by β-strands. *Ann. NY Acad. Sci.*, 726:105–117, 1994.

[134] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[135] J. K. Strauss-Soukup and L. J. Maher, III. DNA bending by GCN4 mutants bearing cationic residues. *Biochemistry*, 36:10026–10032, 1997.

[136] J. K. Strauss-Soukup and L. J. Maher, III. Role of asymmetric phosphate neutralization in DNA bending by PU.1. *J. Biol. Chem.*, 272:31570–31575, 1997.

[137] A. G. Street and S. L. Mayo. Pairwise calculation of protein solvent-accessible surface areas. *Folding & Design*, 3:253–258, 1998.

[138] A. H. Swirnoff and J. Milbrandt. DNA–binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.*, 15:2275–2287, 1995.

[139] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. The clustalx windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nuc. Acids Res.*, 24:4876–4882, 1997.

[140] A. A. Travers. Reading the minor-groove. *Nature Struct. Biol.*, 2:615–618, 1995.

[141] C. D. Waldburger, J. F. Schildbach, and R. T. Sauer. Are buried salt bridges important for protein stability and conformational specificity? *Nature Struct. Biol.*, 2:122–128, 1995.

[142] L. Wang, T. O'Connell, A. Tropsha, and J. Hermans. Energetic decomposition of the α-helix–coil equilibrium of a dynamic model system. *Biopolymers*, 39: 479–489, 1996.

[143] M. H. Werner, A. M. Gronenborn, and G. M. Clore. Intercalation, DNA kinking, and the control of transcription. *Science (Washington, D.C.)*, 271:778–784, 1996.

[144] W. C. Wimley, K. Gawrisch, T. P. Creamer, and S. H. White. Direct measurement of salt-bridge solvation energies using a peptide model system: Implications for protein stability. *Proc. Natl. Acad. Sci. U.S.A.*, 93:2985–2990, 1996.

[145] P. H. Winston. *Artificial Intelligence*. Addison-Wesley, Reading, Massachussetts, 1992.

[146] J. Wiorkiewicz and M. Karplus. Personal communication, 1990.

[147] A.-S. Yang, B. Hitz, and B. Honig. Free energy determinants of secondary structure formation: III. β-turns and their role in protein folding. *J. Mol. Biol.*, 259:873–882, 1996.

[148] A.-S. Yang and B. Honig. Free energy determinants of secondary structure formation: I. α-helices. *J. Mol. Biol.*, 252:351–365, 1995.

[149] A.-S. Yang and B. Honig. Free energy determinants of secondary structure formation: II. Antiparallel β-sheets. *J. Mol. Biol.*, 252:366–376, 1995.

[150] T. J. You and D. Bashford. Conformation and hydrogen ion titration of proteins: A continuum electrostatic model with conformational flexibility. *Biophys. J.*, 69:1721–1733, 1995.

[151] M. Zacharias, B. A. Luty, M. E. Davis, and J. A. McCammon. Poisson–Boltzmann analysis of the λ repressor–operator interaction. *Biophys. J.*, 63: 1280–1285, 1992.