

Computational Approaches for the Design and Prediction of Protein-Protein Interactions

by

Gevorg Grigoryan

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

Doctor of Philosophy

at the

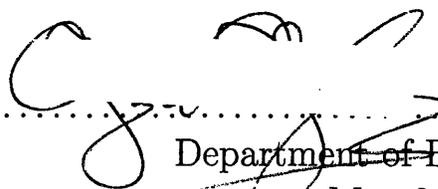
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

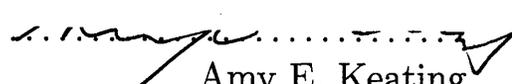
© Gevorg Grigoryan, MMVII. All rights reserved.

The author hereby grants to MIT permission to reproduce and distribute publicly paper and electronic copies of this thesis document in whole or in part.

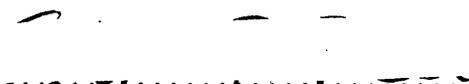
Author

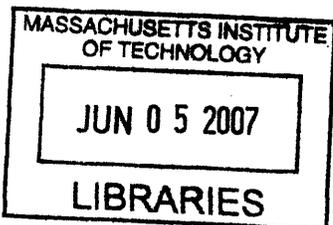

Department of Biology
May 24, 2007

Certified by


Amy E. Keating
Assistant Professor
Thesis Supervisor

Accepted by


Stephen Bell
Chairman, Department Committee on Graduate Students



ARCHIVES

Computational Approaches for the Design and Prediction of Protein-Protein Interactions

by

Gevorg Grigoryan

Submitted to the Department of Biology
on May 24, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

There is a large class of applications in computational structural biology for which atomic-level representation is crucial for understanding the underlying biological phenomena, yet explicit atomic-level modeling is computationally prohibitive. Computational protein design, homology modeling, protein interaction prediction, docking and structure recognition are among these applications. Models that are commonly applied to these problems combine atomic-level representation with assumptions and approximations that make them computationally feasible. In this thesis I focus on several aspects of this type of modeling, analyze its limitations, propose improvements and explore applications to the design and prediction of protein-protein interactions.

Thesis Supervisor: Amy E. Keating

Title: Assistant Professor

Acknowledgments

There are many people without whom my graduate experience would either not be possible or would not be nearly as exciting and rewarding as it was. Above all, I would like to thank my parents for constant support and encouragement as well as a never-ending supply of great advice. They have always helped me keep things in perspective and I consider them an integral part of any success I have had or will have in life. I would also like to thank my sister Armine who is my life-long partner in crime and without whose energy and capacity for laughter life would just be too dull for me. And, of course, I would like to thank my girlfriend Keila for all her support and understanding through these years. Together with me, she has shared all the ups and downs of graduate life and her constant encouragement made it very easy to stay focused and enthusiastic about my work.

Members of the Keating lab have been an integral part of my experience at MIT. In addition to being colleagues we have also become friends, which made the lab a very comfortable and productive environment to work in. I have had numerous stimulating conversations on topics related and unrelated to science with James Apgar. I would also like to thank (in no particular order) Devdoot Majumdar and Aaron Reinke for frequent late-night conversations in lab, Emiko Bare for knowing everything about everything in the department, Shaun Deignan for helping me deal with computer issues and Alejandro Ochoa for being a great UROP. My scientific thinking has benefited greatly from conversations with Christy Taylor, Xiaoran Fu, Nora Zizlsperger and Karl Gutwin.

I have had the pleasure to be classmates with some very bright individuals, many of whom will surely move on to be part of the academic elite. I would like to acknowledge especially Gordon Lu and Francois St-Pierre. Conversations with both have stimulating many new lines of thinking. Francois has also read in great detail every manuscript I have produced as a graduate student and provided wonderfully insightful comments. He has also been a great friend throughout my years at MIT.

It was extremely fortunate that one of my high school classmates, Dmitry Malioutov,

was a PhD student at MIT at the same time with me. Dmitry and I have become the best of friends and this friendship was absolutely central to my everyday existence at MIT. In addition, he happens to be exceptionally bright and, although his field of study is in statistical inference models, conversations with him have made very direct contributions to my research as well as my general knowledge base.

I would like to thank my thesis committee members Bob Sauer, Chris Burge, Mike Yaffe and Leonid Mirny for their continued guidance, advice and suggestions. I would especially like to thank Bob for also meeting with me on numerous occasions to discuss scientific as well as career-related questions. His experience and wisdom have been great assets to have access to. Bob was also nice enough to read and provide useful comments on my manuscripts.

Last, but most certainly not least, I would like to thank my advisor Amy E. Keating. Saying that none of my work would have been possible without her does not even begin to describe the degree to which she has contributed to my development as a graduate student. Her leadership and her friendship have redefined for me the standard for a relationship between an advisor and an advisee. In addition to being a great mentor, Amy also possesses a great sense of scientific insight and intuition, from which I have benefited immensely throughout my graduate career. She has taught me much about how to choose problems of study, how to convey ones work to other scientists, how to write good papers and much more. Although Amy was always more than happy to share with me her insight and provide guidance, she was also absolutely willing to let me pursue my own ideas, which taught me independence. I hope that when the times comes for me to be in charge of a laboratory, I can be as good of a mentor, colleague and friend to my students as Amy was to me.

Contents

1	Introduction	17
1.0.1	Structural Sampling in DSF Models	18
1.0.2	Solvent in DSF Models	19
1.0.3	Implications of Removing Structural Degrees of Freedom . . .	20
1.0.4	The Reference State	21
1.0.5	Systematic Reduction of Complexity	22
2	Structure-based prediction of bZIP partnering specificity	25
2.1	Introduction	26
2.2	Results	30
2.2.1	Testing Framework	30
2.2.2	General Modeling Procedure	31
2.2.3	Explicit Unfolded State	32
2.2.4	Implicit Unfolded State	35
2.2.5	Intra-helix interactions	36
2.2.6	Alternative Models for Core Interactions	37
2.2.7	Comparison to Other Methods	41
2.2.8	Interaction <i>versus</i> non-interaction Discrimination	42
2.2.9	Implicit <i>versus</i> explicit models of side chain-side chain interactions	43
2.3	Discussion	46
2.4	Methods	52
2.4.1	Datasets and testing	52

2.4.2	Repacking and minimization	53
2.4.3	Evaluation of folded energy	55
2.4.4	Helix propensities	56
2.4.5	Full-helix folding energy function	57
2.4.6	Modeling backbone relaxation	57
2.4.7	DFIRE	58
2.4.8	RosettaDesign	58
2.4.9	FOLD-X	58
2.4.10	Evaluating contributions of $\mathbf{a}_i - \mathbf{a}'_i$ interactions to binding . .	58
2.5	Acknowledgements	59
2.6	Supplementary Data	60
3	Ultra-fast Evaluation of Protein Energies Directly from Sequence	61
3.1	Introduction	62
3.1.1	Theory	64
3.1.2	Interpretation of the Expansion	68
3.2	Results	70
3.2.1	Coiled Coil	71
3.2.2	Zinc Finger	77
3.2.3	WW Domain	80
3.2.4	A Design Application and Speedup Analysis	83
3.3	Discussion	84
3.3.1	Conclusion	88
3.4	Materials and Methods	89
3.4.1	Repacking and minimization	89
3.4.2	The coiled-coil unit cell	90
3.4.3	Cluster Expansion fitting	90
3.4.4	Zinc-finger design exercise	92
3.5	Acknowledgements	93

4	Computing van der Waals Energies in the Context of the Rotamer Approximation	95
4.1	Abbreviations	96
4.2	Introduction	96
4.3	Materials and Methods	99
4.3.1	Structural Database	99
4.3.2	Repacking, Design, Minimization and Evaluation	100
4.3.3	van der Waals modifications	103
4.3.4	Statistical Measures	104
4.4	Results	104
4.4.1	Overview of van der Waals modifications	107
4.4.2	Modified van der Waals energies versus NCE	110
4.5	Discussion	121
4.6	Acknowledgements	128
5	A Novel Framework for Specificity Design	131
5.1	Introduction	132
5.2	Results and Discussion	136
5.2.1	Designs that Optimize Stability Hit Off-target Partners	136
5.2.2	Designing Stability and Specificity	140
5.2.3	Proposed Designs for Experimental Testing	143
5.3	Conclusions	146
5.4	Materials and Methods	148
5.4.1	Cluster Expansion	148
5.4.2	Formulation of the Problem as an Integer Linear Program	149
5.4.3	Design Specifications	152
5.4.4	Choosing b, c and f positions	153
6	Conclusions	155
6.1	Coarseness of Structural Sampling	155
6.2	Adjustable Energy Functions	157

6.3	Unfolded States	158
6.4	Summary	158
7	Possible Future Directions	161
7.1	Specificity Design Framework	161
7.2	Structure-based Modeling of Coiled-coil Interactions	162
7.3	Cluster Expansion	166

List of Figures

2-1	Number of coiled-coil pairs AB and AC (A, B, and C correspond to bZIP sequences) consistently satisfying $S_{AB} - S_{AC} > \Delta F$, where is the raw fluorescence signal for dimer XY observed in ref [134].	31
2-2	Performance of different models for predicting dimer stability differences, as a function of ΔF	34
2-3	Optimizing the value of parameter hp_{ref} in the context of model HP using different dimer comparison datasets.	36
2-4	Performance of model HP/S on the dimer comparison dataset with $\Delta F = 6000$ as a function of hp_{ref} and s	38
2-5	Optimizing the value of parameter s on different dimer comparison datasets gave very similar optimality ranges.	38
2-6	Comparison of (a) $\mathbf{g}_i \mathbf{e}'_{i+1}$ and (b) $\mathbf{a}_i \mathbf{a}'_i$ coupling energies measured by Vinson and co-workers [3, 94] with the corresponding computed interactions.	40
2-7	Comparison of $\mathbf{a}_i \mathbf{a}'_i$ coupling energies measured by Vinson and co-workers [3] with corresponding FKS weights.	41
2-8	Performance of model HP/S/C on discriminating between interacting and non-interacting leucine zippers.	43
2-9	ROC analysis of performance predicting interacting and non-interacting leucine zippers.	44
3-1	The procedure for fitting a cluster expansion.	66
3-2	Schematic of a parallel dimeric coiled coil.	72

3-3	The unit cell used for modeling coiled-coil interactions.	73
3-4	Cluster expansion of coiled-coil energies.	76
3-5	Agreement between experimentally measured double-alanine coupling energies for residues E, Q, R and K at $\mathbf{g} - \mathbf{e}' +$ [94] and corresponding pair ECI from the cluster expansion (in kcal/mol).	77
3-6	Cluster expansion of zinc-finger (ZF) energies.	79
3-7	Important triplet clusters for the expansion of zinc-finger energies. . .	80
3-8	Cluster expansion of WW domain energies.	82
3-9	Important higher-order clusters for the expansion of WW-domain energies.	83
3-10	Distribution of the energies of the top 100 sequences from direct design and CE design.	85
4-1	A cartoon representing the RCE and NCE landscapes.	105
4-2	Overview of the computational experiment for comparing the ability of different van der Waals modifications to predict NCE.	108
4-3	Scatter plots of RCE vs. NCE for three vdW modifications - L-J (in a and d), R90 (in b and e) and PRM (in c and f).	112
4-4	Performance of different vdW modifications on predicting the NCE of low-RCE structures resulting from native sequence repacking (left panels) or sequence design (right panels).	114
4-5	Global tests of the RCE energy landscape.	116
4-6	Comparison of different rotamer libraries using either the R90 or L-J potentials.	117
4-7	Mapping from pre-minimization atom-to-atom distances onto post-minimization atomic interaction energies.	119
4-8	The contribution of a given rotamer pair to the NCE of a structure strongly depends on the surrounding structural context.	122
4-9	Agreement between RCE and NCE contributions for clashing atomic interactions using either R90 or L-J.	123

5-1	Illustration of the specificity sweep procedure.	135
5-2	Agreement between structure-based energies explicitly calculated with HP/S/C and a CE sequence-based approximation.	137
5-3	Results of optimizing binding partners against each human bZIP coiled coil.	138
5-4	Improving specificity at the cost of stability.	142
5-5	Specificity sweeps against several targets with model HP/S/C.	145

List of Tables

2.1	Coiled-coil backbone variation in members of the bZIP family.	32
2.2	χ -angle recovery for placing native side chains on bZIP and non-bZIP parallel dimeric coiled-coil backbones.	33
2.3	Calculated contribution of N-N, N-V, or V-V at $\mathbf{a}_i\mathbf{a}'_i$ positions to the rigid-body binding energy of the coiled coil GCN4.	47
4.1	Summary of Characteristics of the Protein Structure Set	99
4.2	Summary of Commonly Used Abbreviations	101
4.3	Changes resulting from minimization of repacked rotameric structures	106
5.1	Final sequences for experimental characterization.	147

Chapter 1

Introduction

Proteins are among the primary macromolecular players of the cell. The ability of proteins to adopt 3-dimensional structure, interact with other proteins, change conformation under various conditions and catalyze reactions are of central interest to biologists. Important to each of the above processes is the fact that proteins are highly structurally flexible. However, proteins are also very complex systems with many degrees of freedom and their structural flexibility is hard to model. They exist somewhere between the quantum world (single atom level) and the macro world (thermodynamic level), for both of which there are well established physical treatments. So an appropriate model for describing protein behavior should lie somewhere between these extremes in the space of physical models, and the most appropriate level of modeling will depend on the application. Many reduced representations of proteins, such as lattice models or beads-on-a-string models, have been used, and in some cases these reduced systems have been rigorously theoretically treated [38, 124, 112]. However, many of the aspects of protein behavior interesting to biologists are determined by higher resolution structural information. Indeed, often it is processes occurring at the atomic level, such as phosphorylation or other chemical modifications, that lead to a biological readout. Similarly, binding of two proteins can lead to atomic-level structural rearrangements exposing previously hidden functional groups and propagating the biological signal. To understand protein behavior in the cell, we must understand such events and thus an atomic-level model of proteins is necessary.

Unfortunately, for many practical applications explicit atomistic models are prohibitively costly. These applications include protein design, homology modeling, interaction prediction, docking, structure recognition and others. Hybrid models need to be formulated that combine elements of atomic-level explicit modeling, and thus retain the necessary resolution, while being computationally more efficient. There is a popular class of models that has been used for this, although it has not been widely recognized as a class. I will refer to these as “discrete structural flexibility” models (DSF models). The idea behind this approach is that rather than modeling proteins as being continuously flexible, the space of possible protein conformation is discretized. This provides several advantages but also necessitates numerous approximations. Here I will review important considerations for such models, lay out the major approximations and assumptions and outline the state of the methods in the field. I will also point out those limitations of DSF models that I have tried to address in my work and put in perspective the methodological improvements I have proposed.

1.0.1 Structural Sampling in DSF Models

Generally, the protein is broken down into its backbone and side-chains portions, which are treated differently in terms of their structural freedom. Side-chain flexibility is restricted to a set of commonly observed conformations, known as rotamers [43, 44, 113]. Backbone flexibility is treated differently in different methods. Often, the backbone is simply fixed to the conformation observed in a native protein. This approach can be used in applications such as protein design [117, 35] and docking [162, 69]. As a generalization of this approach, a finite set of backbones can be used to represent some of the structural flexibility [65, 88, 4]. Finally, in another method the backbone is continually rearranged using a pre-compiled discrete set of structural fragments. Baker and co-workers pioneered this approach and have applied it, with great success, to problems of protein design [95], structure prediction [23], docking [184] and homology modeling [126]. No matter how the backbone is modeled, however, the flexibility of the backbone is treated separately from the flexibility of

side-chains. That is, for each one backbone conformation, side-chain conformations are optimized to minimize energy. This can be viewed as a hierarchical approach to flexibility, where the conformation of the backbone is considered to define most of the structure and side-chain conformations provide the final small adjustments.

Clearly, modeling proteins as being discretely flexible is an approximation. However, from the structural standpoint this approximation is probably not that severe. For example, it is known that most sidechains do in fact occur in conformations close to a rotameric one [153, 113]. Fixing the backbone can, in some instances, be a bad approximation, but results from the Baker lab indicate that naturally observed backbone flexibility can indeed be represented via a library of small structural fragments. Alternatively, for some systems effective parameterizations of backbone freedom can be derived [32, 65]. The real difficulty associated with using DSF models lies in the energetic treatment of resulting discrete structures.

1.0.2 Solvent in DSF Models

When a protein is represented with all of its atoms and surrounded by explicit solvent, the potential energy can be calculated by simply adding pairwise atomic interactions. In this case, for any pair of atoms (or any pair of atom groups) their interaction energy can be unambiguously separated and is independent of the positions of other atoms. However, when some of the system's degrees of freedom are removed or restricted, which is what DSF models do, such explicit pairwise separability is lost. One example of this is related to the treatment of solvent degrees of freedom. In most DSF models solvent molecules are not explicitly treated. In principle, solvent conformational degrees of freedom can also be discretized and an attempt at this through the use of discrete "solvated" rotamers has been made [80]. However, due to the fact that the solvent is much less conformationally restrained than the protein, it is less amenable to conformational discretization. Instead, most DSF models use some sort of an aggregate representation for the solvent that tries to account for the effect of solvent averaged over all of its conformational states. One of the most common approaches is to use a uniform high-dielectric medium in place of water and apply

the Poisson-Boltzmann theory to solve for electrostatic energies numerically [73, 7]. However, although, Coulombic interactions are completely pairwise decomposable, when solvent degrees of freedom are averaged, electrostatic energy can no longer be represented in terms of independent contributions from pairs of atom groups. This significantly complicates many of the aspects of the methods used to approach the problems mentioned above (e.g. protein design, docking, structure prediction, etc.). To circumvent this, pairwise-decomposable electrostatic models have been developed that are either approximations of the complete Poisson treatment [67], are empirical models meant to recapitulate the general hydrophobic/hydrophilic tendencies in proteins [173, 88] or are fit to experimental data [103, 47]. However, approximations of electrostatic energy can lead to problems for such applications as computational protein design [181]. Additionally, because electrostatics frequently plays a major role in determining protein-protein interaction preferences, such approximations can be expected to cause problems in structure-based protein interaction prediction as well. Chapter 2 describes how I used structural modeling to predict interaction preferences among human bZIP proteins. In this work, I used a hierarchical approach to dealing with the inaccuracies associated with approximate electrostatics models. Pairwise approximate treatments of electrostatics can be used to arrive at a reasonable low-energy structure (or an entire list of structures), which can then be re-scored using a higher-accuracy electrostatics models. Others have adopted a similar strategy in computational protein design [59].

1.0.3 Implications of Removing Structural Degrees of Freedom

Just as the solvent degrees of freedom are averaged in DSF models, so too are some of the protein conformational degrees of freedom. For example, because side-chain flexibility is restricted to a finite set of rotamers, each rotameric configuration of a protein actually represents an entire ensemble of conformations structurally close to it. Thus, the energy associated with this configuration should also be represen-

tative of this ensemble. Because of this, even those non-bonded interaction terms that preserve pairwise-decomposability in the DSF framework, such as van der Waals interactions, can not necessarily be modeled as such. A common approach to this problem, in relation to van der Waals energies, is to modify the shape of the van der Waals potential to make it more “fuzzy”, thereby accounting for some of the possible structural relaxations of the protein. However, extensive modifications lead to non-physical energies. In chapter 4, I explore this topic and compare the appropriateness of a range of commonly used modifications. I also show that by adopting the rotamer approximation, one makes the problem of computing appropriate van der Waals extremely non-pairwise decomposable so that no particular modification is expected to work well.

1.0.4 The Reference State

In an explicit model of protein flexibility, where none of the degrees of freedom are frozen, a protein is expected to visit all of its accessible states with probabilities related to the free energies of these states. Thus, the free energy difference between any two states, such as the folded and the unfolded states, can be calculated by simply measuring the fraction of the time the protein spends in either of them. However, such explicit simulations of protein behavior are inaccessible to current computing technology by at least several orders of magnitude [163]. Because of this, free energy differences in conjunction with DSF modeling are normally approximated by treating the two states separately, and subtracting their resulting energies. Unfortunately, the problems associated with using DSF models are even more severe when treating the unfolded state. All of the same assumption and approximations still apply, but an additional problem is caused by the absence of explicit structural information about the unfolded state. Without an explicit backbone structure (or a set of backbone structures), it is hard to account for the contributions to energy arising from pairwise interactions between amino acids. For this reason, this contribution is most often simply ignored with the idea that amino acid-to-amino acid interactions in the unfolded state are negligible due to the lack of persistent structure. Thus, a common way to

model the unfolded state in conjunction with the DSF approach is to only account for local side chain-to-backbone interactions [34, 185]. This is a severe approximation, as even in the absence of persistent structure, the topological constraints imposed by the protein sequence imply that significant contributions from amino-acid pairs can still be present. Additionally, it has been shown that in some instances the unfolded state consists of an ensemble of partially folded structures [81]. Finally, it has been demonstrated that pairwise interactions in the unfolded state can play important roles in protein stability [118]. In my work modeling bZIP interaction preferences (chapter 2), I have shown that unfolded state energies based on such modeling significantly hurt the performance of structure-based methods, relative to not modeling the reference state at all (i.e. all sequences have the same free energy in the unfolded state). Additionally, I have shown that simple scaling of amino acid-to-amino acid interactions in the folded state, based upon whether these interactions can also occur in the unfolded state, may be an appropriate, though crude, way of accounting for some pairwise interactions in the reference state.

1.0.5 Systematic Reduction of Complexity

The approximations in the energy treatment of the DSF representation of proteins, although sometimes quite severe, make the models computationally tractable. However, it is not clear that the particular set of approximations currently used in the field are an optimal tradeoff between tractability and physical realism. In fact, often these approximations, such as the unfolded state approximations, are made in an arbitrary fashion by “necessity”. One would prefer an approach to reducing a model’s complexity that is more rigorous and in which the effect on accuracy of a given assumption is known. In chapter 3, I present a new approach (cluster expansion or CE) that can potentially serve this purpose. Using this method, protein energies can be analytically approximated in terms of reduced representations of proteins, such as the DSF representation. What is attractive about this approach is that instead of making arbitrary assumptions, it systematically expands the quantity of interest (here protein energy) in terms of the reduced degrees of freedom (such as rotamer

states or even amino-acid states). This expansion can be made arbitrarily accurate by accounting for higher-order terms. For example, as I show in chapter 3 and reference [61], fixed-backbone energies from a DSF model can be expressed solely in terms of amino-acid variables, although for some systems this requires the presence of either triplet or even quadruplet interactions between amino acids. In principle, a similar approach can be taken to express a more accurate measure of energy that normally could not be used with the DSF representation (such as the electrostatic energy given by the Poisson equation) in terms of rotameric states of amino acids, thereby making it consistent with the DSF framework.

So far, I have only applied CE to expand standard DSF-like models in terms of amino-acid sequence, which retains all of the problems of accuracy associated with DSF models. However, due their tremendous reduction in complexity, such sequence-based models can be used to solve significantly more challenging problems than the original structure-based DSF models. One example, which I explore in chapter 5, is systematic computational design of protein interaction specificity. Design of specificity involves the selection of protein sequences that preferentially stabilize a structural state relative to a number of competing states. For situations where only one state is considered, efficient algorithms exist that can find the optimal sequence for stabilizing that state as well as its rotameric structure [42, 56, 58, 101, 105, 143]. These algorithms, however, require that the energy of the state be decomposable in terms of rotamer pair contributions. When considering several states, the expression to optimize involves energies for all of the states. Such expressions, in general, are not pairwise decomposable and thus sequence selection in computational specificity design has to be performed with non-optimal searching techniques. By simplifying the energy model to involve only sequence degrees of freedom, I was able to formulate the problem of specificity design in a manner than can be solved exactly (see chapter 5).

In this thesis, I analyze several aspects of DSF modeling, propose methodological improvements, and explore applications of this type of theory to the analysis and design of protein-protein interactions.

Chapter 2

Structure-based prediction of bZIP partnering specificity

Predicting protein interaction specificity from sequence is an important goal in computational biology. We present a model for predicting the interaction preferences of coiled-coil peptides derived from bZIP transcription factors that performs very well when tested against experimental protein microarray data. We used only sequence information to build atomic-resolution structures for 1,711 dimeric complexes, and evaluated these with a variety of functions based on physics, learned empirical weights or experimental coupling energies. A purely physical model, similar to those used for protein design studies, gave reasonable performance. The results were significantly improved when helix propensities were used in place of a structurally explicit model to represent the unfolded reference state. Further improvement resulted upon accounting for residue-residue interactions in competing states in a generic way. Purely physical structure-based methods had difficulty capturing core interactions accurately, especially those involving polar residues such as asparagine. When these terms were replaced with weights from a machine-learning approach, the resulting model was able to correctly order the stabilities of over 6,000 pairs of complexes with greater than 90% accuracy. The final model is physically interpretable, and suggests specific pairs of residues that are important for bZIP interaction specificity. Our results illustrate the power and potential of structural modeling as a method for predicting protein

interactions and also highlight obstacles that must be overcome to reach quantitative accuracy using a de novo approach. Our method shows unprecedented performance predicting protein-protein interaction specificity accurately using structural modeling and suggests that predicting coiled-coil interactions generally may be within reach.

2.1 Introduction

The number of interactions that occur among human proteins has been conservatively estimated as $\sim 40,000$ - $200,000$, and may be many-fold higher [21]. It will be a long time before these interactions are measured directly with reliable methods and even longer until structural detail can be assigned to all protein complexes experimentally. The need for computational methods to address these problems - to predict protein-protein interactions and to provide useful structural models of them - is clear. But there are significant challenges [169]. Although considerable progress has been made in the past 5-10 years, predicting the structure of a protein from its sequence remains an unsolved problem. Even in cases for which the overall fold is known, high-resolution details that determine protein-protein interaction specificity continue to elude state-of-the-art methods [125]. Docking proteins of known structure is now feasible in many cases, particularly in the absence of large conformational changes [127], but this is not yet an approach that has practical utility for supplying new interaction data.

A variety of strategies are being pursued to address these problems. High-accuracy models are likely to require all-atom representations and physics-based energy functions, and several groups have developed such approaches for modeling the energetics of protein-protein interactions [72, 83, 96, 103, 185]. Kortemme *et al.* [89] as well as Guerois *et al.* [62] have presented empirical energy functions that are fast to evaluate and that can be used to predict the effects of point mutations on protein stability or protein-protein interaction affinity, given high-resolution structural data. Both approaches rely on fitting a combination of physical and statistical terms to a dataset of point mutation energies for proteins with available crystal structures. The estimated accuracy of both methods is in the range of ~ 0.8 to 1.1 kcal/mol per single, con-

servative, amino-acid change, which is good enough to be practically useful for some applications. The precision of these approaches, however, comes at the cost of extensive scaling that reduces physical interpretability. Additionally, because the databases used in these methods contain proteins of many different classes, it is difficult to tell whether the non-uniform scaling of energy terms is due to general shortcomings of the models or whether certain underlying assumptions are more appropriate for some structural classes as opposed to others.

A few groups have begun to take a high-resolution homology modeling approach to predicting protein-interaction specificity. Aloy and Russell modeled the interactions of fibroblast growth factors with their receptors and were able to classify the affinity of different hormone/receptor pairs as low or high with some success [6]. Kiel *et al.* modeled the interactions of Ras-binding domains using available Ras/effector crystal structures and found good agreement with experimentally determined binding affinities [84]. Systematic methods for high-throughput modeling of complexes based on known structures are also being developed [5, 114].

We are pursuing a “bottom-up” strategy for predicting protein interaction specificity, in which we consider a single protein motif at a time, in high detail. This general approach has been explored for SH2, SH3 and PDZ domains, with some successes in classifying different types of ligands [12, 24, 82, 187, 191, 195]. We tackle the problem for the α -helical coiled coil, which is possibly the most prevalent interaction/oligomerization domain in all of biology. Coiled coils consist of two or more alpha helices wrapped into a bundle with a slight superhelical twist. The high structural symmetry of the motif is encoded by an underlying amino-acid heptad repeat $(\mathbf{abcdefg})_n$ that contains hydrophobic residues at most **a** and **d** positions. A considerable amount is known about the folding and dynamics of coiled coils as well as the effect of many mutations on their stability and interaction specificity. The coiled coil has also been a popular model system for computationally designing interaction specificity [68, 168]. Among many other biological roles, the coiled coil provides a key structural and dimerization element in two important classes of eukaryotic transcription factors - the bZIP and the bHLHZ proteins [76, 109]. In this paper, we explore

the ability of structure-based modeling to capture the interaction preferences of the bZIP coiled coils, using this example as a model for how motif-specific approaches may provide a route towards computational annotation of the protein interactome.

Basic region leucine zipper (bZIP) proteins bind DNA as homo- or hetero-dimers [76, 180] and have been implicated in numerous processes including cell proliferation [9], response to cytokine stimulation and development [19, 63, 71, 76, 106, 146]. These proteins share a homologous domain consisting of a region rich in basic residues followed by a coiled coil. Several crystal structures have confirmed that the basic region is responsible for binding DNA and that the coiled coil mediates dimerization [48, 55]. The coiled-coil region is frequently referred to as a “leucine zipper” because the majority of heptad d positions in known bZIPs are occupied by leucines. By encoding dimerization preferences, the leucine zipper region helps to determine DNA binding specificity [63, 91]. The human genome contains ~ 53 unique bZIP domains allowing for the potential formation of $\sim 1,431$ unique bZIP dimers [134, 172, 179]. A significant amount of work has been directed towards describing interactions among specific bZIP family members, as well as towards understanding dimerization specificity by experimentally measuring the strengths of key interactions [3, 19, 40, 64, 71, 94, 179]. More recently, Newman and Keating measured the interactions between nearly all pairs of human and some yeast bZIP leucine zippers using a protein microarray technique [134]. They showed that only $\sim 15\%$ of all possible dimers actually form, demonstrating that bZIPs are highly specific in choosing binding partners. Because the dimerization preferences are encoded by the well-studied and structurally conserved coiled-coil domain, this dataset provides an ideal framework for studying structural determinants of protein stability and interaction specificity.

The goal of our present study was to derive a physically realistic model that accounts for the observed pattern of bZIP coiled-coil binding preferences and to use it to understand the physical basis of these interactions and their specificity. Such a model is likely to have utility for treating coiled coils other than those found in bZIP proteins. Several models for predicting the interaction specificity of bZIPs have been proposed. Vinson and co-workers have experimentally measured coupling en-

ergies for many important interactions in two-stranded coiled coils [3, 40, 94] and have shown that these correlate with whether coiled coils homo or heterodimerize [179]. This approach is powerful, and captures many important trends in interaction preferences. A possible weakness is that it cannot easily take into account the context-dependence of residue-residue interaction strengths. In addition, because the required experiments are demanding, not all interactions have been measured. Singh and co-workers have similarly assumed context-independence of a larger set of important pairwise residue-residue interactions and optimized their relative weights with a machine-learning method [52]. This model performs very well for predicting bZIP coiled-coil interactions, but it suffers from a lack of interpretability, as the molecular structure and the physics that give rise to the predicted specificity are not addressed. Further, machine-learning models of this type require large amounts of data for training.

Ideally, the relative stabilities of different complexes could be evaluated using first principles directly from models of their structures. There are many different structure-based energy functions that are commonly used in protein stability prediction [122]. Explicit physical models that capture effects such as packing, electrostatics, hydrogen bonds and desolvation have the potential of being most interpretable. However, explicit consideration is not possible when the underlying structural models are not available or are not sufficiently accurate, limiting the applicability of such an approach. In this work, we have predicted high-resolution structures of bZIP coiled-coil domains and compared the ability of several physically motivated energy functions to explain the interaction specificities of these proteins. We found that the energy of the unfolded state, as well as the strengths of certain core interactions, are particularly difficult to capture with explicit structural considerations and are much better accounted for implicitly. Comparing the predictions of our final model to the experimental results obtained by Newman and Keating [134], we found that for pairs of bZIP dimers observed to have significantly different stabilities, the model predicts the correct order of stability in over 95% of cases. When all pairs of bZIP dimers with a consistent observed difference in stability are considered (regardless of the magnitude

of the difference), the model is correct in over 80% of cases.

2.2 Results

2.2.1 Testing Framework

Throughout this work, the performance of different computational methods for predicting bZIP coiled-coil interaction preferences was tested against the experimental protein microarray data of Newman and Keating [134]. In their experiment, coiled-coil peptides printed on glass slides were probed with fluorescently labeled solution-phase peptides. The resulting data do not provide absolute measures of binding strength but can be used to reliably order the stabilities of different complexes. This is especially true when comparing complexes that shared a common solution-phase probe in the experiments (e.g. dimer AB *versus* dimer AC, where peptide A was the probe). We used those pairs for which relative stability was well determined and tested the performance of our models on the task of correctly predicting this ordering. Because ordering is easier in cases where there is a significant difference in stability, pairs of coiled-coil complexes were classified according to the difference between their fluorescence signals in the assay, ΔF . For example, when comparing complexes AB and AC, ΔF is the absolute value of the difference between the fluorescence observed for complex AB and that observed for complex AC. Twelve datasets were constructed corresponding to ΔF values ranging from 0 to 10,000. The number of data points in each of these datasets is shown in Figure 2-1 and ranges from 801 “easier” comparisons with $\Delta F > 10,000$ to 33,186 “difficult” comparisons with $\Delta F > 0$. Although datasets with lower F contain all of the pairs with higher F values, they are dominated by dimer pairs with low signal differences and similar stabilities. The ability of different computational methods to predict the correct ordering of complex stabilities was evaluated as a function of ΔF .

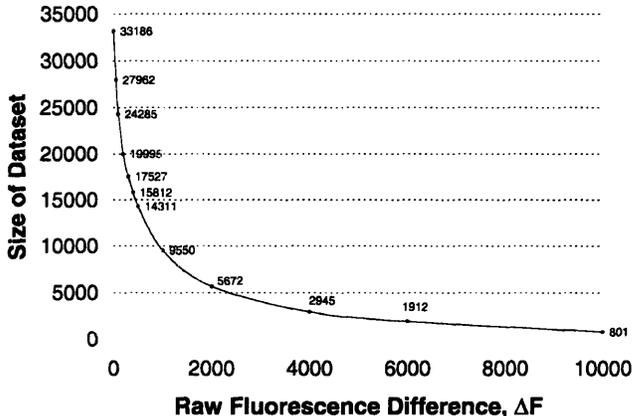


Figure 2-1: Number of coiled-coil pairs AB and AC (A, B, and C correspond to bZIP sequences) consistently satisfying $S_{AB} - S_{AC} > \Delta F$, where ΔF is the raw fluorescence signal for dimer XY observed in ref [134].

2.2.2 General Modeling Procedure

Our approach for computing leucine-zipper interaction preferences involved predicting the structures of complexes, and then evaluating their relative stabilities with a series of physically motivated energy functions. For structure prediction we assumed a constant, ideal GCN4-like backbone and placed side chains onto this scaffold using a pairwise-decomposable energy function and the Dead End Elimination algorithm [42]. There are eight unique bZIP dimers with crystal structures in the PDB, and they all have similar backbone geometries. Table 2.2.2 gives the RMS deviations between these structures and the GCN4 backbone as well as our idealized model backbone. To test the ability of our method to recover a relevant structure given this generic backbone, we constructed models for three of the eight available bZIP crystal structures (those with a resolution of 2.0 Å or better) as well as several other high-resolution parallel dimeric coiled-coil structures. The frequency with which side chains were placed in close-to-native conformations was evaluated on the ideal backbone and on the wild-type backbones. Table 2.2 summarizes the results. As expected, χ -angle recovery was higher at the core **a** and **d** positions than at the surface-exposed **e** and **g** positions, but **a** and **d** sites were also more sensitive to the choice of backbone. Nevertheless, using the ideal backbone resulted in average χ_1 and χ_2 recovery rates of 82% and

Table 2.1: Coiled-coil backbone variation in members of the bZIP family.

	1CI6	1DH3	1FOS	1GD2	1GU4	1JNM	1NWQ
Backbone RMSD w/ 2ZTA (Å)	1.26	1.22	0.93	0.64	0.91	0.87	0.72
Backbone RMSD w/ ideal (Å)	1.16	1.14	0.82	0.39	0.82	0.73	0.68
Alignment length (residues)	26	26	31	26	31	27	28

70% in the core, respectively.

Structures generated using Dead End Elimination were subsequently minimized slightly to remove steric clashes, and were then evaluated using a variety of non-pair-wise-decomposable energy functions. The differences in the models tested came from how the unfolded state was treated, and how core interactions were modeled. The contribution of any amino acid aa to the stability of a protein or protein complex can be broken into two components: $\Delta G_{aa}^{\text{folding}} = (G_{aa,F}^{\text{self}} - G_{aa,UF}^{\text{self}}) + (G_{aa,F}^{\text{pair}} - G_{aa,UF}^{\text{pair}})$. The first component contains the single-residue or “self” contributions of amino acid aa to the energy difference between the unfolded and folded states. This includes changes in intra-amino acid interactions, changes in the entropy of the amino acid, and changes in the interaction between the amino acid and the protein backbone. It also includes mutual desolvation between aa and the backbone, but not between aa and other modeled side chains. The self contribution does not contain any sequence-dependent terms, but does depend on the shape of the folded backbone. The second, sequence-dependent, contribution arises from specific side chain-to-side chain interactions involving amino acid aa in the folded and unfolded structures, and the effects of other side chains on desolvating aa . Modeling the unfolded state is a challenge for computational protein folding and design. Our first four models (defined below, and referred to as EX, PF, HP and HP/S) differ in the treatment of the $G_{aa,UF}^{\text{self}}$ and $G_{aa,UF}^{\text{pair}}$ terms. Our final model, HP/S/C further modifies $G_{aa,F}^{\text{pair}}$ for the **a** and **d** positions.

2.2.3 Explicit Unfolded State

Because energy functions based on explicit structural models provide the most interpretability, we first tested the use of a structurally explicit unfolded state for predicting bZIP coiled-coil partners. Following others, we modeled the unfolded state by

Table 2.2: χ -angle recovery for placing native side chains on bZIP and non-bZIP parallel dimeric coiled-coil backbones.

	Native Backbone		Ideal Backbone	
	a, d ^a	e, g ^b	a, d ^a	e, g ^b
χ_1	94% (149/159)	64% (130/203)	82% (131/159)	62% (126/203)
χ_2	84% (105/125)	57% (102/179)	70% (87/125)	52% (93/179)

^a core **a** and **d** positions that are at least four residues removed from either end of the molecule (to avoid end effects).

^b **g** and **e** positions only.

neglecting residue-residue interactions ($G_{aa,UF}^{pair}$) and accounting only for interactions of side chains with themselves and with a local poly-Gly penta-peptide backbone [34, 185]. The resulting energy model EX is described by:

$$\Delta G^{\text{folding}} = \sum_{i < j} G_{vdW,EEF,Coul,GBscreen}^{sci-scj} + \sum_i G_{vdW,EEF,Coul,GBscreen}^{sci-t} - \sum_i G_{vdW,EEF,Coul,GBscreen}^{sci-ut} \quad (2.1)$$

where $sci - scj$ designates the interaction between side chains i and j , $sci - t$ is the interaction of side chain i with the template (all of the protein excluding modeled side chains) and $sci - ut$ is the interaction of side chain i with the local penta-peptide backbone in the unfolded state.

Figure 2-2(a) shows the performance of model EX on the task of correctly ordering pairs of dimers in terms of stability. Although the success rate of the model is reasonable for the highest fluorescence difference dataset ($\sim 80\%$), it falls off quickly and reaches $\sim 61\%$ for datasets where ΔF is small. This is quite modest given that a 50% rate can be obtained by random guessing.

We also tested the performance of a variant model that assumes single-residue terms to be the same in the folded and the unfolded states (i.e. $G_{aa,F}^{\text{self}}$ and $G_{aa,UF}^{\text{self}}$ cancel). Accordingly, model PF consists only of the term $G_{aa,F}^{pair}$. It has no explicit treatment of the unfolded state, it omits side chain-to-template interactions and only accounts for side chain-to-side chain interactions and side chain-to-side chain desolvation effects in the folded state. This produced considerably better results (see Figure 2-2(a)). A likely explanation is that the unfolded state is modeled poorly with the penta-peptide method. This is particularly interesting given that this is a

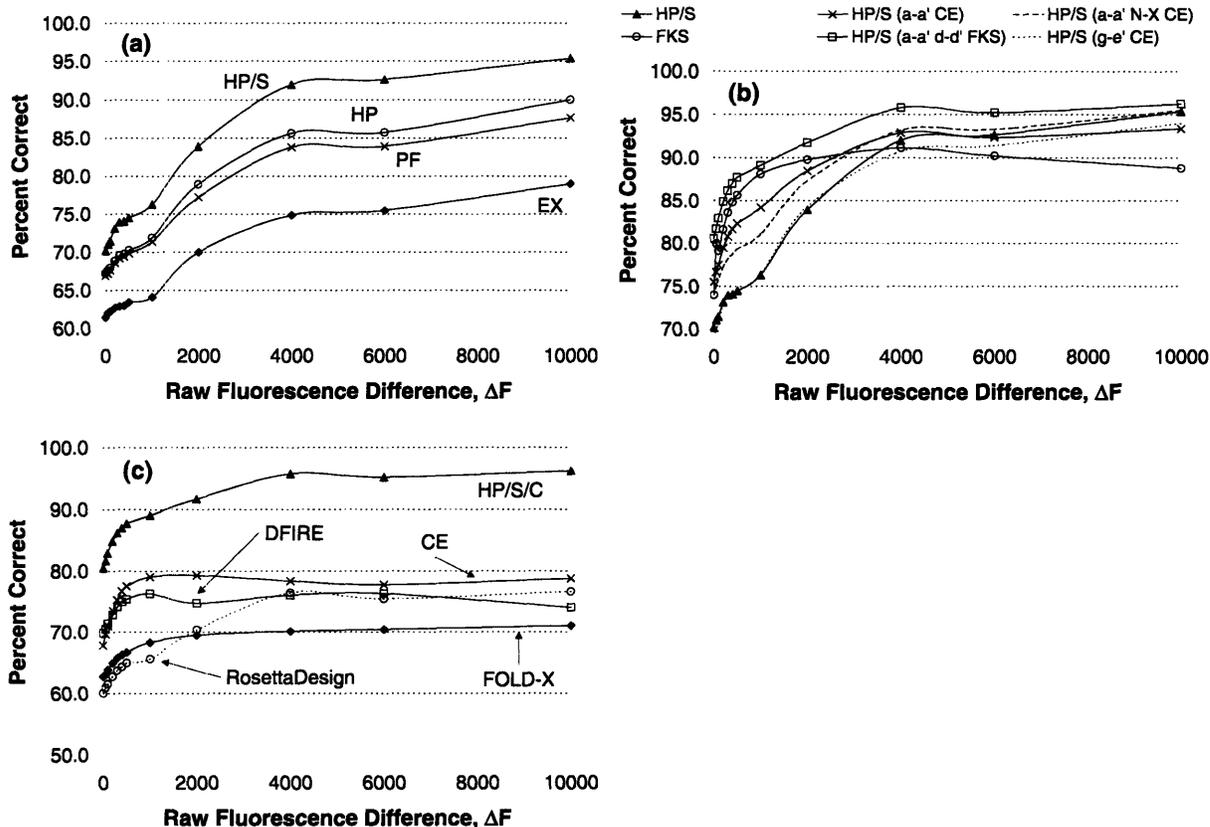


Figure 2-2: Performance of different models for predicting dimer stability differences, as a function of ΔF . The y-axis shows the percentage of correctly ordered dimer pairs in each of 12 datasets defined by a ΔF cutoff. As expected, the performance is better for dimer pairs with large experimentally observed differences in stability. (a) Model EX corresponds to equation 2.1, Model HP to equation 2.2, Model HP/S to equation 2.3 and model PF to evaluation with only the term $G_{aa,F}^{\text{pair}}$. (b) Variants of model HP/S. Model HP/S ($a - a'$ CE) corresponds to substituting all $\mathbf{a}_i \mathbf{a}'_i$ interactions X-Y, where both X and Y are one of [Val, Leu, Ile, Lys or Asn], with coupling energies experimentally determined by Acharya *et al.* [3]. Model HP/S ($a - a'$ N-X CE) is similar except of the above interactions only those involving at least one Asn were replaced. FKS corresponds to the Support Vector Machine-based method of Fong *et al.* using base weights [52]. Model HP/S ($g_i e'_{i+1}$ CE) corresponds to substituting $\mathbf{g}_i \mathbf{e}'_{i+1}$ interactions involving Glu, Gln, Arg, and Lys with coupling energies measured by Krylov *et al.* [94]. Model HP/S ($a - a' d - d'$) has all $\mathbf{a}_i \mathbf{a}'_i$ and $\mathbf{d}_i \mathbf{d}'_i$ interactions replaced with corresponding FKS weights (same as model HP/S/C). (c) The performance of model HP/S/C compared to that of DFIRE [197], Rosetta Design, FOLD-X [62] and a model based on adding $\mathbf{g}_i \mathbf{e}'_{i+1}$ and $\mathbf{a}_i \mathbf{a}'_i$ coupling energies (model CE).

popular approach and has been used in the successful design of numerous proteins [35, 117, 156].

2.2.4 Implicit Unfolded State

As an estimate of $G_{aa,F}^{\text{self}} - G_{aa,UF}^{\text{self}}$ for residues in helical environments, several investigators have measured the influence of amino-acid substitutions on helix stability [18, 29, 74, 116, 139]. The resulting helix propensity scales, though measured in different contexts and with different methods, agree with one another well. We tested two models that use helix propensities to account for changes in amino-acid self energies upon folding. In one, we assumed that these propensities adequately capture all of the self contributions to coiled-coil folding. In another, we introduced a correction factor based on the amount of interaction of each amino acid with the two-helix backbone, which varies according to coiled-coil heptad position (see Materials and Methods). The former approach performed better in all tests and was incorporated into model HP, which is described by the following equation:

$$\Delta G^{\text{folding}} = \sum_{\text{site } i} [hp_{\text{ref}} + hp(aa_i)] + \sum_{i < j} G_{vdW,EEF,Coul,GBscreen}^{\text{sci-scj}} \quad (2.2)$$

where the first term represents $G_{aa,F}^{\text{self}} - G_{aa,UF}^{\text{self}}$ and includes the sum of helix propensities of all amino acids in the dimer, while the second term accounts for all side chain-to-side chain interactions in the folded structure. The parameter hp_{ref} sets the reference point for an absolute scale and was necessary to compare sequences of different lengths. It was adjusted to optimize the number of correctly ordered dimer pairs in the dataset with $\Delta F = 6000$ (this dataset includes only 5.8% of the total number of dimer comparisons), although the particular data set used was not critical (see Figure 2-3). A significant range of hp_{ref} values gave essentially optimal performance, however inappropriate values for hp_{ref} (e.g. $hp_{\text{ref}} > -0.7$) that penalized longer coiled coils relative to shorter ones led to a significant decrease in performance. The optimal value of hp_{ref} resulted in an absolute helix propensity for Gly of 0.61 kcal/mol, favoring the unfolded state. Figure 2-2(a) shows how well this model orders pairs of

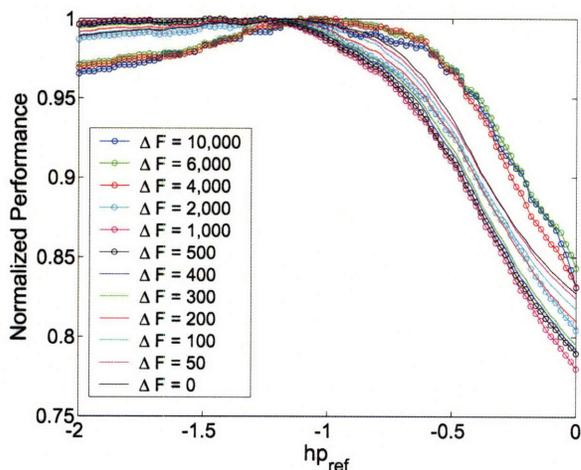


Figure 2-3: Optimizing the value of parameter hp_{ref} in the context of model HP using different dimer comparison datasets. Each line corresponds to a dimer comparison dataset with a specific value of ΔF . The y-axis shows performance (fraction of correctly ordered dimer pairs) normalized by the optimal performance for each dataset. A significant range of values ($-1.3 \leq hp_{\text{ref}} \leq -0.7$) is essentially optimal, indicating that the particular value of hp_{ref} chosen is not very important.

dimers in terms of stability. Model HP performs significantly better than model EX and also somewhat better than model PF. For 861 pairs of dimers with the largest experimentally observed differences in stability, model HP predicts the correct order in over 90% of cases.

2.2.5 Intra-helix interactions

Model HP performs well for dimer pairs that are significantly different in terms of stability. However, performance still falls off quickly as ΔF decreases. To address this, we tested the assumption that $G_{\text{aa,UF}}^{\text{pair}}$ can be ignored. This is a very common assumption made for estimating protein stability, but there is evidence that such a model for the unfolded state may be inappropriate for many coiled coils [81, 118]. We introduced a variable parameter to scale intramolecular side chain-to-side chain interactions relative to intermolecular ones, effectively introducing a pair term to

competing uncomplexed states. Model HP/S is defined as:

$$\Delta G^{\text{folding}} = \sum_{\text{site } i} [hp_{\text{ref}} + hp(aa_i)] + \sum_{\text{inter-helix pairs } i,j} G^{\text{sci-scj}} + s \cdot \sum_{\text{intra-helix pairs } i,j} G^{\text{sci-scj}} \quad (2.3)$$

where s is the intra-chain interaction scale factor. The last two sums in the equation capture $G_{\text{aa,F}}^{\text{pair}} - G_{\text{aa,UF}}^{\text{pair}}$ by assuming that side-chain to side-chain interactions occurring in the folded state also occur, to some degree, in the reference state. Figure 2-2 shows that model HP/S significantly outperforms both of the two previous models. For 2,945 dimer pairs with $\Delta F \geq 4,000$, the model predicts the correct order of stability in over 90% of cases. For all dimer pairs in the test, the model predicts the correct order of stability in 70% of cases.

The two adjustable parameters in equation 2.3, hp_{ref} and s , were chosen by optimizing the performance of the model on the dimer comparison dataset with $\Delta F = 6000$. As before, no clearly optimal value for hp_{ref} was observed, rather, a significant range $-1.8 \leq hp_{\text{ref}} \leq -0.8$ produced near optimal results (the value of 1.08 was used). The same was true for s (Figure 2-4). Interestingly, the optimal range of $-1.5 < s < 0$ ($s = -0.7$ was used) suggested that favorable side-chain interactions within the same helix may actually reduce coiled-coil stability. Optimizations using the remaining eleven dimer comparison subsets gave slightly different optimal values, but the ranges stayed essentially the same, with s always negative (Figure 2-5).

2.2.6 Alternative Models for Core Interactions

Model HP/S exhibits good performance overall, but shows a strong dependence on ΔF below 4,000. To explore the origins of this effect we investigated interactions known to be important for determining coiled-coil stability and specificity. These include $\mathbf{a}_i\mathbf{a}'_i$ and $\mathbf{d}_i\mathbf{d}'_i$ interactions that form the hydrophobic core and are essential for dimer stability, and $\mathbf{g}_i\mathbf{e}'_{i+1}$ interactions, which can contribute to both stability and specificity (the prime denotes the position on the opposing monomer and subscripts refer to heptad index) [179]. Vinson and co-workers have experimentally determined coupling energies for 19 pairs of amino acids at $\mathbf{a}_i\mathbf{a}'_i$ positions [3, 40] and 16 pairs of

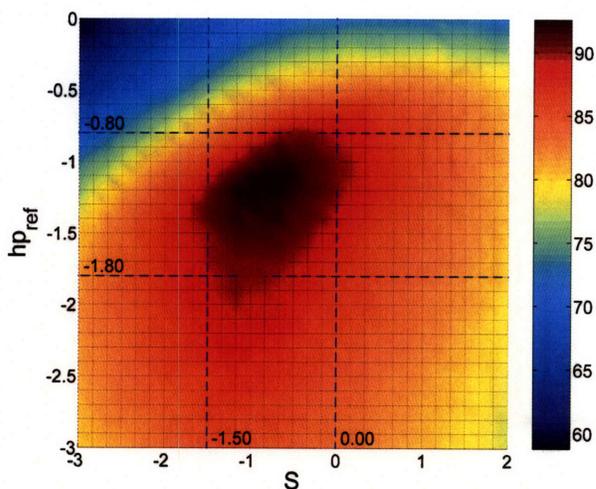


Figure 2-4: Performance of model HP/S on the dimer comparison dataset with $\Delta F = 6000$ as a function of hp_{ref} and s . Color, as shown in the key at right, indicates the percentage of correctly order dimers. The optimal performance is 92.7% and the straight, dashed lines indicate approximately where performance is better than 90%. A range $-1.8 \leq hp_{\text{ref}} \leq -0.8$ gives essentially optimal performance. The same is true for s . Strikingly, however, the range of optimality for s lies entirely in the negative region $-1.5 \leq s \leq 0.0$ implying that intra-helix interactions either do not contribute or contribute negatively to coiled-coil stability.

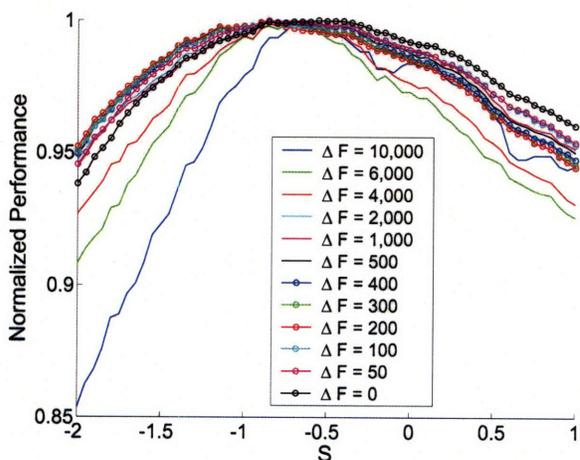


Figure 2-5: Optimizing the value of parameter s on different dimer comparison datasets gave very similar optimality ranges. For simplicity of analysis, hp_{ref} was set to 1.08 kcal/mol (the value used in model HP/S) although optimizing hp_{ref} simultaneously with s produced very similar values of hp_{ref} for different datasets. Each line corresponds to a dimer comparison dataset with a specific value of ΔF . The y-axis shows performance (fraction of correctly ordered dimer pairs) normalized by the optimal performance for each dataset.

amino acids at $\mathbf{g}_i\mathbf{e}'_{i+1}$ positions [94] using alanine double mutant cycle analysis. In Figure 2-6 we compare their values with average values of our computed interaction energies. For $\mathbf{g}_i\mathbf{e}'_{i+1}$ interactions, the agreement is very good, with a correlation coefficient of 0.89. For core $\mathbf{a}_i\mathbf{a}'_i$ interactions, however, the agreement is much worse, particularly for residue pairs involving Asn. To test the influence of errors in $\mathbf{a}_i\mathbf{a}'_i$ interactions, we replaced calculated side chain-to-side chain interaction energies with experimentally determined values (only for those $\mathbf{a}_i\mathbf{a}'_i$ pairs for which these were available), leaving the rest of the energy function the same. Figure 2-2(b) shows the significant improvement in performance that results, particularly for dimer comparison data sets with low ΔF . The improvement comes almost entirely from $\mathbf{a}_i\mathbf{a}'_i$ interactions involving Asn (Figure 2-2(b)). Asn to non-Asn coupling energies are the largest among the ones measured by Acharya et al. [3], and pairing of Asn residues at opposing \mathbf{a} positions is known to be an important determinant of coiled-coil specificity [138, 179]. Surprisingly, substituting experimental coupling energies for $\mathbf{g}_i\mathbf{e}'_{i+1}$ did not result in any improvement and in fact performed slightly worse for higher ΔF datasets (see Figure 2-2(b)).

A machine-learning method for predicting coiled-coil associations based on Support Vector Machine-like optimization has recently been proposed by Fong et al. [52, 159]. This model (referred to here as FKS) assumes that each of seven important types of interactions in a parallel dimeric coiled coil ($\mathbf{a}_i\mathbf{a}'_i$, $\mathbf{d}_i\mathbf{d}'_i$, $\mathbf{a}_i\mathbf{d}'_i$, $\mathbf{d}_i\mathbf{a}'_{i+1}$, $\mathbf{d}_i\mathbf{e}'_i$, $\mathbf{g}_i\mathbf{a}'_{i+1}$, $\mathbf{g}_i\mathbf{e}'_{i+1}$) can be assigned an additive weight based on the amino-acid identities at the interacting sites. These weights are optimized by training the model on known coiled-coil sequences, hypothesized non-interactions and information from biophysical studies, but not on the human bZIP array data [52]. The method performs very well on the pair ordering test, as shown previously and in Figure 2-2(b) [52]. Although model HP/S outperforms the FKS model for dimer pairs with significantly different stabilities, the FKS method does better at separating dimers of more similar stability. We observed above that better core interactions taken from experiment improved the performance of model HP/S for low- ΔF dimer pairs, but experimental values are not available for all of the important core interactions. To test whether remaining inac-

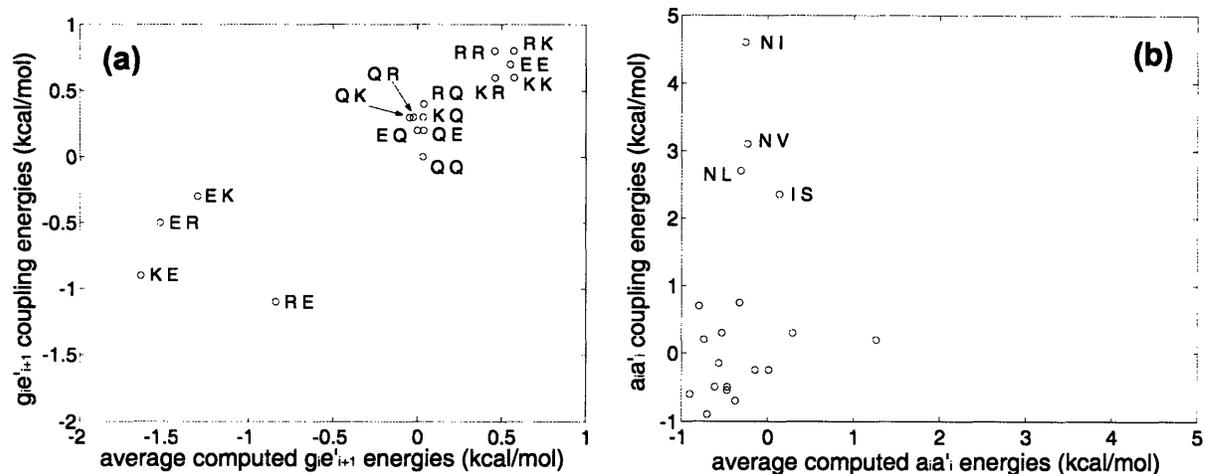


Figure 2-6: Comparison of (a) $g_i e'_{i+1}$ and (b) $a_i a'_i$ coupling energies measured by Vinson and co-workers [3, 94] with the corresponding computed interactions. The latter were calculated as the average of the particular type of interaction in all of the modeled bZIP dimers.

curacies in the model still had to do with inadequate modeling of core interactions, we replaced all computed $a_i a'_i$ and $d_i d'_i$ interactions in model HP/S with weights from the FKS model. Although these weights do not represent physical energies, they do capture the tendency of specific $a_i a'_i$ and $d_i d'_i$ pairs to stabilize coiled-coil complexes. Further, experimental $a_i a'_i$ coupling energies and FKS weights are of similar magnitude and rank ordering, although they do not agree quantitatively (Figure 2-7). Reasonable agreement is expected, because the coupling energies were used in training the FKS model [52, 159]. Figure 2-2(b) shows the performance that resulted from replacing all core $a_i a'_i$ and $d_i d'_i$ interactions in model HP/S with FKS weights (model HP/S/C). Weighting the FKS terms relative to the terms computed based on structure did not give a noticeable improvement in performance. Model HP/S/C outperforms all of the models discussed above as well as the FKS model itself. More than 80% of the 33,186 dimer pairs in the test set are correctly ordered. For 2,945 significantly different dimer pairs (with $\Delta F > 4,000$) the model predicts the correct order in over 95% of cases.

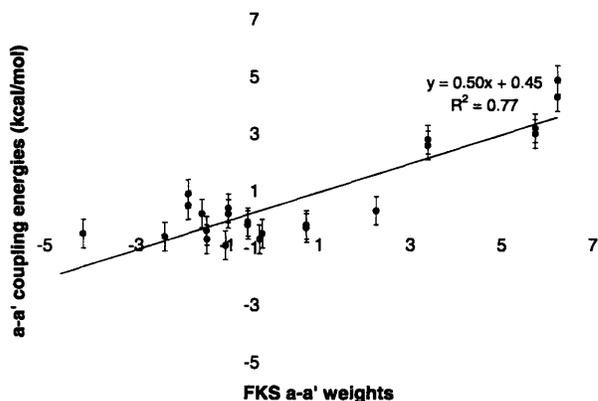


Figure 2-7: Comparison of $a_i a'_i$ coupling energies measured by Vinson and co-workers [3] with corresponding FKS weights. Vertical bars represent the estimated experimental error of 0.5 kcal/mol.

2.2.7 Comparison to Other Methods

Several general-purpose methods have recently been developed to evaluate the binding energies of protein complexes given their structures. We applied some of these to the coiled-coil pair ordering test to provide an unbiased measure of the difficulty of the problem. DFIRE is a statistical potential that has been reported to predict protein-protein binding energies in good agreement with experimental measurements [111]. Figure 2-2(c) shows the performance of DFIRE on ordering bZIP dimer pairs using our predicted structures. Model HP/S/C significantly outperforms DFIRE in this test.

FOLD-X is a method for estimating protein stability and protein-protein interaction strength developed by Guerois *et al.*, who have shown it to be effective in predicting single amino-acid mutation energies in proteins as well as in protein complexes [62]. We used FOLD-X to score and order structures predicted with our methods. Figure 2-2(c) shows that model HP/S/C performs significantly better than FOLD-X for ordering bZIP coiled-coil pairs. It is possible that DFIRE and FOLD-X would show better performance on native structures, but these are not available.

We also compared the performance of model HP/S/C with that of the RosettaDesign algorithm of Baker and co-workers [95, 37]. The energy function underlying this method is very effective for both protein structure prediction and design [96, 133].

Figure 2-2(c) compares the performance of model HP/S/C and RosettaDesign on predicting dimer order stability. In this test RosettaDesign was used to predict the coiled-coil structures given a fixed ideal backbone. Thus, unlike DFIRE and FOLD-X, this method was not constrained to predict the energies of structures obtained with our procedure. Nevertheless, Model HP/S/C significantly outperforms RosettaDesign at all values of ΔF .

Vinson and co-workers have proposed that coiled-coil dimerization preferences can be explained in terms of experimentally measured coupling energies of key $\mathbf{g}_i\mathbf{e}'_{i+1}$ and $\mathbf{a}_i\mathbf{a}'_i$ interactions [3, 40, 94, 179]. In Figure 2-2(c) we also show the performance of a model that estimates stability using the sum of the available $\mathbf{g}_i\mathbf{e}'_{i+1}$ and $\mathbf{a}_i\mathbf{a}'_i$ coupling energies for each dimer (model CE). This model performs slightly better than the three structure-based models described above but is still significantly worse than Model HP/S/C.

2.2.8 Interaction *versus* non-interaction Discrimination

As an additional test of model HP/S/C we examined its ability to discriminate interacting from non-interacting dimers. This is a more challenging test than ordering dimer pairs by stability. The dimer comparison datasets consisted only of pairs involving a common peptide partner, because these are most reliably ordered by the experiments. Discriminating interactions from non-interactions more generally requires comparison of dimers of completely different composition. Figure 2-8 shows the ability of model HP/S/C to differentiate interactions from non-interactions. Each column corresponds to a particular peptide interacting with all tested peptides. There is a clear overall tendency for interacting dimers to have lower energies by model HP/S/C, but there is no single energy cutoff that cleanly separates interacting pairs from non-interacting pairs. At a cutoff of -50 , 60% of true interactions are detected with 79% of predicted interactions correct. Higher coverage can be achieved at the cost of specificity: at a cutoff of -45 , 79% of true interactions are detected with 52% of predicted interactions correct.

Within a bZIP sequence family, cutoffs for distinguishing interactions from non-

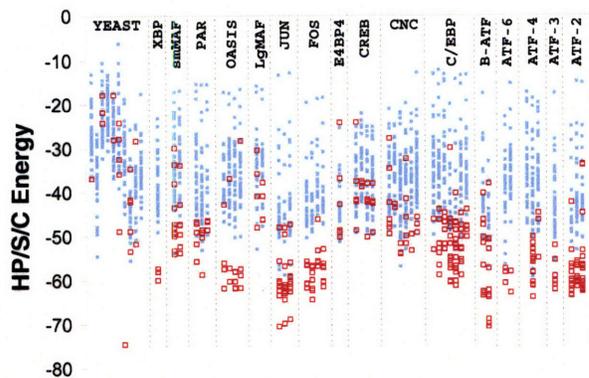


Figure 2-8: Performance of model HP/S/C on discriminating between interacting and non-interacting leucine zippers. Red dots signify interactions and blue dots non-interactions, as determined experimentally. Each column represents interactions with a single probe. Probes are sorted by family along the x-axis. The energy from model HP/S/C is on the y-axis.

interactions are easier to define. In Figure 2-8, dimers are sorted by the sequence family of the probe, illustrating that model HP/S/C discriminates extremely well between interactions and non-interactions within most families. To provide a quantitative analysis of the model’s performance in the discrimination test, we constructed receiver-operator characteristic (ROC) curves for each family of interactions and computed the average ROC curve, shown in Figure 2-9. We compared this to the performance of the FKS and CE models. Model HP/S/C performs comparably to the FKS model and significantly better than the CE model.

2.2.9 Implicit *versus* explicit models of side chain-side chain interactions

We investigated further why our structurally explicit model does not describe $\mathbf{a}_i\mathbf{a}'_i$ and $\mathbf{d}_i\mathbf{d}'_i$ interactions accurately, using experimental data that characterize the important role of Asn at a sites. Using a designed coiled coil with a Val-Val (V-V) interaction at a central pair of $\mathbf{a}_i\mathbf{a}'_i$ sites, Vinson and colleagues found that substitution of a single Asn residue (giving V-N) is destabilizing by 5.2 kcal/mol while substitution of V-V with N-N is destabilizing by only 3 kcal/mol. Further analysis gave coupling energies for V-V, V-N and N-N of -0.7 , $+3.0$ and -0.5 , respectively [3]. Other studies have

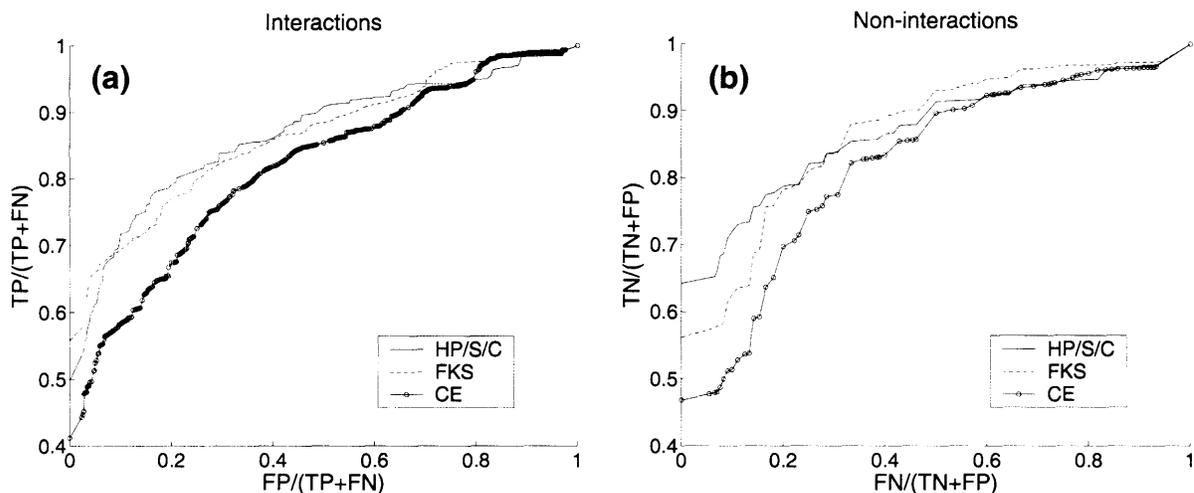


Figure 2-9: ROC analysis of performance predicting interacting and non-interacting leucine zippers. The family-averaged ROC curves are shown for model HP/S/C, the FKS model and a coupling energy model (CE) in solid lines, dashed lines, and lines with open circles, respectively. **(a)** Fraction of interacting dimers identified by the models ($TP/TP+FN$) as a function of the number of dimers incorrectly identified as interacting ($FP/FP+TN$). **(b)** Fraction of non-interacting dimers identified as such by the models ($TN/TN+FP$) as a function of the number of dimers incorrectly identified as non-interacting ($FN/FN+TP$). TP: true positives; FN: false negatives; TN: true negatives; FP: false positives.

also made it clear that there is a strong preference for Asn residues to be paired at $\mathbf{a} - \mathbf{a}'$ positions in parallel, two-stranded coiled coils [137, 196]. We tested the ability of different models to capture the relative stabilities $N-N > V-N$ and $V-V > V-N$.

We used the well-studied bZIP homodimer GCN4 (PDB ID 2ZTA) and calculated the contribution of \mathbf{a} and \mathbf{a}' residues to the rigid-body binding energy of this complex. Although this is a poor approximation of real folding pathways, it isolates deficiencies of the energy function from deficiencies of an unfolded-state approximation. We constructed a series of models where V-V, V-N or N-N pairs were substituted at either site 9 (V-V in native GCN4) or site 16 (N-N in native GCN4). The V-N pair is best accommodated at site 16. Table 2.3 shows a decomposition of van der Waals, electrostatic and desolvation contributions to binding. A possible explanation for the observed preference for N-N over V-N, that is also consistent with the greater stability of V-V than N-N, is that the desolvation penalty associated with burying an Asn group is large, but that this cost is more than compensated for if the buried Asn can interact with another Asn at the opposing \mathbf{a}' position. However, electrostatic interaction and desolvation energy differences between N-N and V-N fail to support this argument. It costs ~ 1.3 kcal/mol in solvation energy to bury the single Asn residue in a V-N pair, but the desolvation energy to bury a second Asn in an N-N pair ($\sim 1.7-1.9$ kcal/mol) almost exactly cancels the favorable interaction between the two Asn side chains, indicating that N-N and N-V are similar in terms of electrostatics. The desolvation costs would likely be even larger given a more realistic model of the unbound state. Including electrostatic interactions with the rest of the protein favors N-V pairing over N-N slightly, increasing the discrepancy with experimental observations. Poisson-Boltzmann calculations done on the same system by Hendsch and Tidor support the idea that the Asn-Asn interaction energy barely compensates (at best) for the cost of desolvating a second Asn residue at site 16 [70].

Including van der Waals packing energies fails to bring the calculations into agreement with experiment. Our energy function favors Asn over Val at site 16, where this is the native residue in GCN4: one Asn is good, two are better. At site 9, N-N and V-V are equally good, and N-V is less favorable. Notably, extreme sensitivity

to backbone structure makes it difficult to evaluate the van der Waals contributions accurately. There is nothing in the packing analysis, however, to suggest that V-V or N-N are strongly preferred over V-N, as is observed experimentally.

Poor performance modeling $\mathbf{a}_i\mathbf{a}'_i$ and $\mathbf{d}_i\mathbf{d}'_i$ interactions could be due to the fixed backbone approximation [83]. These interactions represent points of closest contact between the two monomers and slight backbone variation could lead to large energy differences. Side-chain minimization relieved this problem to some extent, but to test whether further relaxation gave improved performance we tested two different backbone relaxation methods: 1) unconstrained minimization in CHARMM and 2) fixed-backbone minimization using a family of ideal coiled-coil backbones. Neither of the approaches gave a significant improvement in $\mathbf{a}_i\mathbf{a}'_i$ and $\mathbf{d}_i\mathbf{d}'_i$ interactions or the performance of the overall model.

Unlike the situation for core interactions, for surface $\mathbf{g}_i\mathbf{e}'_{i+1}$ interactions our explicit physical model gave better performance than either experimentally measured coupling energies or statistical weights. Replacing the computed $\mathbf{g}_i\mathbf{e}'_{i+1}$ interactions with experimental ones does not improve the performance of the model (Figure 2-2(b)). An explanation for this could be that the strength of surface interactions is context-dependent, however we found that the average over all computed $\mathbf{g}_i\mathbf{e}'_{i+1}$ interactions for a given pairs of residues can be substituted for the explicitly calculated ones with little or no degradation in performance.

2.3 Discussion

We have developed a physically motivated method for predicting relative coiled-coil association strengths and specificities. Our models perform better than several others, some of which were developed for scoring coiled coils and others for more general purposes. Our final model HP/S/C separates interacting from non-interacting coiled coils and can order a large number of dimer pairs in terms of stability, in agreement with experimental protein microarray experiments (Figures 2-2(c) and 2-8). It accomplishes this without experimental knowledge of the structures of the complexes.

Table 2.3: Calculated contribution of N-N, N-V, or V-V at $\mathbf{a}_i\mathbf{a}'_j$ positions to the rigid-body binding energy of the coiled coil GCN4.

Sites ^a	Minimiz. procedure ^b	van der Waals		Electrostatics		EEF1 desolvation
		$(i-j)+(i-t)+(j-t)^c$	$(i-j)^c$	$(i-t)+(j-t)^c$	i,j self ^c	$(i-j)+(i-t)+(j-t)^c$
A16-B16 native N-N	None	-10.3	-2.2	-2.8	3.3	20.6
	EEF conv	-16.2	-1.9	-2.7	3.1	20.4
A16-B16 rep N-N	None	-13.7	-1.4	-2.0	3.2	21.4
	EEF conv	-14.2	-1.8	-2.0	3.0	21.2
A16-B16 rep N-V	None	15.6	0.0	-2.6	1.4	15.8
	EEF conv	-10.2	0.0	-3.4	1.3	14.9
A16-B16 rep V-V	None	15.0	0.0	0.0	0.0	9.8
	EEF conv	-5.8	0.0	0.0	0.0	9.0
A9-B9 native V-V	None	-9.2	0.0	0.0	0.0	7.1
	EEF conv	-8.0	0.0	0.0	0.0	7.9
A9-B9 rep V-V	None	-7.9	0.0	0.0	0.0	8.3
	EEF conv	-9.1	0.0	0.0	0.0	8.0
A9-B9 rep N-V	None	-6.0	0.0	-2.8	1.3	11.6
	EEF conv	-8.0	0.0	-2.8	1.1	11.5
A9-B9 rep N-N	None	-6.9	0.1	-4.3	3.3	18.8
	EEF conv	-9.2	0.0	-4.0	2.7	16.8

^a Residue positions in the structure 2ZTA, and the amino acids modeled at these sites: native indicates residues from the x-ray structure, rep indicates models repacked as specified in the Methods.

^b Relaxation procedure used: either no minimization or minimization in CHARMM using the EEF1 energy function until convergence.

^c i and j are the two modeled sites in column 1. $i-j$ indicates the interaction between side chains at the opposing $\mathbf{a} - \mathbf{a}'$ positions considered. $i-t$ indicates the difference between the interaction of side chain i with the remainder of the structure (the template) in the dimeric and the monomeric states. For EEF1, the notation $x-y$ indicates the mutual desolvation of x and y . For the electrostatic term i self is the change in the total side-chain self electrostatic energy upon binding for the two modeled side chains (reaction field plus screened intra-side chain interactions).

Our development of these models highlighted several limitations of existing computational methods and ultimately led us to an approach that combined explicit and implicit treatments. Three main insights led to significant increases in performance.

First, a popular penta-peptide model for the unfolded state did not perform well, as evidenced by a significant improvement upon either simply removing it, or replacing it with a helix propensity term (Figure 2-2(a)). The unfolded state is an ensemble of many conformations, some of which may have significant local structure. Thus, the penta-peptide model likely underestimates interactions and overestimates solvent accessibility. Furthermore, the “unfolded” state is not the only alternative state that must be considered; any significantly populated competing structure can affect the stability of a bZIP complex [129]. The simple explicit unfolded state models commonly used for design also ignore other potentially important effects. For example, it is difficult to accurately estimate loss of side-chain entropy upon folding. Rather than try to address such complex contributions explicitly, we incorporated experi-

mental measurements of approximately the same terms [65]. We found that a helix propensity model for the unfolded state worked much better than an explicit one.

Second, we found that intra-helix interactions between side chains may have an important role not only in the coiled-coil state but also in competing dissociated states (such as the unfolded state). Single-residue helix propensities contributed positively to overall coiled-coil stability in our models, improving overall performance. However, full-helix folding energies (similar to those computed with AGADIR [97], see Methods), which included side chain-to-side chain interactions in addition to intrinsic helix propensities, showed a decreased ability to order dimer pairs by stability (data not shown). In fact, when we scaled intra-helix side chain-to-side chain interactions relative to inter-helix ones, we found that optimal performance arose from negative scaling (Figures 2-3 and 2-5). This suggests that intra-chain interactions in alternative states may be an important consideration for modeling coiled-coil folding.

There is experimental and theoretical support for this idea. Marti and Bosshard [118] have demonstrated that repulsive electrostatic interactions in the unfolded state give rise to a large difference in how a heterodimeric coiled coil and its disulfide-linked counterpart respond to changes in pH. Kammerer *et al.* have shown that certain regions of at least some coiled-coil monomers are structured before the dimerization step [81]. Theoretical calculations by Myers and Oas suggest that significant partial helicity in monomers of the yeast transcription factor GCN4 is likely, and a model based on this agrees well with folding kinetics and the effects of several Ala to Gly substitutions [132]. Thus it appears that the random-coil model may not be appropriate for the unfolded state of coiled coils. More generally, Fitzkee and Rose have suggested that ensembles of primarily structured proteins can capture experimentally measurable properties of the unfolded state as well as a random-coil model [51].

The fact that a model based on scaling down intra-chain interactions improves performance suggests that this strategy may be an effective way of accounting for pair interactions in the collective reference state (i.e. the set of all competing states). Model HP/S does not necessarily assume that the competing states are helical. Rather it maintains that residues interacting within the same chain in the folded structure

may also interact in other states (such as the unfolded state) and therefore need to be down-weighted. The fact that the optimal value for s is negative indicates that the amount of intra-chain interactions in the competing states may be as great or greater than in the folded coiled-coil state, perhaps because there are no competing inter-chain side-chain interactions. It is likely that the appropriate value for s will depend on the system studied and the actual sequence considered. However, our analysis of bZIP coiled coils shows that although the optimal value for the parameter s does depend somewhat on the particular subset used for fitting, these differences are small, and the optimal s is always negative.

Third, we found that $\mathbf{a}_i\mathbf{a}'_i$ and $\mathbf{d}_i\mathbf{d}'_i$ interactions, particularly those involving polar residues, are not captured well in our explicit model even when residues at these sites are allowed to relax using side-chain minimization. Performance improved significantly when interaction terms for these sites were replaced with values from experiments or from a machine-learning method. This was not true for $\mathbf{g}_i\mathbf{a}'_{i+1}$, $\mathbf{d}_i\mathbf{e}'_i$, $\mathbf{a}_i\mathbf{d}'_i$, $\mathbf{d}_i\mathbf{a}'_{i+1}$, or $\mathbf{g}_i\mathbf{e}'_{i+1}$ interactions, which were best represented explicitly. The fact that the physical model performs well for surface interactions but not for the core seems at odds with the observation that side-chain conformation recovery is much better in the \mathbf{a} and \mathbf{d} positions than in the more exposed \mathbf{g} and \mathbf{e} positions (Table 2.2). However, structure prediction at the resolution of χ -angle recovery and evaluation of stability are very different challenges. Physical energy functions that include van der Waals and electrostatic interactions are very sensitive to atomic position at close distances. Because the core is more crowded, it requires a finer resolution of possible conformations to give accurate estimates of these energies. Additionally, properly calculating the cost of burial, and balancing it with gained interactions, is difficult for core and partially buried positions. For surface positions, the total energy is more forgiving of variations in atomic position. Also, burial of surface residues compared to the unfolded state is less significant, making the desolvation/interaction tradeoffs less critical.

An accurate structure-based *denovo* model for predicting protein-protein interactions would have great utility. It would not require large amounts of experimental

data for determining parameters, and it would be interpretable in terms of basic physical principles. However, generic energy functions that have been developed for homology modeling and protein design do not perform well for the challenging problem of predicting coiled-coil interaction specificity. Despite the fact that they are effective for other purposes, DFIRE [197], RosettaDesign, and FOLD-X [62] perform poorly in our test. Although our best model so far contains implicit terms derived from a learning method, we have made considerable progress towards an accurate, physically interpretable, structure-based model. Model HP/S, with no FKS terms included, outperforms the methods mentioned above, and for $\Delta F \geq 2,000$ performs significantly better than a coiled-coil specific scoring function derived from experimental coupling energies. Model HP/S/C shows the best performance of any method. Thus, it is worth considering what models HP/S and HP/S/C suggest about the physical origins of coiled-coil interaction specificity.

Models HP/S and HP/S/C predict specific interactions important for bZIP coiled-coil stability. They capture the recognized importance of $\mathbf{g}_i\mathbf{e}'_{i+1}$ charge complementarity, and predict residue-residue interaction energies in quantitative agreement with measured coupling energies (Figure 2-6). Interestingly, our models predict that $\mathbf{g}_i\mathbf{a}'_{i+1}$ interactions are very important for establishing stability and interaction specificity. These terms contribute significantly to the predictive ability of the model, as evidenced by a strong reduction in performance upon selectively removing them (data not shown). Additionally, calculated magnitudes for several $\mathbf{g}_i\mathbf{a}'_{i+1}$ interactions are larger than those for $\mathbf{g}_i\mathbf{e}'_{i+1}$ interactions. Many of the same charged amino acids with long side chains that participate in strong $\mathbf{g}_i\mathbf{e}'_{i+1}$ interactions are also involved in strong $\mathbf{g}_i\mathbf{a}'_{i+1}$ ones, particularly Lys/Asp, Glu/Lys and Lys/Glu. Based on our success in recapitulating experimental $\mathbf{g}_i\mathbf{e}'_{i+1}$ coupling energies, we propose that such top-ranked $\mathbf{g}_i\mathbf{a}'_{i+1}$ pairs are excellent candidates for experimental measurement. A list of predicted strong $\mathbf{g}_i\mathbf{a}'_{i+1}$ interactions is given in the Supplementary Material.

Our results support the frequently made assumption that coiled-coil stability can be accurately expressed as a sum of residue-pair interactions. Although the explicit structures used to compute energies in our method have the potential to capture

context-dependent effects, we found no evidence that this contributed significantly to predictive ability. In fact, there was essentially no difference in overall performance between using contextually explicit interactions and the averages of interactions over all environments in which they were modeled. Further, the context-independent $\mathbf{a}_i\mathbf{a}'_j$ or $\mathbf{d}_i\mathbf{d}'_j$ terms from the FKS gave very good performance when added to model HP/S.

There is clearly room for improvement in these methods, as highlighted by mediocre performance distinguishing interactions from non-interactions in the YEAST, large Maf and CREB bZIP families (Figure 2-8). We investigated why some large Maf and CREB non-interacting pairs were predicted to be favorable whereas other pairs that interact experimentally were given high energies. For several cases examined, the problem was a mismatch between the energy function used for structure minimization and that used for final energy evaluation. In particular, the use of different electrostatic models (a distance-dependent dielectric for minimization and a Generalized Born treatment for evaluation) led to an imbalance between van der Waals and electrostatic interaction terms. Both of these terms are very sensitive to small changes in interatomic distances, so structures relaxed with one function sometimes give rise to unrealistically strong repulsions, or insufficiently attractive interactions, when evaluated with another. Structure minimization using only the van der Waals energy improved performance for the large Maf and CREB families quite significantly (data not shown) but did not improve overall results.

Although the performance of model HP/S/C for predicting interaction specificity based on sequence is already very good, improvements will likely come from general advances in the methods used for protein design and homology modeling, as well as from a better understanding of specific features of the coiled-coil motif. Table 2.2 shows that our structure-prediction performance is compromised by the use of an ideal backbone. Our studies of $\mathbf{a}_i\mathbf{a}'_j$ interactions involving Asn and/or Val (Table 2.3) show that computed energies are strongly dependent on the site in GCN4 at which these residues are modeled, illustrating how important it can be to capture the role of local backbone flexibility. Thus, better structural sampling techniques are needed. Electrostatic interactions remain difficult to describe accurately in a

pair-wise function suitable for use with standard search algorithms, and re-ranking strategies designed to address this suffer from discrepancies between approximate and more detailed energy surfaces. It is significant that our use of experimental data, introduced in the form of learned weights from the FKS model, was essential for achieving the highest accuracy predictions in this test. For the purpose of predicting coiled-coil interactions, a combined approach that uses structure and learned weights can probably be further optimized, perhaps within the FKS learning framework itself.

In the absence of a general method for predicting protein structure or protein-protein interactions, domain- or motif-specific models represent a promising way forward. Narrowing the focus of a study to a particular motif forces one to identify features that are important for that structure and to capture these as accurately as possible. Such an approach can lead to interesting insights, as well as to better performance. There is precedent for the value of this perspective in the extensive experimental and modeling work describing the DNA-binding specificity of zinc-finger proteins, and in the development of statistical models to recognize certain protein motifs from sequence data [13, 22, 49, 82, 190]. Further, domain-specific approaches lend themselves readily to experimental testing, which is key to making significant progress [134, 157, 170, 199]. The ability to predict the interaction specificity of various motifs and domains would be extremely useful for structure prediction, structural genomics applications, and annotation of the protein-protein “interactome”. At the same time, developing this capability promises to uncover important physical principles governing protein-protein interactions and teach us about the deficiencies of general-purpose models for capturing them.

2.4 Methods

2.4.1 Datasets and testing

Models were tested using experimental data for 58 peptides from Newman and Keating. [134] Duplicate sequences and one of a pair of peptides with identical sequence

at **a**, **d**, **e**, and **g** positions were removed. We assessed performance in ordering the stability of bZIP dimer pairs as a function of the difference, ΔF , between the raw fluorescence intensities measured for two compared interactions, as defined in Fong *et al.* [52] This yielded 33,186 experimental orderings for $\Delta F = 0$ out of a maximum of 95,847 comparisons possible (1,653 pairs of interactions for each probe). This is significant coverage, given that most of the possible comparisons are between non-interacting pairs.

For additional tests (Figure 2-8) all coiled-coil pairs were classified into 186 interactions and 849 non-interactions in a manner similar to Fong *et al.* [52]. For each pair of coiled coils, AB, there are two possible consensus Z-scores defined in ref [134] using a binomial test over all data sets – one arising from the experiment where A was the probe in solution and the other where B was the probe. An interaction was assigned if both consensus Z-scores were > 2.5 , whereas a non-interaction had Z-scores < 1.0 in at least 75% of the measurements in all experiments. Such definitions produce high-confidence sets of interactions and non-interactions, given the experimental data, and cover 60.5% of all possible pairings. All interaction/non-interaction predictions were grouped according to the sequence family of the peptide used as the solution probe in the microarray experiment. ROC curves were constructed for each family separately and the resulting curves were averaged in Figure 2-9.

2.4.2 Repacking and minimization

An initial side-chain placement study was carried out using known structures. Of the eight available bZIP dimer crystal structures, we used only three with resolutions of at least 2.0 Å (PDB IDs 2ZTA, 1GD2 and 1GU4). To obtain additional structures, all of the proteins listed in the SCOP database in the “leucine zipper domain” family with at least 2.0 Å resolution were manually examined (103 molecules). Those structures with a dimeric coiled-coil domain sufficiently separated from the rest of the molecule were included (PDB IDs 1GK6, 1UII, 1UIX, 1PI9, 1IC2). Different structures of the same molecule or mutant variants were ignored. In cases where other domains were present, the coiled-coil region was manually excised from the overall structure. The

repacking procedure described below for structure prediction was followed (except that all positions were repacked), and angles were classified as native-like if they were within $\pm 40^\circ$ of the crystal structure.

When predicting bZIP structures, an ideal, parallel, dimeric coiled-coil backbone with Crick parameters [33, 66] $R_o = 4.9\text{\AA}$, $\omega_o = 3.67^\circ$, $\phi = 21.2^\circ$ was used in all calculations except where specifically noted otherwise. The coiled-coil sequences, registers and heptad alignments were taken from Newman and Keating [134]. For a pair of sequences, a dimer of length equal to that of the shorter sequence was modeled. Only **a**, **d**, **e**, and **g** heptad positions were considered; **b**, **c**, and **f** were fixed as Ala. Given the fixed backbone and the sequences, side chains from the Richardson penultimate rotamer library [113] were placed using Dead End Elimination (DEE) followed by an A^* branch and bound search [42, 56, 58, 101, 105, 143]. The energy function used in conjunction with this consisted of the following terms: $\Delta G = \Delta G^{vdW} + \Delta G^{elec} + \Delta G^{des} + \Delta G^{dih}$. All terms were calculated in CHARMM using the param19 force field [25]. ΔG^{vdW} is the van der Waals energy modeled as a 6-12 Lennard-Jones potential using 90% radii. ΔG^{elec} is the water-screened electrostatic interaction energy calculated using a distance-dependent dielectric $\epsilon = \kappa r$. $\kappa = 4.0$ was used for side chain-to-template interactions and for non-polar to non-polar side-chain interactions, $\kappa = 16.0$ for polar to polar side-chain interactions and $\kappa = 8.0$ for all other side chain-to-side chain interactions. In coiled coils, this particular set of constants reproduces interactions calculated with the Poisson-Boltzmann equation well (data not shown). ΔG^{des} is the desolvation energy calculated with the EEF1 model of Lazaridis and Karplus [103] and ΔG^{dih} is the rotamer torsion energy. Energies calculated in the folded state included intra-residue interactions, interactions of side chains with the entire template as well as pairwise side chain-to-side chain interactions. The unfolded state was modeled as a set of non-interacting GGxGG penta-peptides with native backbone geometry, one per design site (in this case all **a**, **d**, **e**, and **g** positions), with the appropriate amino acid substituted at x. Energies calculated in the unfolded state, therefore, capture only intra-residue interactions as well as local side chain-to-backbone interactions. Predicted structures were indepen-

dent of the unfolded state energies. However, the resulting penta-peptides with their optimal rotamers were used for folding energy evaluations in model EX.

Once the optimal combination of rotamers was obtained for each bZIP dimer, the structures were allowed to relax through side-chain minimization using CHARMM. The energy function used for this included van der Waals energy with 100% radii, a distance dependent dielectric of $1/r$ as well as all bond, angle, dihedral and improper dihedral terms. In order to avoid biasing the resulting structures by the crude electrostatic function, only 10 steps of steepest descent followed by 10 steps of adopted basis Newton-Raphson minimization were used. This amount of minimization was found to be sufficient to relieve unrealistic van der Waals clashes without significantly changing the structure.

2.4.3 Evaluation of folded energy

The energy function for evaluating final structures was not constrained to be pairwise decomposable. We used the Generalized Born model (GB) with perfect radii computed using PEP [15], which is essentially as accurate as full treatment with the Poisson-Boltzmann (PB) equation [140]. A disadvantage of GB or PB models is that the reaction-field term cannot be properly expressed as a sum of contributions from different groups (such as the backbone or other side chains). This is problematic for models with an implicit unfolded state that accounts for the amino acid-to-backbone desolvation upon folding, where it is necessary to remove over-counted terms. Further, we suspected that the large reaction-field energies that result from GB, which uses a vacuum-like reference state with a low dielectric, did not provide accurate desolvation estimates in the absence of a realistic unfolded state structure. Therefore, we replaced the GB reaction field term with the approximate excluded volume-based solvation model from Effective Energy Function 1 (EEF1) by Lazaradis and Karplus [103] (calculated in CHARMM [25]). This energy function uses an aqueous small-molecule reference state and has the additional benefit of accounting for the hydrophobic effect [103]. It performed similarly to GB in model EX tests, where the unfolded state was explicitly considered and no implicit cancellation of local backbone effects was

necessary. The total electrostatic and desolvation energy consisted of Coulombic interaction energy in a uniform dielectric of 4, electrostatic screening from the GB model due to transfer to a medium of external dielectric of 80 and internal dielectric of 4, and atomic desolvation energy from the EEF1 model. The packing energy was modeled using the param19 van der Waals potential with 100% radii [25]. In all calculations only polar hydrogen atoms were considered explicitly and atomic parameters were derived from param19 [25]. In all models, changes in intra-side chain and intra-backbone interactions upon folding were ignored (except in cases where the latter is partly accounted for by helix propensities). Intra-backbone changes are difficult to model, and were assumed to largely cancel in comparisons of bZIP pairs with similar lengths. Changes in intra-side chain interactions were found to be rather small and strongly dependent on the choice of rotamer (or rotamers) in the unfolded state. Additionally, we observed that explicitly accounting for changes in intra-side chain interactions upon folding did not improve the performance of the models.

2.4.4 Helix propensities

Although only **a**, **d**, **e**, and **g** positions were considered for building dimer structures, all positions were included when calculating helix propensities. We used two models of helix propensity. In the first, we assumed that all heptad positions of a coiled coil have helix propensities equal to those in a “generic” helix and used the values of Munoz *et al.* [130]. These have been shown to correlate well with the average of several experimentally obtained scales [130]. In the second, we allowed for the possibility that single-residue contributions to helix stability depend on heptad position. To capture this, we used helix propensities for an isolated coiled-coil **f** position measured by O’Neil and DeGrado [139] and applied a correction factor derived as follows. First, we modeled each amino acid in an **f** position using an ideal coiled-coil backbone with the same sequence background as used by O’Neil *et al.* (structures were obtained using the same protocol as for bZIP coiled coils). Then, for any amino acid at heptad position **x** in a modeled bZIP coiled coil, we computed the difference between its total interaction with the coiled-coil backbone (both helices) and that of the same

amino acid when modeled in the **f** position. The helix propensity of this amino acid at this position was then corrected by the resulting difference. In this model the change in amino-acid self energy upon folding to the **f** position is captured by helix propensities, while the change in self energy due to going from position **f** to any position **x** is captured by side chain-to-backbone interactions.

Because helix propensities are a relative scale (usually referenced by Ala or Gly) we introduced an adjustable parameter hp_{ref} that shifts the entire scale by the same amount in order to be able to compare $G_{\text{self}}^{\text{folding}}$ for proteins of different lengths.

2.4.5 Full-helix folding energy function

Our implementation of the energy function underlying the AGADIR method by Serano and co-workers was based on the parameters for version 1s-2 given in Lacroix *et al.* [97]. Each sequence was scored as a difference between the energy of the folded helical state, where the entire sequence forms a helix, and the random-coil reference state, as described in reference [97].

2.4.6 Modeling backbone relaxation

We used two approaches to introduce backbone flexibility. In the first, we subjected structures resulting from the side-chain placement procedure to 2,000 steps of unconstrained continuous minimization in CHARMM. This resulted in a slight deformation of the backbone and improvement in mostly the van der Waals energy. For the second method, we considered a family of eight ideal coiled-coil backbones (each representing the best fit to one of the eight representative structures of bZIP coiled coils in the PDB – see Table 2.2.2). Each dimer was repacked on each of the backbones and the resulting structures were subjected to the same protocol as used for a single ideal backbone. The best energy according to model HP/S/C was used as the score for each dimer.

2.4.7 DFIRE

The executable and parameter files for DFIRE corresponding to the version used in reference [111] were obtained from the Zhou lab. The executable was run on all of the dimer structures predicted using our protocol. The value of the binding energy from the output was used to score each dimer. We tried using DFIRE on the set of structures obtained either with or without continuous side-chain minimization in CHARMM. The latter gave a slightly better performance and corresponds to the values in Figure 2-2(c).

2.4.8 RosettaDesign

A standalone version of RosettaDesign was obtained from the Baker lab. The method was used with the same GCN4-like ideal coiled-coil backbone as in our model. We tried models with either wild-type or Ala residues at the **b**, **c**, and **f** positions with nearly identical performance. The default rotamer library and parameters were used. Only one solution was requested. The total energy in the output file was used to score each dimer.

2.4.9 FOLD-X

Fold-X version 2.0.1 for Windows XP was downloaded from <http://foldx.embl.de/>. The executable was applied with default parameters to the structures predicted using our protocol either with or without side-chain minimization. The latter gave better performance and corresponds to the values in Figure 2-2(c). The “Stability” command was used and the total energy from the output file was used to score each dimer.

2.4.10 Evaluating contributions of $a_i - a'_i$ interactions to binding

Similar to Hendsch *et al.* [70] we modeled binding as rigid docking. All of the energies listed in Table 2.3 are differences between the dimeric and the monomeric states.

Using the crystal structure of GCN4, 2ZTA, as a model, either sites A16 and B16 (naturally occupied by Asn) or sites A9 and B9 (naturally occupied by Val) were considered. At each site either the native structure or models in which the targeted sites were replaced with N-N, V-N, N-V or V-V were examined. Models were generated by optimally placing mutant side chains onto the backbone in the presence of all native side chains, following the procedures outlined above. Of the two complex structures for N-V or V-N, the one producing the lower binding energy was chosen. All structures were evaluated without side-chain minimization and also after four different procedures in CHARMM: 1) 10 steps of minimization using only the van der Waals potential (with 100% param 19 radii), 2) minimization until convergence using only the van der Waals potential 3) 10 steps of minimization using the full EEF1 [103] energy function or 4) minimization until convergence using the full EEF1 energy function. The convergence criterion was that 10 steps of minimization changed the total energy by less than 0.1 kcal/mol. The different minimization procedures were found to give similar results and only (4) is shown in Table 2.3.

A variety of energy terms were computed to evaluate the final structures. The van der Waals energy was calculated in CHARMM using full param19 radii. Electrostatic energy was calculated as a sum of the Coulomb energy in a uniform dielectric ($\epsilon = 4$) and solvation energy from the GB model. Desolvation energy from the EEF1 model was calculated in CHARMM.

2.5 Acknowledgements

This chapter is based on a manuscript whose authors are Gevorg Grigoryan and Amy E. Keating, published in the *Journal of Molecular Biology*, vol. 355, pages 1125–1142, 2006. We thank the CSBi high-performance computing platform for computer time and support, the Baker, Serrano and Zhou labs for making programs available, J. Apgar for implementing the energy function underlying AGADIR, J. Fong and M. Singh for pre-processing of datasets, and M. Singh, R. Sauer, F. St-Pierre, and X. Stowell for comments on the manuscript. This work was supported by the NIH

(GM67681), and by NSF career award to A.K. (MCB: 0347203).

2.6 Supplementary Data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/doi:10.1016/j.jmb.2005.11.036>.

Chapter 3

Ultra-fast Evaluation of Protein Energies Directly from Sequence

The structure, function, stability and many other properties of a protein in a fixed environment are fully specified by its sequence, but in a manner that is difficult to discern. We present a general approach for rapidly mapping sequences directly to their energies on a pre-specified rigid backbone, an important sub-problem in computational protein design and in some methods for protein structure prediction. The cluster expansion (CE) method that we employ can, in principle, be extended to model any computable or measurable protein property directly as a function of sequence. Here we show how CE can be applied to the problem of computational protein design, and use it to derive excellent approximations of physical potentials. The approach provides several attractive advantages. First, following a one-time derivation of a CE expansion, the amount of time necessary to evaluate the energy of a sequence adopting a specified backbone conformation is reduced by a factor of $\sim 10^7$ compared to standard full-atom methods for the same task. Second, the agreement between two full-atom methods that we tested and their CE sequence-based expressions is very high (RMSD 1.1– 4.7 kcal/mol, $R^2 = 0.7 - 1.0$). Third, the functional form of the CE energy expression is such that individual terms of the expansion have clear physical interpretations. We derived expressions for the energies of three classic protein design targets – a coiled coil, a zinc finger and a WW domain

– as functions of sequence, and examined the most significant terms. Single-residue and residue-pair interactions are sufficient to accurately capture the energetics of the dimeric coiled coil, whereas higher-order contributions are important for the two more globular folds. For the task of designing novel zinc-finger sequences, a CE-derived energy function provides significantly better solutions than a standard design protocol, in comparable computation time. Given these advantages, CE is likely to find many uses in computational structural modeling.

3.1 Introduction

Protein structure prediction, homology modeling, fold recognition and design, including the prediction and design of macromolecular interactions, are among the most complex and essential problems in contemporary computational structural biology. Proteins are critical players in the cell and their function is dictated by their structure. Because the number of proteins with known sequence far exceeds the number with known structure, an ability to predict structure from sequence would be extremely valuable. On the other hand, designing proteins with specific structure and function is also important because of the usefulness of proteins as reagents and therapeutics [102].

At the heart of any computational approach to protein design or structure prediction lies the problem of determining the fitness (effective energy) of a particular protein in a given conformation or state. Depending on the method used, this effective energy may correspond to different physical quantities, e.g. stability, solubility, binding affinity, catalytic efficiency or a combination thereof. In protein design, the goal is to optimize this fitness in the large space of possible amino-acid sequences. In the fold-recognition approach to structure prediction (also called threading), the goal is to identify the most suitable structure for a particular sequence, given a library of known folds. In both cases the complexity of the problem imposes two sometimes conflicting requirements on the energy function used: physical accuracy and computational efficiency.

There are two major classes of fitness functions used in the fields of structure prediction and design. Lazaridis and Karplus [104] refer to these as statistical effective energy functions (SEEFs) and physical effective energy functions (PEEFs). SEEFs are derived from databases of proteins with known structures and describe the distribution of residues (or atoms) at different distances, solvent exposure, and sometimes more complicated measures, such as local atom density or relative orientation of secondary structure elements [151]. These terms are treated as effective potentials for calculating the energy of a protein in a given conformation. Most statistical energy functions include up to pair interactions [54, 158, 197]. However, it has been suggested that pairwise statistical energy functions may not be suitable for protein design or fold prediction [124, 178], so some SEEFs include higher order terms [27, 124, 149]. The advantages of SEEF methods lie in their computational efficiency, simplifying abstraction from details, and ability to implicitly capture effects such as desolvation, loss of entropy, and the hydrophobic effect, which are hard to account for explicitly. To gain these benefits, accuracy and physical interpretability are compromised.

Physical effective energy functions use atomic-level representations to capture underlying physical phenomena and approximate the free energy of the studied system. Some of the terms commonly included in PEEFs are van der Waals interactions, electrostatic interactions, hydrogen bond energies, dihedral angle torsion energies, atomic desolvation energies and solvent-accessible-surface-area or volume-dependent estimates of the hydrophobic effect [57, 93, 104, 144]. Some attempts have also been made to model side-chain entropy [31]. The advantage of PEEFs is that they have the potential to provide a more comprehensive understanding of the observed phenomena. The disadvantages are that much of the underlying physics is difficult to account for quantitatively, and when it is possible to do so, it is usually computationally expensive. An optimal energy function would have the simplicity and computational efficiency offered by SEEFs while retaining the theoretical rigor and physical interpretability of PEEFs.

A protein's behavior is a function of its sequence, given a defined environment. In particular, the energy required for a protein to fold to a given state or conformation (a

quantity of central importance for protein design and structure prediction problems) is a function of its sequence regardless of the complexity of the underlying physics that determines that energy. In this paper we present a general method by which the energy of a protein on a fixed backbone, given by an arbitrary energy function, can be accurately expressed as a simple function of its sequence. In principle, this method can be applied in conjunction with any energy function, the only limitation on the complexity being that energies for enough training sequences can be generated, at reasonable computational effort. We illustrate an application in which the calculated molecular mechanics energy of a protein, with a continuum treatment of solvation, can be mapped to a simple function of sequence that is extremely fast to evaluate and that maintains high accuracy. We find that the number of training sequences required to compute this mapping is significantly lower than would normally be adequate for sequence-space searches done in protein design. Furthermore, the resulting expansion retains, and in certain ways enhances, physical interpretability.

In the following sections, we first present an overview of theory of cluster expansion and detail its application to protein structural modeling. We point out how the expansion consists of terms that are conceptually familiar to biochemists. We then go on to apply the method to three protein systems: the α -helical coiled coil, the zinc finger and the WW domain. For each domain, we show that CE can derive useful yet highly simplified energy expressions. We conclude with a direct demonstration of the power of CE in protein design.

3.1.1 Theory

We seek to express the energy of a protein folding to a particular conformation as a function of its sequence. To attain this goal we employ the technique of cluster expansion (CE). CE is a method for representing a property (in this case, energy) that depends on discrete and topologically ordered degrees of freedom in a system [152]. The method finds its origin in alloy theory, where very expensive *ab initio* calculations are required to accurately capture material properties, and only computations on a small number of atomic arrangements with relatively small unit cells are possible

[39, 152]. The cluster expansion is essentially a parameterization of the energy in terms of discrete variables that give the occupancy of each lattice point in the crystal. When the occupation variable is a spin variable ($\sigma_i = +1$ or -1), the CE takes on the form of a generalized Ising model. This approach has proven itself highly accurate in predicting alloy phase diagrams [11, 175, 177], and in identifying novel low energy crystal structure [28, 176].

In its more general form, CE is an expansion of the energy in a set of linearly independent basis functions that span the relevant configuration space (e.g. all possible distributions of atoms A and B on a crystal lattice, or all possible amino-acid sequences on a protein backbone). In most forms, the basis set of the cluster expansion is mathematically complete by construction, and a full expansion will result in a perfect representation of the energy. Truncated expansions may have practical utility, however. The use of a truncated cluster expansion to model the energy is analogous to using any truncated expansion in basis functions (e.g. plane waves or spherical harmonics) to represent a complex unknown function. The goal in developing an effective CE is to identify a truncated expansion that, when fit to a training set of data, provides an accurate mapping between degrees of freedom and energy using a minimal number of parameters.

We have recently pioneered the use of CE for describing protein energetics [198]. To do so, we make a correspondence between an alloy lattice and a protein backbone and between alloy constituent elements and amino acids. Whereas alloy problems are typically solved for two or three possible species per site, the complete collection of natural amino acids requires twenty species per site. Such a dramatic increase in phase space requires some reformulation of the CE implementation typically used for alloys. The general idea is to define a set of basis functions that correspond to the energetic contributions of single amino acids at single sites, pairs of amino acids at pairs of sites, triples of amino acids at sets of three sites, and so on. If intuition holds, the lower-order terms in this expansion will be more important than the higher-order ones, and a truncated expansion will be sufficient to represent the energy. In practice, given a set of training sequences and their energies, the CE is derived by starting with

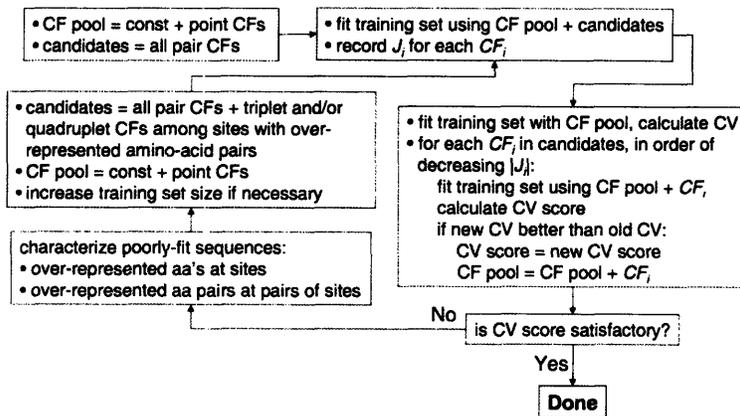


Figure 3-1: The procedure for fitting a cluster expansion. Cluster functions (CFs) capture the contribution of a particular set of amino acids at a set of sites. Point, pair, and triplet CFs contain the contributions of amino acids at single sites, pairs of sites, triplets of sites, etc. The energetic contribution of any cluster function CF_i is denoted by the variable J_i . CV score designates the cross-validation root mean square error (i.e. the average error with which the energy of each sequence is predicted when left out of the fit), and its behavior serves as a measure of parameter significance. The goal of the fitting procedure is to find an optimal pool of cluster functions with which to expand the energy. Point and constant CFs are always included and thus form an initial pool of CFs. In the next step, all pair CFs are considered as candidates. In order to assess the relative importance of candidate CFs, they are initially all added into the fit and their corresponding J_i 's are stored. The candidates are then visited one-by-one in the order of decreasing $|J_i|$ and considered for inclusion into the CF pool. Candidates are included if they reduce the CV score. If the final CV score upon trying all pair CFs is not satisfactory, the list of candidates is appended with higher order terms and the procedure is repeated. Details are provided in Materials and Methods.

lower order terms and successively considering higher order contributions until a fit of the expansion to the data gives adequate performance when tested under cross validation. This process is outlined in the flowchart in Figure 3-1 and elaborated in the Materials and Methods. A formal description of the theory of CE as we have applied it to protein energetics follows.

Given a discrete variable σ that can take on M different values ($\sigma = 0 \dots M - 1$), any function of it can be expanded using a basis set of M linearly independent functions $\Phi = \{\phi_0 \equiv 1, \phi_1, \dots, \phi_{M-1}\}$:

$$f(\sigma) = \sum_{a=0}^{M-1} J_a \phi_a(\sigma) \quad (3.1)$$

where J_a are constants. A similar statement can be made about any function $f(\vec{\sigma})$ of N discrete variables $\vec{\sigma} = \{\sigma^1 = 0 \dots M - 1, \dots, \sigma^N = 0 \dots M - 1\}$, because $\vec{\sigma}$ can be thought of as a discrete variable with M^N possible values. Thus, to expand $f(\vec{\sigma})$ exactly, a basis set with M^N functions is needed. Let vector $\vec{\sigma}$ represent an amino-acid sequence with element indices of the vector corresponding to sites on the protein under study. Thus, we consider N sites on a protein with M amino acids possible at each site. Further, let function $f(\vec{\sigma})$ be the optimal energy of sequence $\vec{\sigma}$ on a given backbone. According to the cluster expansion formalism [152], a particularly convenient basis set for expanding $f(\vec{\sigma})$ can be obtained by considering all the possible products between functions in the N point basis sets $\Phi^i = \{\phi_o(\sigma^i) \equiv 1, \phi_1(\sigma^i), \dots, \phi_{M-1}(\sigma^i)\}$ each completely describing the sequence space at site i . Thus, a basis set suitable for expanding $f(\vec{\sigma})$ is defined in the product space of the point functions:

$$\Phi' = \Phi^1 \otimes \Phi^2 \otimes \dots \otimes \Phi^N = \left\{ \begin{array}{l} [1], \\ [\Phi_1(\sigma^1)], \dots, [\Phi_{M-1}(\sigma^1)], [\Phi_1(\sigma^2)], \dots, [\Phi_{M-1}(\sigma^2)], \\ \dots \dots [\Phi_1(\sigma^N)], \dots, [\Phi_{M-1}(\sigma^N)], \\ [\Phi_1(\sigma^1) \Phi_1(\sigma^2)], \dots, [\Phi_1(\sigma^1) \Phi_{M-1}(\sigma^2)], \dots \dots, \\ \dots \dots \dots [\Phi_{M-1}(\sigma^1) \Phi_{M-1}(\sigma^2)], \dots \dots \dots, \\ \dots \dots \dots [\Phi_{M-1}(\sigma^{N-1}) \Phi_{M-1}(\sigma^N)], \\ [\Phi_1(\sigma^1) \Phi_1(\sigma^2) \Phi_1(\sigma^3)], \dots \dots \dots \\ \dots \dots \dots [\Phi_{M-1}(\sigma^{N-2}) \Phi_{M-1}(\sigma^{N-1}) \Phi_{M-1}(\sigma^N)], \\ \vdots \\ [\Phi_1(\sigma^1) \Phi_1(\sigma^2) \Phi_1(\sigma^3) \dots \Phi_1(\sigma^N)], \dots \dots \dots \\ \dots \dots \dots [\Phi_{M-1}(\sigma^1) \Phi_{M-1}(\sigma^2) \Phi_{M-1}(\sigma^3) \dots \Phi_{M-1}(\sigma^N)] \end{array} \right. \quad (3.2)$$

where in each row, the subscripts that index functions ϕ independently run through $1 \dots M - 1$ and the superscripts indexing protein sites take on all possible combinations of $1 \dots N$, without duplicates. Each basis function in this set (expressions in square brackets in equation 3.2) depends on the amino-acid identity at either no

sites (constant term), one site, two sites and so on. We call a set of specific sites a *cluster*. Each cluster has several basis functions, or cluster functions (CFs), associated with it. For instance, any point cluster i (a cluster consisting of site i) has $M - 1$ cluster functions associated with it (functions $\phi_1(\sigma^i), \dots, \phi_{M-1}(\sigma^i)$ but not $\phi_o(\sigma^i) \equiv 1$, which is attributed to the constant cluster). Therefore, there are a total of $N \cdot (M - 1)$ point cluster functions (the second row in equation 3.2) because there are N point clusters. Similarly, each pair cluster $\{i, j\}$ has $(M - 1)^2$ cluster functions associated with it ($\phi_o(\sigma^i) \phi_k(\sigma^j) \equiv \phi_k(\sigma^j)$ and $\phi_k(\sigma^i) \phi_o(\sigma^j) \equiv \phi_k(\sigma^i)$ are associated with point clusters i and j for $k > 0$ and with the constant cluster for $k = 0$). Because there are $N \cdot \frac{(N-1)}{2}$ pair clusters, the total number of pair cluster functions is $(M - 1)^2 N \cdot \frac{(N-1)}{2}$ (the third row in equation 3.2). For a size- k cluster, there are $\binom{N}{k} \cdot (M - 1)^k$ cluster functions. Therefore, the total number of cluster functions is $\sum_{k=0}^N \binom{N}{k} \cdot (M - 1)^k = M^N$ and there are as many linearly independent cluster functions in the basis set as there are possible values of the discrete variable $\vec{\sigma}$. Given the constructed basis set, we can exactly expand the energy of a sequence on the modeled backbone as:

$$f(\vec{\sigma}) = \sum_I \sum_A J_A^I \Psi_A^I \quad (3.3)$$

where I is a cluster of sites, Ψ_A^I is the A -th cluster function associated with cluster I , and the coefficients J_A^I are referred to as effective cluster interactions (ECI).

3.1.2 Interpretation of the Expansion

Because the point basis set at a single AA site $\Phi = \{\phi_0 \equiv 1, \phi_1, \dots, \phi_{M-1}\}$ can be any set of linearly independent functions, we choose for simplicity $\phi_a(\sigma) = \delta(a \cdot (\sigma - a))$. In other words $\phi_o(\sigma)$ is always one, and for $a > 0$, $\phi_a(\sigma)$ is always zero unless it is applied to the amino acid with index a . For any particular sequence $\vec{\sigma} = \{\sigma^1, \dots, \sigma^N\}$ the only CFs that remain in the expansion are of the form $\phi_{\sigma^i}(\sigma^i) \dots \phi_{\sigma^j}(\sigma^j)$ where

$\sigma^i \dots \sigma^j \neq 0$ (see equation 3.2) and thus $f(\vec{\sigma})$ is expressed as:

$$\begin{aligned} f(\vec{\sigma}) &= J_o + \sum_{\sigma_i \neq 0}^i J_{\sigma_i}^i \phi_{\sigma_i}(\sigma_i) + \sum_{\sigma_i, \sigma_j \neq 0}^{i \neq j} J_{\sigma_i \sigma_j}^{ij} \phi_{\sigma_i}(\sigma_i) \phi_{\sigma_j}(\sigma_j) + \dots \\ &= J_o + \sum_{\sigma_i \neq 0}^i J_{\sigma_i}^i + \sum_{\sigma_i, \sigma_j \neq 0}^{i \neq j} J_{\sigma_i \sigma_j}^{ij} \end{aligned} \quad (3.4)$$

The first term in the expansion is constant and J_o can be thought of as the energy of a reference sequence. Indeed, for a hypothetical sequence $\vec{\sigma} = \{\sigma^1 = 0, \sigma^2 = 0, \dots, \sigma^N = 0\}$, the only surviving part of the expansion is the constant term. The amino acid assigned index zero at each site defines the reference sequence; for simplicity, we will take this to be alanine. The ECI corresponding to higher order terms in the expansion then define additional contributions to the energy of a sequence relative to poly-alanine. For example, $J_{\sigma_i}^i$ corresponds to the point contribution of amino acid σ^i at site i relative to alanine at that site. This is the sequence context-invariant portion of an alanine-mutation energy. If there were no interactions among amino acids, point contributions and Ala-mutation energies would be equivalent. The context-dependent effects are captured by higher-order terms. For example, when interactions are present, the ECI corresponding to the terms $J_{\sigma_i \sigma_j}^{ij}$ capture the effective interaction between amino acids σ^i at site i and σ^j at site j relative to an Ala-Ala pair. Notice, however, that for amino-acid pairs Ala-X at sites i - j , where X corresponds to any amino acid, the value of $J_{\sigma_i \sigma_j}^{ij}$ is zero. The contribution of this interaction is captured in the point energy for amino acid X at site i . Therefore, the ECI corresponding to $J_{\sigma_i \sigma_j}^{ij}$ represents the pure effective interaction between the two amino acids, devoid of self contributions. This is conceptually identical to a double mutant coupling energy - a measure well known to biochemists [3, 94, 154]. Coupling energies measure the change in stability brought about by a double mutation, corrected by the change in stability due to each of the two single mutations. If the reference sequence in our cluster expansion is poly-alanine, pair ECI correspond to double alanine mutant coupling energies.

Even though the physics determining the conformational energy of a protein in solution is frequently modeled with only single-atom energies and pairwise atomic interactions, higher order contributions may arise if one integrates out some degrees

of freedom. For example, when modeling molecular solvation, if individual solvent molecules are replaced with a continuum high-dielectric medium, higher order interactions are necessary to accurately describe electrostatics as a function of conformational changes in the solute [73]. Similarly, integrating out side-chain degrees of freedom and expressing energy as a function of sequence can lead to higher order interactions between sequence variables, even though on the atomic level no more than pairwise interactions are present.

As shown in equations 3.3 and 3.4, the CE formalism allows for arbitrarily high order interactions (up to N-tuples) of residues. If all of the M^N terms have to be accounted for, such an expansion is not very useful. However, intuition dictates that for physical systems higher order interactions should be less important, and thus that ignoring them may be appropriate. If the expansion is truncated, the remaining coefficients J_A^I can be fit to minimize the error between the correct value of some desired fitness function and its CE approximation. Given a set of training sequences $\vec{\sigma}_1$ to $\vec{\sigma}_n$ with known energies $E(\vec{\sigma}_1)$ to $E(\vec{\sigma}_n)$, equation 3.3 defines a system of linear equations with J_A^I as the unknowns (each equation corresponding to one sequence).

$$\begin{bmatrix} E(\vec{\sigma}_1) \\ \dots \\ E(\vec{\sigma}_n) \end{bmatrix} = \begin{bmatrix} 1 & \dots & \Psi_A^I(\vec{\sigma}_1) \\ \dots & \dots & \dots \\ 1 & \dots & \Psi_A^I(\vec{\sigma}_n) \end{bmatrix} \cdot \begin{bmatrix} J_o \\ \dots \\ J_A^I \end{bmatrix} \quad (3.5)$$

If there are more sequences than cluster functions, the linear system in equation 3.5 becomes over-determined and it is possible to use least squares fitting to find the optimal values of J_A^I .

3.2 Results

In principle, the method of cluster expansion can be applied to any property of a protein sequence that can be computed or measured experimentally for a large set of training examples. In this work we expanded the energy of a sequence adopting a particular backbone conformation, which is a necessary component for protein design

and some methods for fold recognition. We computed this energy in two different ways. First, using a side-chain repacking scheme and a molecular mechanics potential (giving $E_{\text{repack}}^{\text{fold}}$) and second, subjecting every repacked structure to a short continuous side-chain relaxation procedure and then re-evaluating it with a more accurate energy function that included a non-pairwise decomposable electrostatics treatment (giving $E_{\text{min,GB}}^{\text{fold}}$) – see Materials and Methods.

In the Results we describe the application of CE to model the energetics of three different protein folds - the parallel dimeric coiled coil (an extended periodic structure), the zinc finger, and the WW domain (both aperiodic). These three structures, though small, are each of significant biological importance and have been the subject of previous protein design efforts using a variety of techniques [35, 68, 92, 150, 166].

3.2.1 Coiled Coil

The method of cluster expansion is particularly well suited for systems dominated by local interactions, because this limits the number of clusters that need to be included. CE also has an additional benefit in periodic systems, where modeling the energetics of a repeating unit cell can capture the behavior of the entire system. Both conditions are usually true in alloy theory, where the method is used extensively. Although proteins are rarely periodic, there are instances in which they are approximately so. An example of such a system is the α -helical coiled coil. The coiled coil is a common structural motif estimated to be present in approximately 5% of all proteins [189]. It consists of two to five right-handed helices that wrap around each other in a left-handed manner to form a super helix [32, 119]. Because of this super-coiling, the backbone geometry is repeated every seven residues - a unit that is referred to as a heptad, with its residues labeled **abcdefg**. Coiled coils can either be parallel (all N termini at one end), anti-parallel (N and C termini at opposite ends) or mixed (in higher order oligomers). In a parallel dimeric coiled coil (see figure 3-2), positions **a** and **d** are located in the core of the dimerization interface, whereas positions **e** and **g** are largely solvent exposed and can form salt bridges between strands of the coiled coil. Positions **b**, **c**, and **f** are solvent exposed on the side of the helix opposite to the

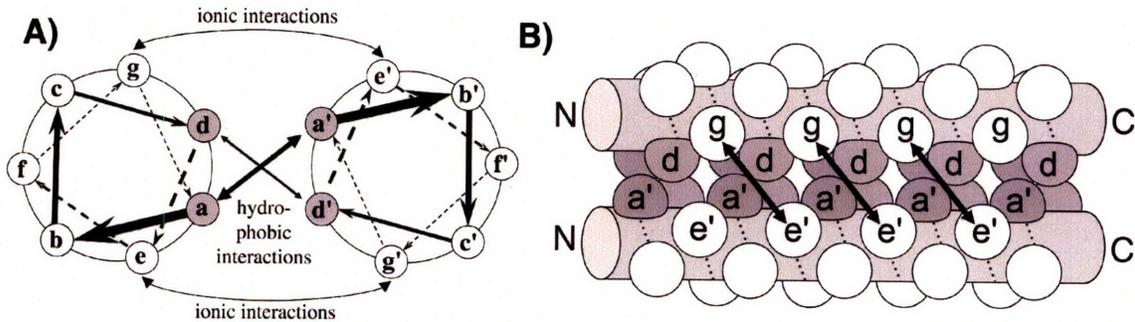


Figure 3-2: Schematic of a parallel dimeric coiled coil. **A)** Helical-wheel representation shows an end-on view of the structure. Opposing **a** and **d** residues interact in the core while opposing **e** and **g** residues frequently participate in electrostatic interactions. **B)** Cartoon representation of the coiled coil, viewed from the side. Residues are represented as spheres. An **e** position is better located for interaction with the **g** position of the previous heptad on the opposing strand than with the **g** position of the next heptad (bold arrows). This interaction is denoted $g - e'$ and coupling energies for it have been determined experimentally [94].

dimerization interface.

The parallel dimeric coiled coil is an extended structure, so it is reasonable to expect that only local clusters will contribute significantly to the energy expansion. Additionally, it is a periodic structure, so by accurately modeling the interactions of one structural subunit (unit cell), we can describe a coiled coil of arbitrary length. The unit cell must contain within it all interactions likely to be important. We postulated that interactions between amino acids more than one heptad apart are not significant. Thus, we modeled the unit cell as the central two-heptad section of a six-heptad dimeric coiled coil, where the flanking sequences were copies of the unit cell sequence (see figure 3.2.1). Because it is generally assumed that positions **b**, **c**, and **f** play only a minor role in determining the dimerization properties of coiled coils, we set these to alanine in our model. Positions **a**, **d**, **e**, and **g** were allowed to be one of 19 amino acids (all natural ones except proline).

We expressed the folding energy of a parallel dimeric coiled coil (i.e. the difference between the dimer state and the unfolded monomers state) as a function of its sequence. In order to be tractable, the expansion in equation 3.3 must be truncated. Consistent with our unit cell approximation, we included only clusters involving sites

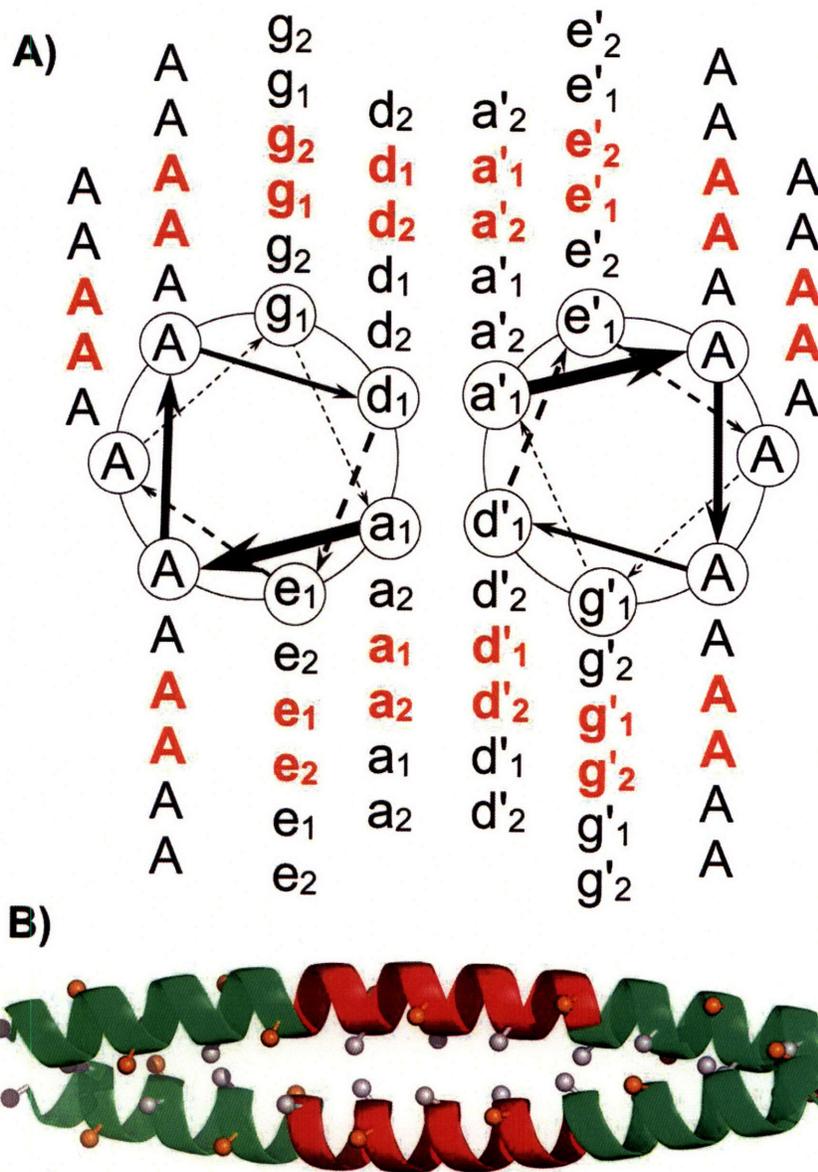


Figure 3-3: The unit cell used for modeling coiled-coil interactions. The entire structure consists of three copies of the sequence of the central unit cell, which is $a_1AA d_1e_1Ag_1a_2AA d_2e_2Ag_2$ on the one strand and $a'_1AA d'_1e'_1Ag'_1a'_2AA d'_2e'_2Ag'_2$ on the other, marked in red (A = alanine). Only positions **a**, **d**, **e**, and **g** were allowed to vary. The energy of the central unit cell is calculated as the sum of its internal interactions and half of its interactions with the bounding structure. **A)** Helical wheel diagram corresponding to the entire structure modeled, with sites in the central unit cell colored red. **B)** Ribbon diagram representation of the modeled system viewed as in 3-2B with the central unit cell colored red. Grey and orange balls represent locations of side-chain C_β atoms of **a/d** and **e/g** sites respectively.

no more than seven residues apart in the expansion. Further, as a starting point, we included only up to pair clusters, resulting in a total of 137 clusters. Taking into account coiled-coil symmetry (since ECI for symmetry equivalent clusters are identical [39, 152]), this was reduced to 1 constant, 4 point and 36 pair clusters with unique ECI. To find appropriate values for coefficients J_A^I , we considered $\sim 30,000$ randomly generated sequences (i.e. approximately 2.5 times as many sequences as J_A^I parameters being fit) and computationally predicted their structures under the assumption of a constant ideal backbone and discretized side-chain conformations [60]. This involved searching a conformational space of 10^{53} structures for an average sequence. Given optimized structures, we calculated $E_{\text{repack}}^{\text{fold}}$ for each and used these as a training set to find optimal values for J_A^I (see Materials and Methods and Figure 3-1). Figure 3-4A shows the progress of the fit accuracy, measured by cross-validation, as a function of the number and type of clusters functions (CFs) included in the expansion. The largest drop in error, per cluster function, is due to point CFs. This is intuitive and consistent with the fact there are strong amino-acid preferences at different coiled-coil heptad positions [115, 183]. A few important pair cluster functions further reduce the error significantly, and many less-important pairs drive the error down slowly.

Figure 3-4B shows the performance of the resulting CE on predicting coiled-coil energies for a test set of $\sim 4,000$ sequences not present in the training set. When deriving the expansion, we considered only the energy of a two-heptad unit cell, so training-set sequences were periodic with a two-heptad sequence repeated three times (see figure 3.2.1 and Materials and Methods). The test set, however, contained non-periodic six-heptad sequences and allowed us to evaluate not only the accuracy of the cluster expansion, but also the validity of the unit-cell approximation. The overall root mean square deviation (RMSD) is 1.96 kcal/mol, whereas that for more relevant sequences (those with calculated energies below -5 kcal/mol) is 1.08 kcal/mol. This is a very small error and is in fact comparable to or better than the accuracy of the underlying energy function. Thus, for a six-heptad coiled coil, the CE formalism reduces a sequence-structure space of 10^{115} possibilities to a search of 10^{61} sequences

with minimal cost in accuracy. The reduction of search space grows exponentially with the length of the coiled coil modeled.

Given the accuracy and simplicity of the CE functional form, the task of evaluating the energy of a sequence is reduced to several interaction table lookups. However, the CE formalism is also convenient because the functional form implies that individual ECI have clear physical interpretations. Specifically, pair ECI correspond to double mutant coupling energies. Figure 3-5A shows the agreement between experimentally measured $\mathbf{g} - \mathbf{e}' +$ coupling energies [40, 94] (the prime designates the opposite strand and “+” indicates the next heptad) and the corresponding pair ECI from the above cluster expansion. The excellent agreement illustrates the physical interpretability of the cluster expansion.

One of the strengths of the CE approach is that, in principle, any energy function can be expanded as a function of sequence. In a previous study we found that more reasonable coiled-coil energies were obtained by allowing the structures resulting from discrete side-chain repacking to relax via several steps of continuous side-chain minimization [60]. In addition, we derived a specific physics-based energy model (HP/S) that performed well in predicting coiled-coil dimerization preferences [60]. Unlike the original energy function used above, HP/S is not pairwise decomposable at the atomic level, due to its more accurate treatment of electrostatics. We fit a cluster expansion for the HP/S energy using the same training set sequences as before. Figure 3-4C shows the progress of the fit as a function of the number and type of included cluster functions. Again, constant, point and pair clusters are sufficient for reasonable accuracy. Figure 3-4D shows the performance of the resulting cluster expansion on a set of $\sim 4,000$ test sequences not included in the training set. The error for relevant sequences (those with energies below 0 kcal/mol) is 1.96 kcal/mol. Note that these energies are not strictly on an experimental scale. Our previous work has determined that stable coiled coils of 5-6 heptads have energies varying over 15 kcal/mol using this energy function [60] and random sequences span a range of over 40 kcal/mol; this is surely larger than the range of experimental free energies of folding. Figure 3-5B shows the agreement of experimental $\mathbf{g} - \mathbf{e}' +$ coupling energies with the

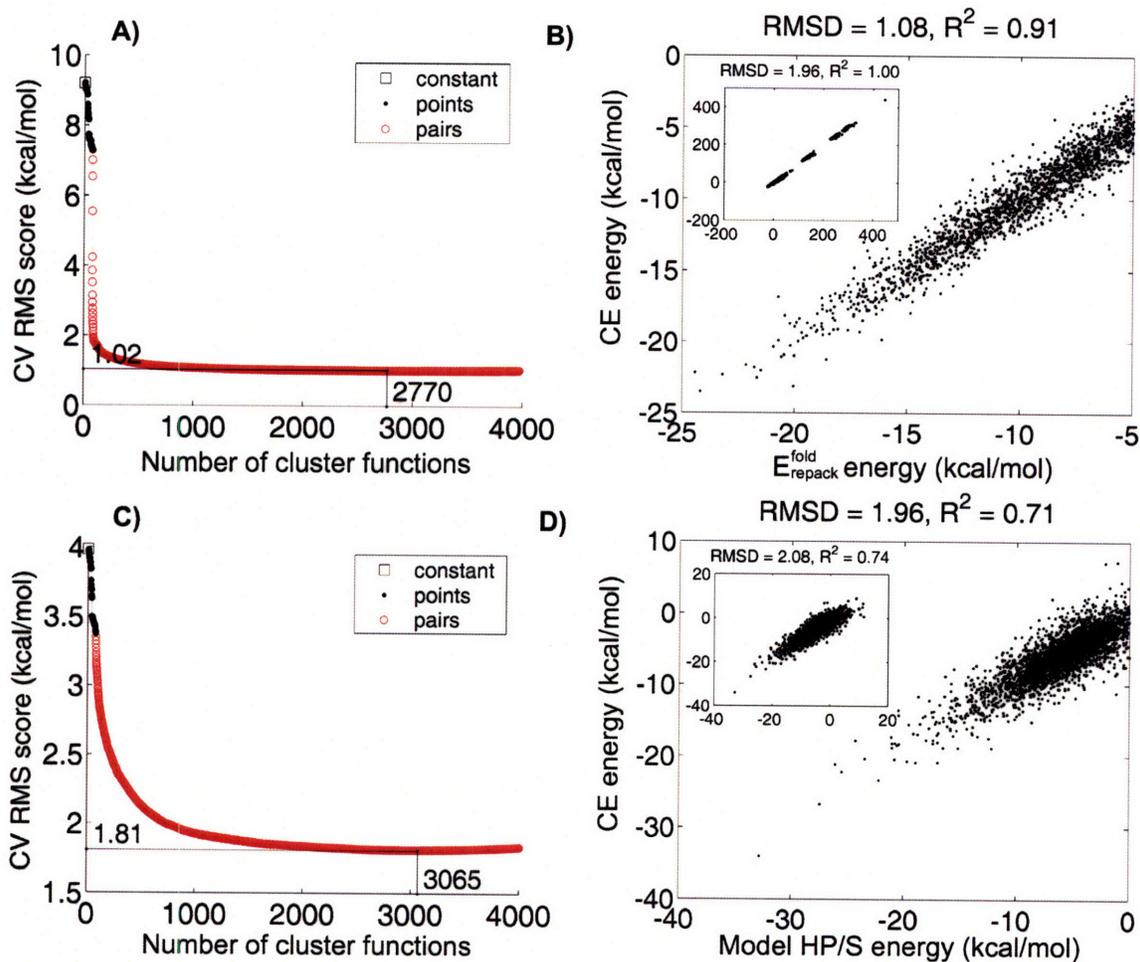


Figure 3-4: Cluster expansion of coiled-coil energies. Panels A) and B) refer to the cluster expansion of $E_{\text{repack}}^{\text{fold}}$; in panels C) and D) energies from model HP/S [60] were used. Panels A) and C) represent the evolution of the CV score (the progress of the fit) as the number of CFs is increased, with the type of CF added at each point (i.e. constant, point, pair) indicated by color. The ordering of the points is described in Materials and Methods. The set of cluster functions and ECI in the final expansion is taken from the point with the minimal CV score, which is indicated on the graphs. Panels B) and D) show the performance of the respective cluster expansions on predicting energies of 4,000 random sequences not included in the training sets. Insets show the entire range of energies, whereas only sequences with reasonably low energies are shown in the main plots.

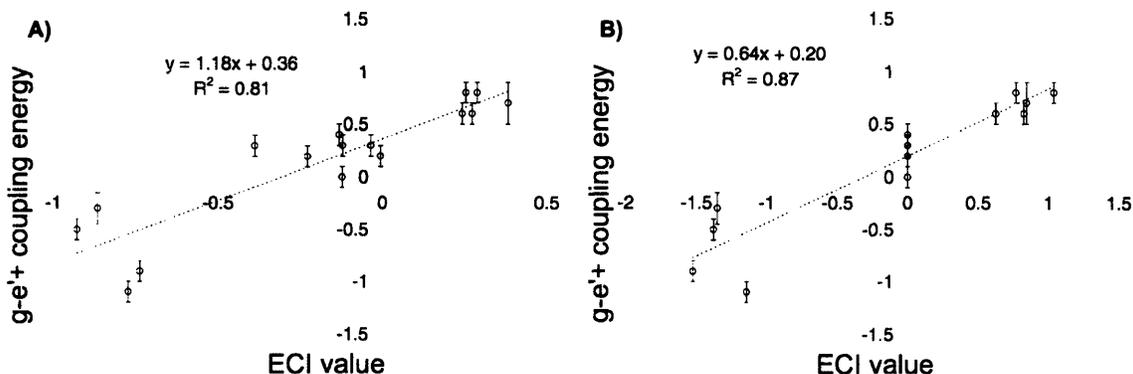


Figure 3-5: Agreement between experimentally measured double-alanine coupling energies for residues E, Q, R and K at $g - e' +$ [94] and corresponding pair ECI from the cluster expansion (in kcal/mol). **A)** Energies from repacking calculations, $E_{\text{repack}}^{\text{fold}}$, were used to fit the CE. **B)** $E_{\text{min,GB}}^{\text{fold}}$ energies were used to fit CE.

corresponding pair ECI obtained from this cluster expansion. This comparison with experimental values is more meaningful, due to cancellations in error in the double mutant cycle that put the calculations on a similar scale.

3.2.2 Zinc Finger

A cluster expansion including only up to residue-pair interactions works well for the coiled coil, an extended fold where only local interactions are likely to be important. To test whether this is a unique property of the coiled coil and whether higher order interactions are important in more globular folds, we examined the zinc-finger motif. Zinc finger domains are found in a diverse set of proteins that require coordination of one or more zinc ions to stabilize their structure [98]. Cys₂His₂ zinc fingers coordinate a zinc ion with two cysteine and two histidine residues and are found in many DNA-binding proteins. Among these, the murine zinc finger Zif268 has been extensively studied [142]. To derive a CE describing the Zif268 fold, we defined the backbone using coordinates from the PDB entry 1ZAA, residues 33-60. The amino acids allowed at each site were based on the classic design by Dahiyat *et al.* [35] and were such that 1 core site was chosen from 7 aliphatic amino acids, 18 surface sites varied among 10 amino acids and 7 interface sites were selected from 16 amino acids (a sequence space of $\sim 10^{27}$). This restriction gives design sequences with better physical properties

while retaining a large and diverse protein design search space. Side-chain repacking was used to calculate folding energies $E_{\text{repack}}^{\text{fold}}$ for $\sim 60,000$ random training sequences and a cluster expansion was derived. The progress of the fit is shown in Figure 3-6A, where the order in which triplet and pair cluster functions were added is defined in Figure 3-1 (see Materials and Methods). In this case, triplet cluster functions are necessary to attain good accuracy, and it is not strictly true that pair terms contribute more significantly than triplet terms. Additionally, the contribution of point terms is relatively larger than for the coiled coil, indicating that an amino acid's contribution to the overall energy is affected significantly by the 3-dimensional template of the molecule. Figure 3-6B shows the accuracy of the derived cluster expansion when tested on a set of $\sim 4,000$ random sequences not included in the training set. The RMSD of 15.3 kcal/mol over the entire range of energies is quite high, but this is due to the large spread in energies (over 1,000 kcal/mol) caused by many of the sequences producing van der Waals clashes. As a percentage of the range, the error is quite low ($< 1.5\%$), and for the more realistic zinc-finger sequences (those with negative energies) the error is only ~ 2.5 kcal/mol. In this case, CE reduces a sequence-structure space of $\sim 10^{60}$ to $\sim 10^{27}$ sequences.

To expand a more physically meaningful energy, we used $\sim 30,000$ structures to calculate $E_{\text{min,GB}}^{\text{fold}}$ for each and used these for training. The progress of the resulting cluster expansion fit is shown in figure 3-6C. Once again, triplet terms are important for attaining good accuracy. Most of the triplet cluster functions arise from the two triplet clusters shown in figure 3-7. These are structurally compact, with CFs of significant magnitude mostly corresponding to large amino acids (such as Y, F, and W). Such clusters represent close-range interactions of bulky residues. Figure 3-6D shows the performance of the CE on a test set of $\sim 4,000$ sequences not included in the training set. Though the agreement is still very good ($R^2 = 0.85$), the error is larger than in other cases (4.61 kcal/mol for sequences with energies ranging between 0 and -60 kcal/mol) indicating that the more complicated geometry of the domain may make the energy a more complex function of sequence.

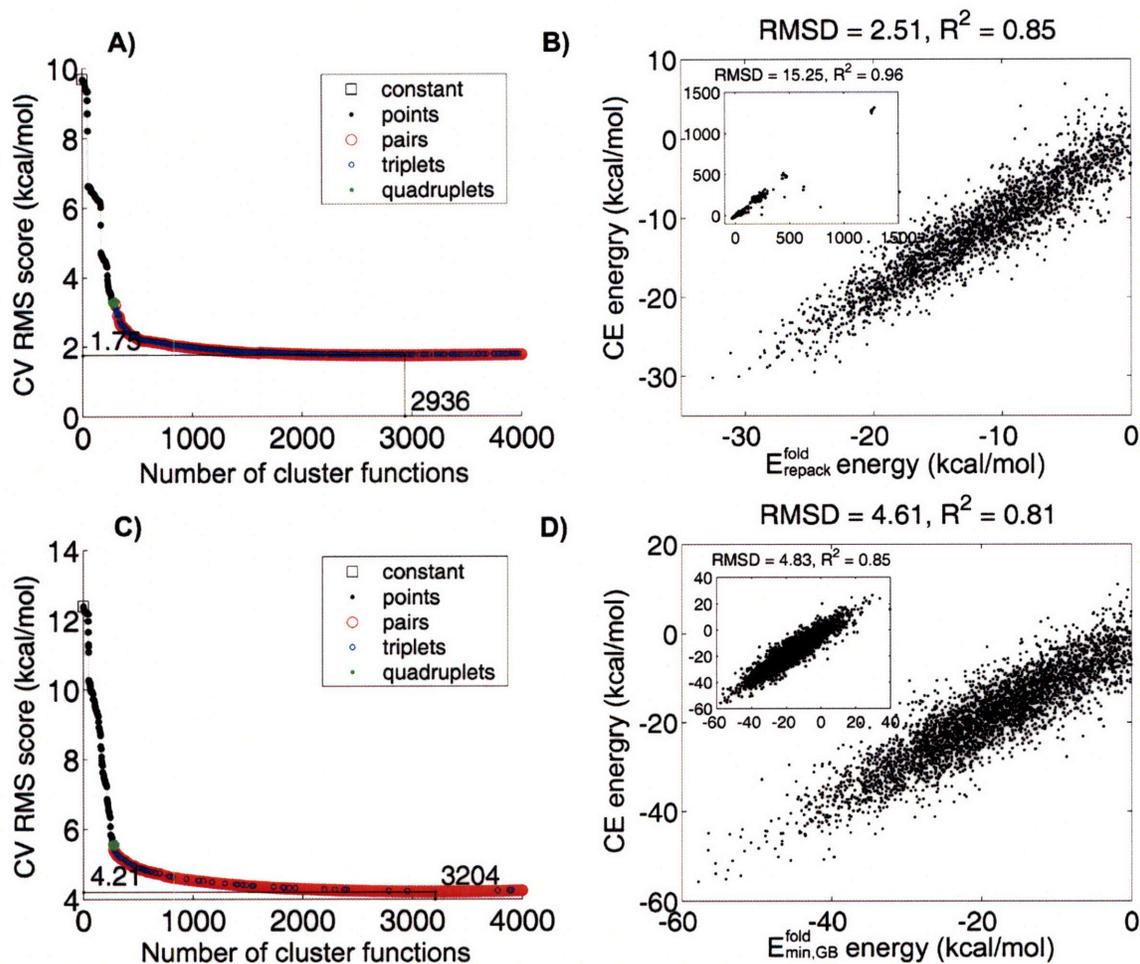


Figure 3-6: Cluster expansion of zinc-finger (ZF) energies. Panels A) and B) refer to the cluster expansion of $E_{\text{repack}}^{\text{fold}}$; in panels C) and D) $E_{\text{min,GB}}^{\text{fold}}$ is expanded. Panels A) and C) represent the evolution of the CV score (the progress of the fit) as the number of CFs is increased, with the type of CF added at each point (i.e. constant, point, pair) indicated by color. The ordering of the points is described in Materials and Methods. The set of cluster functions and ECI in the final expansion is taken from the point with the minimal CV score, which is indicated on the graphs. Panels B) and D) show the performance of the respective cluster expansions on predicting energies of 4,000 random sequences not included in the training sets. Insets show the entire range of energies, whereas only sequences with reasonably low energies are shown in the main plots.

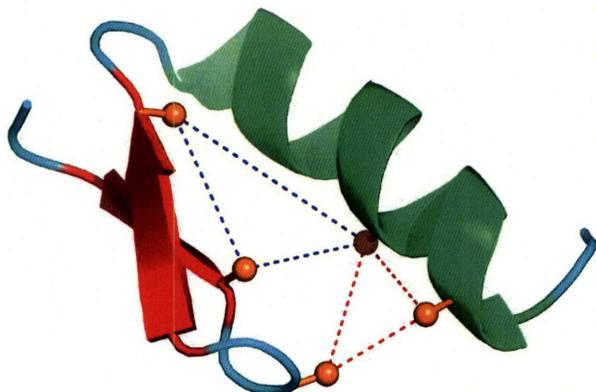


Figure 3-7: Important triplet clusters for the expansion of zinc-finger energies. Orange balls represent the location of the C_{β} atoms of side chains. Two clusters are shown, one in red and one in blue.

3.2.3 WW Domain

The WW domain is a protein-protein interaction motif composed of 35-40 residues. It forms the smallest known independently stable triple-stranded antiparallel β -sheet. WW domains bind proline-rich or proline-containing ligands [77]. A defining feature of this motif, from which its name is derived, is the presence of two tryptophans spaced 20-22 residues apart. Under the assumption that the statistical information encoded in multiple sequence alignments of WW domains reflects evolutionary constraints, Ranganathan and co-workers have used these statistics to engineer artificial WW domains with specific binding properties [150, 164]. Protein design methods using energy functions similar to those we employ here have also been applied to this domain [92].

We derived a cluster expansion for the WW domain that captures relationships between sites that are important for folding energetics. We used the structure of human PIN1 WW domain (PDB ID 1PIN) to define backbone coordinates and chose an alphabet of amino acids at each site using a multiple sequence alignment of WW domains from the SMART database (accession number SM00456). The choices at each position covered at least 90% of all naturally occurring residues. Thus the search space is very diverse while at the same time it excludes sequences that are grossly incompatible with the WW domain fold and not worth searching. The resulting

problem had an average of 7.5 amino acids per position and a total of $\sim 1.1 \times 10^{27}$ possible sequences. We explicitly computed structures for $\sim 42,700$ random sequences and estimated their folding energies.

Figure 3-8A shows the progress of expanding $E_{\text{repack}}^{\text{fold}}$ for the WW domain as a function of the number and type of cluster functions in the expansion. Similar to the Zn finger, we found that higher order terms (11 triplet clusters and 1 quadruplet cluster) were necessary for good agreement. Figure 3-8B shows the performance on a set of $\sim 4,000$ test sequences not included in the training set. The error of only 1.76 kcal/mol over a range of ~ 40 kcal/mol is impressively low and the correlation is good. Here CE reduces a sequence-structure space of 2.6×10^{65} to 1.1×10^{27} sequences.

Figure 3-8C shows the progress of expanding $E_{\text{min,GB}}^{\text{fold}}$ for the WW domain. Once again, higher order interactions contribute significantly to the expansion. However, the relative contribution of point terms as compared to the case where no minimization was done (figure 3-8A) is much larger. This is likely due to the fact that many high energy side chain-to-side chain interactions were relieved upon minimization. Several triplet clusters contribute many cluster functions of considerable magnitude. However, unlike for the zinc finger, for the WW domain there are two types of triplet clusters. One consists of structurally compact sites, and CFs arising from these clusters are mostly positive and correspond to large amino acids (see Figure 3-9A for an example). In the other, sites are more structurally dispersed and combinations of residues producing significant CFs consist mostly of charged and polar amino acids (see Figure 3-9B). These two types of clusters roughly correspond to the two main classes of interactions we model - van der Waals (short-range) and electrostatics (which can be long-range). Additionally, there is one quadruplet cluster that seems to be important for overall accuracy - it is shown in Figure 3-9C. The set of amino acids at this cluster that give rise to large CFs is diverse and it does not have a clear structural or energetic interpretation. The error of the fit, 4.7 kcal/mol (Figure 3-8D), is higher than before but, considering the energy range of over 300 kcal/mol, this is sufficiently accurate to be very useful.

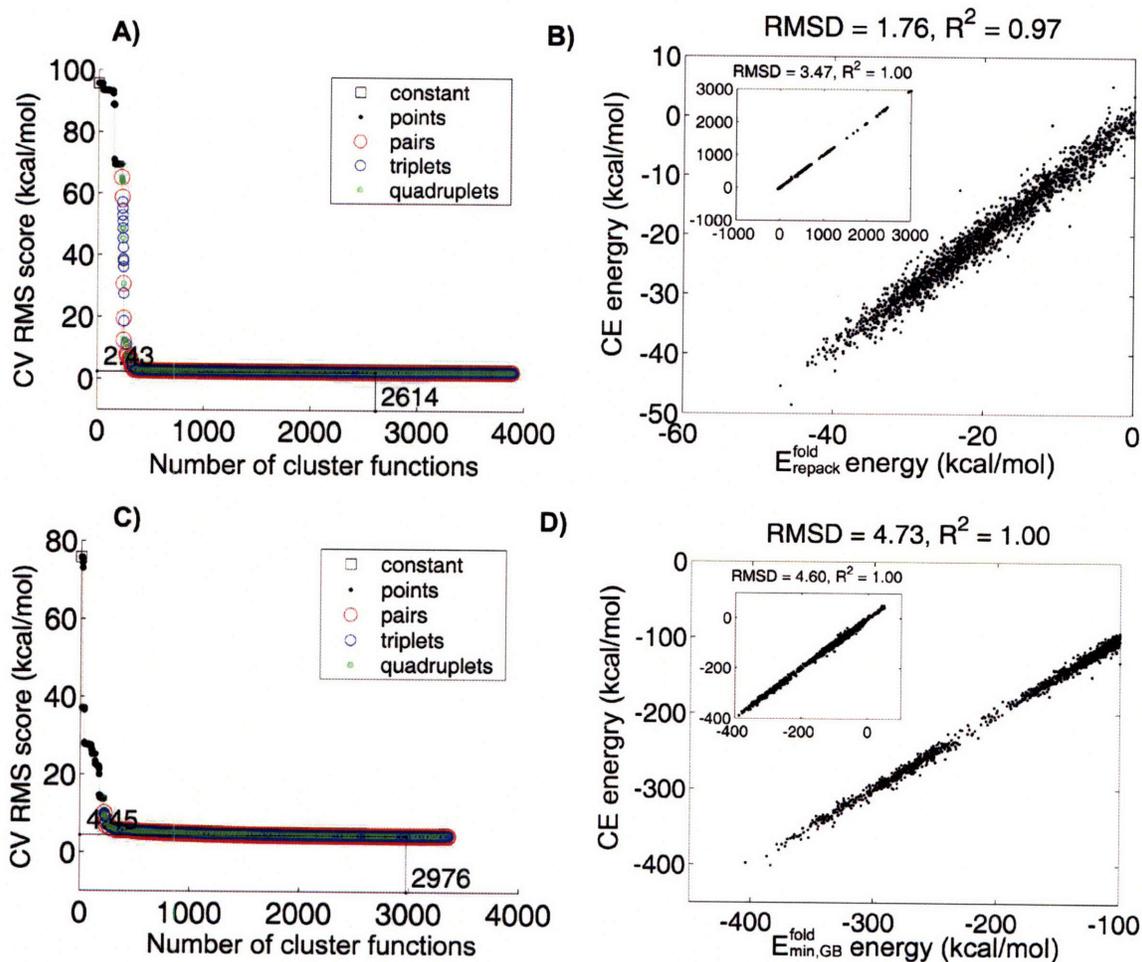


Figure 3-8: Cluster expansion of WW domain energies. Panels A) and B) refer to the cluster expansion of $E_{\text{repack}}^{\text{fold}}$; in panels C) and D) $E_{\text{min,GB}}^{\text{fold}}$ is expanded. Panels A) and C) represent the evolution of the CV score (the progress of the fit) as the number of CFs is increased, with the type of CF added at each point (i.e. constant, point, pair) indicated by color. The ordering of the points is described in Materials and Methods. The set of cluster functions and ECI in the final expansion is taken from the point with the minimal CV score, which is indicated on the graphs. Panels B) and D) show the performance of the respective cluster expansions on predicting energies of 4,000 random sequences not included in the training sets. Insets show the entire range of energies, whereas only sequences with reasonably low energies are shown in the main plots.



Figure 3-9: Important higher-order clusters for the expansion of WW-domain energies. Orange balls represent the location of the C_{β} atoms of side chains. **A)** A structurally compact cluster corresponding to short-range interactions. **B)** A more disperse cluster arising from long-range electrostatic interactions. **C)** Quadruplet cluster with many contributing cluster functions corresponding to a wide range of amino-acid types.

3.2.4 A Design Application and Speedup Analysis

Because the sequence-dependent energy function provided by CE is enormously simplified relative to the full physical model, it takes significantly less time to evaluate the energy of one sequence. This parameter is of critical importance in protein design, where very large sequence spaces need to be searched. We compared the amount of time it takes to evaluate the energy of one sequence either with the direct structural method or using CE (see Materials and Methods; all computations were run on 2.4 GHz CPU machines with 2GB of memory, although memory was not a limiting factor). For the coiled-coil system considered above (a total of 48 variable sites) it took ~ 360 seconds on average to repack, minimize and re-evaluate one sequence. Using CE, it took $\sim 4 \times 10^{-5}$ seconds to evaluate an approximation of that same energy, a speedup of 9×10^6 . For the zinc finger (a total of 26 variable sites) on average it took ~ 70 seconds per sequence for the structural method and $\sim 7 \times 10^{-6}$ seconds with CE - a speedup of 10^7 . And finally, for the WW domain (34 variable sites) the corresponding times were ~ 70 and $\sim 6 \times 10^{-6}$ seconds - a speedup of 1.2×10^7 .

The large speed advantage of CE comes at the cost of an error in energy. In addition, deriving a cluster expansion relies on evaluating a set of training sequences with the slower, atomic-level methods and carrying out the fitting procedure. To

assess the overall advantage that CE brings to protein design, we used the zinc finger system as an example and carried out two design procedures. One was a sequence search driven by the “exact” energies obtained by repacking, minimizing and evaluating every sequence (direct design). The other consisted of using the same evaluation procedure to calculate energies for a training set of random sequences, deriving a cluster expansion and performing a sequence search guided by CE energies (CE design). In an approximation of a head-to-head competition, the two methods were allowed the same amount of wall-clock time (~ 2 days), and up to 20 processors, as follows. Direct design was allowed to sample a total of 60,000 sequences by performing 20 independent Monte Carlo runs each with 3,000 steps (with the temperature linearly falling from 1000K to 298K and the acceptance of each step governed by the Metropolis criterion [123]) and took 2 days on 20 processors. Fitting the cluster expansion required explicit modeling of $\sim 30,000$ sequences, which took 1 day on 20 processors. In addition, the fitting procedure (run in serial) took approximately a day of mostly human operational time (see Materials and Methods for details of the fitting procedure). Upon completing the fit, CE design was given 12 minutes on 1 processor to run 100 Metropolis Monte Carlo searches guided by CE energy, each with 10^6 steps and the same temperature range as above. The best sequences from each of these 100 runs were then explicitly repacked, minimized and evaluated using the original, direct energy function. Figure 3-10 compares energy histograms corresponding to these sequences (with their energies evaluated with the direct energy function) and the 100 best sequences from direct design. Clearly, due to its ability to cover a considerably larger sequence space, CE discovers significantly better sequences.

3.3 Discussion

We successfully adapted the method of cluster expansion [152], often used in alloy theory, to express the energies of proteins in several backbone conformations directly as functions of their sequences [198]. The resulting energy functions are a tremendous simplification relative to the underlying physical model, and as such offer an

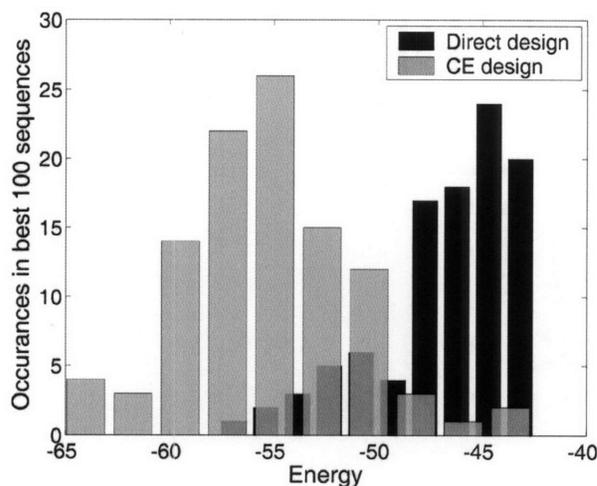


Figure 3-10: Distribution of the energies of the top 100 sequences from direct design and CE design. The best solutions from CE design were modeled and repacked using the direct method for comparison purposes. Thus, the reported energy is that computed using the direct method for both cases. The best sequence found with CE design is significantly better than the best one from direct design. Also, the ensemble of best sequences found with CE is significantly more stable than that from direct design. This indicates that its greater speed allows CE design to reach and sample a lower energy sequence space.

enormous computational speedup compared to explicit atomic-level calculations. Despite their simplicity, these functions produce energies in close agreement with those obtained through explicit calculations. Additionally, the functional form associated with the cluster expansion formalism ensures that the individual terms of the final expression are easily interpreted physically. The fact that this approach can be used in conjunction with any theoretical or experimental energy function, regardless of its complexity, makes this a very powerful general method that is likely to prove useful for many computational structural approaches.

We successfully applied CE to three model systems and illustrated its potential for computational protein design. Figure 3.4 shows the results for the parallel dimeric coiled coil. We found that including only up to pair interactions in the cluster expansion was sufficient for excellent agreement, giving an error of just 1 – 2 kcal/mol. Interestingly, several methods of scoring coiled-coil dimerization have assumed that pair interactions in sequence space are sufficient to describe the fold [52, 115, 120]. Additionally, many experimental studies of coiled-coil interactions have made the

assumption that a pair of amino acids at a pair of sites has a roughly constant contribution, regardless of the remaining sequence environment [40, 179]. The finding that a cluster expansion with only up to pair terms is sufficient to accurately describe the energy of the entire structure supports these assumptions.

One of the strengths of the cluster expansion approach is the transparency of the functional form and the consequent interpretability of the fitting coefficients. Supporting this, we demonstrate good agreement between experimentally measured coiled-coil $\mathbf{g} - \mathbf{e}' +$ coupling energies [94] and the corresponding pair ECI from the cluster expansion (see Figure 3.5). These measures are not exactly equivalent, as coupling energies are measured in a specific context, whereas ECI capture an effective interaction between two residues that is independent of surrounding sequence. Practically, however, much of the context-dependence probably cancels in corrections for single-site effects.

There is a less direct correspondence between point ECI and Ala-mutation energies, which are very sensitive to environment. Additionally, self contributions to folding are more sensitive than coupling energies to the nature of the unfolded state, and modeling the unfolded state is a challenge. However, we do find qualitative agreement between point ECI and experimentally observed positional amino-acid preferences. Leucine has the most favorable point ECI at \mathbf{d} positions according to the cluster expansion derived from minimized structures. Analysis of parallel dimeric coiled-coil sequences shows that Leu is by far the most common residue at position \mathbf{d} [115, 120, 183]. Moitra *et al.* have further shown that in at least two slightly different sequence backgrounds Leu is the most stabilizing aliphatic amino acid at the \mathbf{d} position [128]. Based on these results, it is reasonable to propose that the observed preference for Leu at \mathbf{d} positions in parallel dimeric coiled coils comes from a favorable single-body energetic contribution, as captured in the cluster expansion. Sequence analysis also suggests that Leu is the most common amino acid at the \mathbf{d} position [115, 120]. Accordingly, Leu has the second best point ECI at \mathbf{a} according to the cluster expansion. In fact, six of the top seven most favorable amino acids based on point ECI are also among the top seven most frequently observed amino acids at \mathbf{a}

positions [115].

We also applied the CE approach to two more compact folds - the Zn finger and the WW domain, and these differ from the coiled coil in several respects. First, higher order cluster functions are necessary for a good fit. Important triplet clusters can be either structurally compact or disperse. In compact triplets, the largest ECI correspond to combinations of large hydrophobic amino acids engaged in short-range van der Waals interactions. Examples of such clusters are shown in Figures 3-7 and 3-9A. Disperse clusters arise from long-range electrostatic interactions, and most significant ECI arise from triplets of charged and polar amino acids (see Figure 3-9B).

Another difference between the coiled coil and the two more globular systems is that the accuracy of the fit is better for the coiled coil. Cluster expansion can attain an arbitrary degree of accuracy provided enough terms are included. However, to derive statistically meaningful ECI for high-order interactions, enough sequences are needed to provide several instances of that interaction. Thus, it was easier to derive a good fit for the coiled coil, where only up to pair clusters were required, than to identify and fit the triplet and quadruplet terms necessary to describe the Zn finger and the WW domain folds. Ultimately, the desired target accuracy is dictated by the application. For protein design, where the goal is to find one or several good sequences, the magnitude of the error in all three systems is amply compensated by a sizeable increase in the accessible sequence space, especially given that the underlying full-detail physical models are only approximations themselves and do contain significant errors. For other applications, higher accuracy may be obtained by including more cluster functions and training on larger datasets, and/or by iteratively improving the CE fit by generating biased training datasets enriched with poorly fit sequences. Theoretically, because the complete expansion is exact, any desired level of accuracy can be attained. However, the cost of this (i.e. in time and memory requirements) depends on the specifics of the system under study, which is already apparent from the examples considered here. Alternatively, in cases where the accuracy of the expansion is not high enough for direct application, CE can be used as a highly efficient filter followed by evaluation with a higher resolution energy function.

A trend seen in all three systems is that the accuracy of the CE fit is worse after minimizing the structures and evaluating them with a non-pairwise decomposable energy function. This indicates that the energy resulting from this procedure is a more complicated function of sequence. Interestingly, in these cases fewer important higher order interactions are detected. This might indicate that structure relaxation reduces the importance of each high order interaction, so they are harder to detect, but there could be more of them. Even though the error is larger for cases with minimization, the actual energies are more informative because they are devoid of the unphysical van der Waals clashes that often result from optimization in discrete side-chain space. In addition, the computational speedup is especially significant here, as minimization and re-evaluation are computationally expensive.

3.3.1 Conclusion

The advantages offered by the cluster expansion methodology should make it widely useful in computational structural biology. We have demonstrated the application of CE to protein design problems in sequence spaces up to 10^{27} . Application to fold-recognition problems of similar size should be straightforward, although the best energy function to expand may differ from that used here. In both design and fold-recognition, CE can be applied to help relieve the fixed backbone approximation by expanding energies for several variants of the same structure. Once expansions are complete, evaluation of a sequence, or of all sequences in a proteome, on each of the backbones is extremely fast. Additionally, given the interpretability of CE, cluster expansions of many closely related structures may reveal key structure determinants.

The prospect that CE may be able to provide a general tool for approaching problems in protein structure prediction and design, beyond the initial demonstrations that we present here, is exciting. Where the limits of the approach lie remains to be explored. We have shown that the type of expansion required will be sensitive to the protein fold studied and to the nature of the energy function being expanded. Large proteins will require more parameters, and possibly more memory-efficient fitting procedures. It is easy to imagine many promising heuristics for choosing which

parameters to fit strategically, however, and/or for partitioning larger problems into smaller ones. We hope that the modeling community will join us in exploring the boundaries of CE for their own problems of interest. The potential payoffs, as we have demonstrated here, are very large.

3.4 Materials and Methods

3.4.1 Repacking and minimization

Energies for repacking were calculated in CHARMM based on parameter set 19 [25]. The energy function consisted of van der Waals energy (with atomic radii scaled to 90%), dihedral angle torsion energy, screened electrostatic interactions given by a distance dependent dielectric model and desolvation energy given by the EEF1 model [4, 103]. We treated the unfolded state by ignoring all side-chain-to-side-chain interactions and treating each side chain on a 5-residue stretch of its local native backbone. Rotamers were taken from the Dunbrack 2002 rotamer library [43]. We used our implementation of the dead end elimination (DEE) and A* branch and bound algorithms [42, 56, 58, 101, 105, 143] to find the optimal structure for each sequence. Given this structure, we calculated its folding energy $E_{\text{repack}}^{\text{fold}}$ using the potential used for repacking. To compute more accurate energies (devoid of large uninterpretable steric clashes and with better electrostatics), we subjected the solutions obtained with DEE to continuous side-chain minimization in CHARMM (10 cycles of steepest-descent minimization and 10 cycles of adopted basis Newton-Raphson minimization). The resulting structures were evaluated with an alternate energy function, in which 100% radii were used for van der Waals calculations and screening of electrostatic interactions was modeled using the Generalized Born model with “perfect” Born radii [140] computed using the program PEP [16] ($E_{\text{min,GB}}^{\text{fold}}$). For the zinc finger and WW domain, the same penta-peptide representation of the unfolded state as before was used for calculating reference energies. For the coiled-coil system, additional modifications were made to the unfolded state according to an

energy model previously shown to perform well in recognizing coiled-coil dimerization preferences (model HP/S) [60].

3.4.2 The coiled-coil unit cell

To derive a scoring function for coiled coils of arbitrary length, we expanded the energetics of a repeating structural element (unit cell). We postulated that interactions between amino acids more than one heptad apart in a coiled coil would not be appreciable and did not include clusters corresponding to these interactions in the CE. The unit cell was chosen to be a two-heptad dimeric parallel coiled coil (see figure 3.4). Additionally, to avoid edge effects, we used a periodic boundary condition for the backbone structure and sequence (see figure 3.3). Each periodic six-heptad training-set sequence was repacked as specified above. Cluster expansion was fit to just the energy of the central unit cell (all of the unit cell self energy and half of all interactions between the unit cell and the rest of the molecule), which allowed each interaction type to be counted exactly once. Thus the resulting ECI map exactly onto the energies of the corresponding interactions and can be applied for non-periodic sequences.

3.4.3 Cluster Expansion fitting

If energies for enough sequences are available, J_A^I can be solved for by standard fitting procedures (see equation 3.5). We used the method of pseudo-inverse [188] to perform least-squares fitting with an exponential weighting reducing the contributions of the less meaningful high-energy sequences. Therefore, for n cluster functions, the fitting procedure has an asymptotic running time of $O(n^3)$ and memory requirement of $O(n^2)$. Determining which of the M^N cluster function terms to keep in the fitting is not trivial (M is the number of residues possible at each site and N is the number of sites; for simplicity, we assume all sites to have the same number of possibilities). Although one may be guided by the notion that point terms are more important than pairs, which in turn are more important than triplets and so on, this is not always

true. We address the problem using the cross-validation (CV) score rather than the root mean square deviation (RMSD) to guide the fitting procedure. The CV score is the average error with which each sequence is predicted when left out of the fitting, and is a good measure of predictive power. When more CFs are included, the RMSD score decreases, while the CV score might increase (i.e. possible over-fitting) if the CFs are not physically relevant.

The fitting procedure used was as follows (see Figure 3.1). The number of sequences in the training set was chosen to be in the range of ~ 1.5 – 2.5 times the expected number of parameters in the fit (i.e. the number of parameters required to model up to all pair interactions). The constant and point CFs were initially included in the CF pool and used to compute a base-line value of the CV score; all pair CFs were considered as candidates for inclusion into the pool. For each pair cluster $\{i, j\}$ we considered all CFs associated with it (each corresponding to the contribution of a pair of amino acids) one at a time, and only those pair CFs that decreased the CV score were added to the pool. Because the contribution of a new CF (and its effect on the CV score) in general depends on the CFs that are already present, the order in which pair CFs are considered for inclusion into the pool is important. To determine a meaningful order, we first performed a fit with all pair CFs (in addition to the constant and points) to obtain fitting parameters J_i for each CF_i . Pair CFs were then considered in the order of decreasing $|J_i|$. Once all pair CFs were considered for inclusion, it was determined whether the quality of the fit (i.e. the magnitude of the CV error) was satisfactory. If it was not, we used the characteristics of poorly fit sequences $\Omega : |\Delta E| > D$ kcal/mol (i.e. those sequences with error larger than D kcal/mol, where D was 10 for unrelaxed cases and ~ 5 – 6 for relaxed ones) to locate important higher-order clusters (triplets and quadruplets). We calculated the information content $I^i = \ln(M) - S(p(\sigma^i|\Omega))$ for each site i and $I^{i,j} = \ln(M^2) - S(p(\sigma^i\sigma^j|\Omega)) - I^i - I^j$ for each pair of sites $\{i, j\}$ out of the amino-acid distribution in Ω . The terms $p(\sigma^i|\Omega)$ and $p(\sigma^i\sigma^j|\Omega)$ are the amino-acid distributions at site i and at the pair of sites $\{i, j\}$ in the sequence profile Ω , respectively, and $S(p) = -\sum_{\{p\}} p \ln p$ denotes the entropy of a probability distribution. Usually only

a few sites had significant point information content. Triplet and/or quadruplet CFs among sites with significant pair information content were manually added to the pool. The number of training sequences was increased (i.e. energies for more sequences were explicitly calculated) if the number of fitting parameters exceeded the number of sequences. For the un-relaxed cases with the Zn finger and the WW domain, the newly considered sequences were biased to include the amino-acid pairs over-represented in poorly fit sequences. All pair CFs in addition to the selected higher order CFs formed the new set of candidates. The procedure for considering candidate CFs one at a time was repeated as above and a final CV score was derived.

3.4.4 Zinc-finger design exercise

The energy models employed in this study do not account for protein solubility. Additionally, the rather crude unfolded state models make it difficult to properly estimate the overall relative point contributions of different amino acids at a given site. To get around these problems, we performed fixed composition design – an optimization problem in which amino-acid composition is held constant, but the sequence is free to change under this constraint [87]. This allows one to specify a reasonable composition that ensures likely solubility while relying on the optimization process to pick a permutation particularly well suited for the given backbone. An additional advantage is the cancellation of the unfolded state energy (assuming a strict composition dependence) across different sequences.

We used the zinc-finger sequence designed by Mayo and co-workers [35] (QQYT AKIK RTFR NQKQ LRDF IEKF KR), which has been experimentally characterized, to fix the amino-acid composition of our design. Note that because this sequence is quite heterogeneous, the search space of all unique permutations, 8.6×10^{20} , is very large and the design problem is still challenging. Each step of a Monte Carlo search in this fixed composition space amounted to picking two sites at random and swapping amino acids between them (if they were not already the same). Two Monte Carlo searches were run - one using repacking, minimization and re-evaluation according to $E_{\min,GB}^{\text{fold}}$ to score each sequence (direct design) and the other using cluster expansion

equivalent of the same energy function (CE design). The DEE and A* branch and bound algorithms for repacking [42, 56, 58, 101, 105, 143] were implemented in C. CHARMM [25] was used for continuous side-chain minimization and calculation of the van der Waals and EEF1 portions of the potential. PEP [16] was used to calculate atomic Born radii. A wrapper script that combined these steps for each sequence was written in perl. Sequence design code was written in C to use MPI and was distributed over 20 CPUs. The program for searching using cluster expansions was written in C without parallelization.

3.5 Acknowledgements

This chapter is based on a manuscript whose authors are Gevorg Grigoryan¹, Fei Zhou, Steve R. Lustig, Gerbrand Ceder, Dane Morgan and Amy E. Keating, published in PLoS Computational Biology, vol. 2:e63, 2006. We would like to thank the CSBi high-performance computing platform for computer time and support and S. Sia, K. Gutwin, X. Stowell, J. Apgar, and F. St-Pierre for comments on the manuscript. This work was supported by the NIH grant GM67681 to AK and by funding from the DuPont-MIT Alliance to GC and DM. The work used computing resources purchased with NSF equipment grant 0216437.

Chapter 4

Computing van der Waals Energies in the Context of the Rotamer Approximation

The rotamer approximation states that protein side-chain conformations can be described well using a finite set of rotational isomers. This approximation is often applied in the context of computational protein design and structure prediction to reduce the complexity of structural sampling. It is an effective way of reducing the structure space to the most relevant conformations. However, the appropriateness of rotamers for sampling structure space does not imply that a rotamer-based energy landscape preserves any of the properties of the true continuous energy landscape. Specifically, because the energy of a van der Waals interaction can be very sensitive to small changes in atomic separation, meaningful van der Waals energies are particularly difficult to calculate from rotamer-based structures. This presents a problem for computational protein design, where the total energy of a given structure is often represented as a sum of pre-calculated rigid rotamer self and pair contributions. A common way of addressing this issue is to modify the van der Waals function to reduce its sensitivity to atomic position, but excessive modification may result in a strongly non-physical potential. Although many different van der Waals modifications have been used in protein design, little is known about which perform best, and

why. In this paper we study ten ways of computing van der Waals energies under the rotamer approximation, representing four general classes, and compare their performance using a variety of metrics relevant to protein design and native-sequence repacking calculations. Scaling van der Waals radii by anywhere from 85 to 95% gives the best performance. Linearizing and capping the repulsive portion of the potential can give additional improvement, which comes primarily from getting rid of unrealistically large clash energies. On the other hand, continuously minimizing individual rotamer pairs prior to evaluating their interaction works acceptably in native-sequence repacking, but fails in protein design. Additionally, we show that the problem of predicting relevant van der Waals energies from rotamer-based structures is strongly non-pairwise decomposable and hence further modifications of the potential are unlikely to give significant improvement.

4.1 Abbreviations

vdW, van der Waals; RCE, rotameric conformational energy; NCE, neighborhood conformational energy; MAD, median absolute deviation; AAD, average absolute deviation; R60-95, modifications in which van der Waals radii are scaled by 60 to 95%; L-J, Lennard-Jones; LR_{90} , linearly repulsive van der Waals with 90% radii; PRM, pairwise rotamer minimization; LR_{90}^A , linearly repulsive van der Waals using 90% radii with all non-bonded terms capped; RR00, Richardson and Richardson penultimate rotamer library; RRexp, Richardson and Richardson penultimate library with expanded aromatics; RRX1, Richardson and Richardson penultimate rotamer library with expanded χ_1 ; Db02, Dunbrack rotamer library from 2002; Db99 Dunbrack rotamer library from 1999.

4.2 Introduction

It has long been known to chemists that molecules tend to adopt staggered, rather than eclipsed, dihedral conformations [50]. When the first few crystal structures of

proteins were solved, it became apparent that the same is true for amino-acid side chains [30]. Side-chain χ angles do not vary over all possible values, but rather cluster in tight distributions around conformations called rotamers (rotational isomers). Beginning in the 1970's, rotamer libraries were compiled to represent side-chain conformations observed in proteins of known structure [17, 30, 78]. Ponder and Richards developed the first complete rotamer library by examining 19 high-resolution crystal structures [145], and many variants of this work based on larger structural datasets have been published since then [45, 121, 171] (reviewed by Dunbrack [43]). The differences between most rotamer libraries lie in their size (number of rotamers per amino acid), the procedure used for discarding potentially bad experimental data and whether or not rotamers are defined as a function of backbone conformation. The rotamer libraries developed by Dunbrack and Cohen [44] and by Lovell *et al.* [113] are among the most commonly used today.

Most protein side-chains are observed to occur in conformations very close to library rotamers, a concept referred to as “rotamericity”. Shrauber *et al.* have shown that although significant outliers from rotameric conformations do exist, between 70 and 95% of all side chains in protein structures have χ angles within 20° of a rotamer [153]. Similarly, Richardson and Richardson estimated the rotamericity of their rotamer library, which they defined as the fraction of observed residues with χ -angles within 30° of a rotamer, to be 94.5% [113]. This, coupled with the fact that rotamers tremendously reduce the difficulty of sampling conformational space and allow for the application of many discrete optimization algorithms, makes it clear that the rotamer concept is very useful from a structural perspective.

It is not so clear that such a decomposition of structure space is justified from the energetic point of view. The landscape of protein conformational energies is very rugged - small changes in coordinates often lead to large changes in energy. Sampling this landscape at discrete structural points can lead to significant loss of information, because the apparent shape of the potential surface, and hence the locations of local and global minima, can change significantly depending on the sample points. This sensitivity of structure space-to-energy space mapping presents a challenge for many

problems to which the rotamer approximation is applied.

Computational protein design in particular is very sensitive to the rotamer approximation, especially when carried out on a fixed backbone. This type of calculation is based on evaluating the energies of various amino acid sequences adopting a given backbone structure. Scoring the compatibility of a sequence with a backbone is a sub-problem of protein structure prediction that requires placing side chains in appropriate conformations. This is often referred to as the side-chain packing (or repacking) problem, and the rotamer approximation is applied at this step by restricting side chains to rotameric conformations. Discrete optimization algorithms such as Dead End Elimination [42] can be applied to find the combination of rotamers giving the globally lowest energy for a given sequence [141]. Therefore, in protein design, the rotamer approximation is used not only to reduce the structure space but also to guide optimization on a very rugged energy landscape.

Most of the roughness of the protein energy landscapes comes from the van der Waals energy, due to its strongly repulsive nature at close distances. The standard approach in the field for addressing this problem is to modify the van der Waals potential to make it less sensitive to atomic position. This results in a less rugged energy landscape and potentially reduces the problems associated with discrete structural sampling. However, a potential disadvantage is that the resulting energy is less physical, so conformations found with this modified potential may be less relevant. Many different van der Waals modifications have been used in the field of computational protein design [68, 96, 100]. However, even though the possible advantages and disadvantages of using such potentials are recognized, little is known about how they compare with one another. In this study we test several widely used van der Waals modifications in various side-chain packing and design calculations. We find that modifications that scale van der Waals radii by $\sim 90\%$ perform best in most tests. Additionally softening the repulsive portion of the potential by linearizing it (hence introducing an energy cap) together with appropriate capping of all non-bonded terms improves performance further. We discuss key aspects of the problem and suggest some limitations for the performance of any pairwise-decomposable potential.

Table 4.1: Summary of Characteristics of the Protein Structure Set

PDB ID	Functional Class	Res	Lng	A	B	O	SCOP
1AMM	Crystallin	1.20	174	5	82	87	b
1BKR	Actin-Binding	1.10	109	62	0	47	a
1EW4	Unknown Function	1.40	106	34	35	37	a+b
1FUS	Hydrolase (Endoribonuclease)	1.30	106	17	29	60	a+b
1G8A	RNA-Binding Protein	1.40	227	57	72	98	a/b
1H4A	Eye Lens Protein	1.15	173	0	82	91	b
1IFC	Lipid-Binding Protein	1.19	132	16	79	37	b
1IFR	Immune System	1.40	121	0	53	68	b
1KNG	Oxidoreductase	1.14	156	47	35	74	a/b
1LU4	Oxidoreductase	1.12	136	43	32	61	a/b
1NG6	Structural Genomics, N/A	1.40	148	113	0	35	a
1O8X	Electron Transport	1.30	146	45	33	68	a/b
1P5F	Unknown Function	1.10	189	76	39	74	a/b
1QAU	Oxidoreductase	1.25	112	15	51	46	b
1QGV	Transcription	1.40	142	40	29	73	a/b
1R26	Electron Transport	1.40	125	45	28	52	a/b
1R29	Transcription	1.30	127	67	13	47	a+b
1UKF	Hydrolase	1.35	188	83	51	54	a+b
1X6X	Structural Protein	0.96	123	25	32	66	-
2LIS	Cell Adhesion	1.35	136	91	0	45	a

Res = resolution in Angstroms; Lng = length in residues; A, B, O = number of sites classified respectively as alpha helix, beta sheet, or other by STRIDE [53]; SCOP = SCOP classification of proteins [131].

4.3 Materials and Methods

4.3.1 Structural Database

Structures were selected from Protein Data Bank (PDB) entries that were determined by X-ray crystallography and had a resolution of 1.4 Å or lower. All structures had a single chain with at most 300 amino acids and none contained non-natural amino acids, metals, or other non-peptide chemicals in regions other than the surface. Out of approximately 900 structures that met these criteria, 20 were chosen manually to cover a diverse range of amino-acid composition, sequence, secondary and tertiary structure. Table 4.1 summarizes the structure set.

Compact structural regions were defined for each considered protein for further calculations. For each structure, an initial seed residue i was chosen at random and

this selection was then expanded to include all residues with at least one atom within 6 Å of vdW surface-to-vdW surface distance of any atom of residue i . This selection constituted a region. The residues within this region were then excluded from the set of candidates for choosing the seed for the next region. The procedure was repeated until no residues were left for selecting a seed for a new region. This resulted in a total of 208 regions with an average of 27 ± 10 sites per region. Based on the relative solvent accessibilities of residues in the native structure, each site was classified as core (< 13% accessible), boundary (between 13 and 49% accessible) or surface (over 49% accessible). NACCESS [75] was used to calculate relative solvent accessibilities.

4.3.2 Repacking, Design, Minimization and Evaluation

Regions defined as described above were used as templates in repacking and design calculations. Residues outside of the considered region were held constant in their native conformations. In sequence design, for reasons of computational time, the number of sites per region was limited to 20 (if the region had more than 20 sites, the first 20 closest to the seed residue were chosen). Each position was allowed six amino-acid possibilities, which included the wild-type amino acid along with five others drawn randomly from a set that depended on the burial classification of the site. For core sites the set to draw from was {C, G, A, V, L, I, F, Y, W, M} for surface sites it was {C, G, A, S, T, H, D, N, E, Q, K, R} and for boundary sites the union of these two sets was used. In all calculations, the 1999 release of the Dunbrack rotamer library was used [44], except where different rotamer libraries were compared.

The energy function used in repacking and design was as follows: $\Delta G = \Delta G^{vdW} + \Delta G^{elec} + \Delta G^{des} + \Delta G^{dih}$. All terms were calculated using the CHARMM param19 parameter set [25]. ΔG^{vdW} is the van der Waals energy modeled using the appropriate modification. ΔG^{elec} is the water-screened electrostatic interaction energy calculated using a distance-dependent dielectric (DDE) $\epsilon = 8r$, where r is atom-atom separation in Angstroms. ΔG^{des} is the desolvation energy calculated with the EEF1 model of Lazaridis and Karplus [103] and ΔG^{dih} is the torsion energy. Energies considered in

Table 4.2: Summary of Commonly Used Abbreviations

van der Waals Modifications	
R60 – R95	6-12 Lennard-Jones potential with radii scaled by 60 – 95%.
L-J	6-12 Lennard-Jones potential with standard (100%) radii.
LR_{90}	<u>L</u> inearly <u>R</u> epulsive van der Waals with <u>90%</u> radii. The function is linear from 0 to 10 kcal/mol in the repulsive range and uses 90% van der Waals radii.
LR_{90}^A	<u>L</u> inearly <u>R</u> epulsive van der Waals with <u>90%</u> radii with <u>A</u> ll non-bonded terms capped. Same as LR_{90} , but all non-bonded terms are capped at the value for the distance where the van der Waals energy is zero.
PRM	<u>P</u> airwise <u>R</u> otamer <u>M</u> inimization. A procedure in which each rotamer or pair of rotamers is minimized briefly in the context of the template prior to evaluation of self-energy or pair-energy terms.
Calculated Quantities	
RCE	<u>R</u> otameric <u>C</u> onformational <u>E</u> nergy. The energy obtained by directly evaluating a rotameric configuration.
NCE	<u>N</u> eighborhood <u>C</u> onformational <u>E</u> nergy. The energy of the structure reached by continuous side-chain relaxation of a rotamer-based conformation.
MAD	<u>M</u> edian <u>A</u> bsolute <u>D</u> eviation.
AAD	<u>A</u> verage <u>A</u> bsolute <u>D</u> eviation.

the folded state included intra-residue interactions, interactions of side chains with the template (all of the protein excluding the designed side chains) and pairwise side chain-to-side chain interactions. The unfolded state, which affects only sequence design calculations, was modeled as a set of non-interacting GGxGG penta-peptides with native backbone geometry, one per design site, with the appropriate amino acid substituted at x.

The side-chain packing problem was solved using the Dead End Elimination (DEE) algorithm [42, 56, 101, 143] followed by an A^* branch-and-bound search [58, 105]. The design procedure involved performing a Monte Carlo search in sequence space using the energy obtained from side-chain packing to score each sequence. For each region, 10 searches with 1,000 steps each were performed with the temperature annealing linearly from 1,000 to 200 K. The 100 sequences with lowest energy were kept for each region. Additionally, for each region 100 random sequences were considered and repacked. To generate non-optimal structures using the native sequence, 100

Monte Carlo searches, each with 1,000 steps, were performed in rotamer space. The lowest-energy structure from each of the 100 searches was kept for analysis for each region. Although most of these 100 structures had reasonably low energies, due to the ruggedness of the conformational energy landscape, high-energy structures did infrequently result from the Monte Carlo sampling. These rare structures introduced a considerable amount of noise when evaluating average absolute deviations (AAD, defined below), which made it difficult to compare different modifications. To remove this effect, only the 90 lowest-energy structures from each region (out of 100) were considered for calculating the average within-region AAD in Figure 4-4c. The outlier-insensitive median absolute deviation (MAD, defined below) was still calculated using all 100 structures per region. AAD, MAD and other abbreviations used commonly in this paper are defined in Table 4.2 for easy reference.

Given a particular rotameric solution, its energy was extracted directly from the pre-calculated energy tables for design as the van der Waals component (evaluated using the appropriate modification) of the total energy in the folded state. This energy is referred to as the rotameric conformational energy, or RCE. Rotameric structures were subjected to 10 steps of steepest descent followed by 10 steps of adopted basis Newton-Raphson side-chain minimization in CHARMM. The resulting energy is referred to as the neighborhood conformation energy, or NCE. We used a short minimization procedure because the predominant change in van der Waals energy occurs in the very beginning of side-chain minimization. We repeated some of the tests presented in this paper (those having to do with native-sequence repacking) using minimization to convergence with no significant changes in results (data not shown). The standard 6-12 Lennard-Jones potential, along with all bond, angle, dihedral and improper dihedral terms, was used in this minimization.

All modifications except LR_{90}^A involved changing only the van der Waals component of the total energy. With LR_{90}^A , in order to avoid side effects associated with capping exclusively the van der Waals energy, DDE electrostatic and EEF1 desolvation terms were capped as well. If a pair of atoms with opposite charges had a vdW energy above zero (atoms were closer to each other than $R_{min}/\sqrt[6]{2}$, where R_{min}

is the equilibrium van der Waals distance), the DDE electrostatic interaction energy between these atoms was calculated using $R_{min}/\sqrt[6]{2}$ as the interatomic distance. A similar capping was done for EEF1 desolvation except that it was applied regardless of whether the mutual atomic desolvation was favorable or unfavorable. Note that according to the theory of the solvent-exclusion model underlying EEF1 [103], the maximal desolvation energy of any given atom, defined as the integral over all space of the solvation free energy density function, is finite and hence the mutual desolvation of any two atoms can not diverge even at zero distance. However, the manner in which the integration is approximated computationally (the free energy density function in the center of the excluding atom is multiplied by the volume of the atom) does cause it to diverge because the density function itself tends to infinity towards the center of the desolvated atom.

4.3.3 van der Waals modifications

In order to allow for arbitrary alterations of the potential used, a program was implemented in C to calculate non-bonded self and pair terms. In the absence of modifications, the program produced van der Waals, distance-dependent dielectric, and EEF1 energies in excellent agreement with CHARMM (within machine error). All radius scaling modifications were calculated by changing the parameter files. For LR_{90} and LR_{90}^A special modifications were implemented. For LR_{90}^A these modifications included capping the attractive DDE and all EEF1 interactions. Additionally, 90% radii were used with both LR_{90} and LR_{90}^A .

Because modification PRM involved continuous side-chain minimization, interactions for it were calculated directly in CHARMM with 100% radii from parameter set 19 [25]. Before evaluating the self energy for each rotamer, the rotamer was subjected to 5 steps of continuous steepest descent minimization in the presence of a fixed template (backbone and side chains of non-design sites). Similarly, before evaluating the interaction energy for each pair of rotamers, the pair was also minimized for 5 steps in the presence of the fixed template.

4.3.4 Statistical Measures

Because of the presence of significant outliers in much of the data analyzed in this study, we adopted a median-based statistical measure of correlation - the median of the absolute value of prediction errors (Median Absolute Deviation or MAD). For a given set of true and predicted values (e.g. T and P) the MAD was calculated as the minimum of $\mathbf{median}(T - s \cdot P)$ with respect to slope s . In order to find the optimal slope, a grid search over all angles from 0 to 90° was performed by considering 6 focusing iterations, each breaking the current range into 100 intervals and zooming in on the lowest point \pm one interval. Once this slope was found, effectively defining the lowest-median-error line, the MAD was calculated with the formula above. Additionally, the average absolute deviation (AAD) was also calculated using the same slope as $\mathbf{median}(T - s \cdot P)$. For within-region RCE-to-NCE agreement, to account for the effect that some regions may have a constant RCE offset, the MAD was calculated as the minimum of $\mathbf{median}(T - s \cdot P - b)$ with respect to the intercept b . In these cases, the slope s was not optimized and was taken from the least-MAD line for the given modification in the corresponding cross-region case, so that each modification had a characteristic slope that was independent of structural region. In general, we found that optimal slopes for within-region agreement were very close to the optimal one for cross-region agreement. To assist in analyzing the raw data describing RCE-to-NCE agreement, we used the lowest-MAD slope to automatically generate plots zoomed in on the relevant region of data (where most of the data points lay), by effectively ignoring outliers. The procedure for generating these plots entailed setting the upper limits of the y and x axes to be equal to the highest NCE in the dataset and the RCE corresponding to it according to the lowest median error line, respectively.

4.4 Results

The rotamer approximation breaks protein structure space into discrete bins, each representing conformations closest to a particular rotameric configuration. This works well from a structural perspective, in the sense that most low-energy conformations

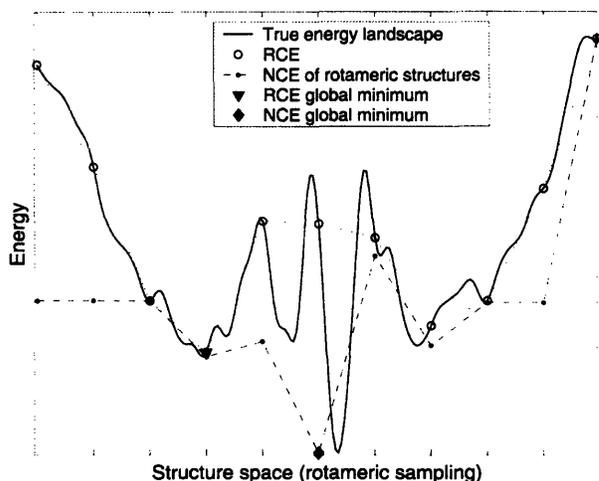


Figure 4-1: A cartoon representing the RCE and NCE landscapes. The solid line represents the true rugged protein energy landscape and open circles indicate points of discrete sampling using rotamers. For each rotameric structure, the NCE assigned to it is the energy of the closest local minimum (filled circles). In this example, the NCE represents much better than the RCE the shape of low-energy regions in the true energy landscape and, in particular, preserves the global minimum.

adopted by proteins have a very close rotameric representation [113, 153]. However, this means that when we score the energy of a particular rotameric configuration, assigning it a rotameric conformational energy, or RCE, we are actually assigning that same energy to an entire local structural region. If the RCE is unrepresentatively large, e.g. because of a slight steric clash, the entire neighborhood may be unduly excluded in a search. The energy that we want to compute is one that is characteristic of the entire structural neighborhood that the rotameric conformation approximates (hereafter referred to as the neighborhood conformational energy, or NCE). A suitable definition for the NCE would be the ensemble-averaged energy over all protein conformations that fall within the same structural bin as the rotamer-based conformation in question. Assuming that this is dominated by the local minimum in this region, in this study we define NCE as the energy of the structure reached by continuous side-chain relaxation of a rotamer-based conformation, a quantity that can be easily calculated using any standard force field. Figure 4-1 shows a cartoon of an energy landscape with the NCE and RCE indicated.

We sought to examine the difference between the NCE and RCE. If this difference

Table 4.3: Changes resulting from minimization of repacked rotameric structures

Quantity	Mean Diff.	Median Diff.	Stdev
6-12 Lennard-Jones energy	$1.3 \cdot 10^5$ kcal/mol	305 kcal/mol	$6.0 \cdot 10^5$ kcal/mol
Coulombic energy with $\epsilon_{int} = \epsilon_{ext} = 4^b$	6.2 kcal/mol	6.1 kcal/mol	3.5 kcal/mol
Torsion energy	16.5 kcal/mol	15.5 kcal/mol	14.2 kcal/mol
Hydrogen bonding energy	3.2 kcal/mol	2.2 kcal/mol	3.4 kcal/mol
EFF1 desolvation energy	17.8 kcal/mol	17.0 kcal/mol	6.3 kcal/mol
Delphi [73] polarization energy upon moving from $\epsilon_{int} = \epsilon_{ext} = 4$ to $\epsilon_{int} = 4$, $\epsilon_{ext} = 80^b$	3.8 kcal/mol	3.1 kcal/mol	2.8 kcal/mol
Solvent accessible surface area multiplied by 10 cal/mol·Å ²	0.2 kcal/mol	0.2 kcal/mol	0.2 kcal/mol
Sidechain position (all atom) ^a	0.16 Å	0.15 Å	0.06 Å

For each region, the sidechain root mean square deviation between pre- and post-minimized structures along with absolute differences in several energy terms upon minimization were calculated and the average, median and the standard deviation over all regions are reported. ^a – difference here is defined as RMSD. ^b – ϵ_{int} and ϵ_{ext} ext represent internal (protein) and external (solvent) dielectric constants.

is small, using the energy of a rotamer-based structure as a measure of its NCE is justified. To investigate this, we defined a set of structurally compact regions from high-resolution crystal structures that included a diverse collection of secondary and tertiary structure environments (see Table 4.1). In total, 208 regions were selected with an average of 27 sites per region. Each of these regions was subjected to native-sequence repacking and the lowest-RCE rotameric solutions were then relaxed using continuous side-chain minimization, as shown in Figure 4-2. Table 4.3 summarizes the change in structure and energy resulting from this minimization. While the amount of structural change is very small (average side-chain RMSD 0.16 Å), conformational energies do change significantly. In particular, the van der Waals energy, computed with the standard 6-12 Lennard-Jones potential, changes by many orders of magnitude on average, with the median deviation of 305 kcal/mol also much higher than that for other terms. Thus, in accord with intuition, most of the energy landscape ruggedness comes from the van der Waals term, which presents a problem for estimating NCE from rigid rotamer-approximated structures. Because it is the greatest source of error resulting from the rotamer approximation, the rest of this paper will address van der Waals energies and we will use RCE and NCE to refer to only the van der Waals component of the total energy.

4.4.1 Overview of van der Waals modifications

The first uses of van der Waals potential modifications in computational structural biology predate the rotamer approximation in protein design. For example, in 1983 Levitt used a 6-12 Lennard-Jones potential with repulsion capped at 10 kcal/mol to improve convergence properties of molecular dynamics and energy minimization simulations [107]. Later this potential was adapted by Koehl and Delarue in a method for side-chain repacking using a self-consistent mean field approach [86]. Several investigators have since modified the repulsive portion of the Lennard-Jones potential in computational protein design. Desjarlais and Handel capped the standard 6-12 Lennard-Jones potential at 100 kcal/mol [41]. Kuhlman and Baker used a modification in which the repulsive Lennard-Jones region was replaced with a linear ascent

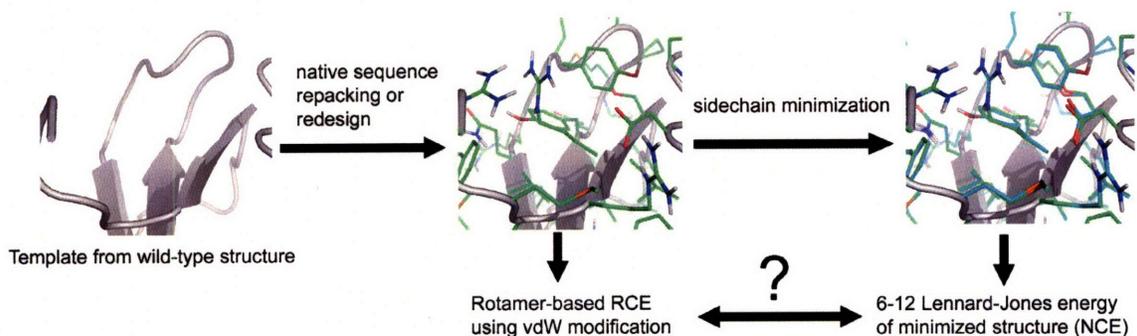


Figure 4-2: Overview of the computational experiment. Given a particular structural region, the native backbone was used as a template and either the native sequence was repacked or alternative low-energy sequences were designed using a specific van der Waals modification. The solutions resulting from these calculations were then subjected to 20 steps of continuous side-chain minimization using the L-J potential, which had only a small effect on side-chain geometry. The van der Waals energy of the minimized structure, evaluated with the 6-12 Lennard-Jones potential (NCE), was compared to the energy of the rotamer-based structure given by the van der Waals modification used in repacking and design (RCE).

to 10 kcal/mol [96]. Another version of the same modification had the linear portion taking effect at $\frac{3}{4}$ of the minimum energy van der Waals distance and having a slope identical to that of the Lennard-Jones potential at that point [95].

Van der Waals energies can also be modified by altering atomic parameters, rather than adjusting the functional form. The most common type of parameter modification is the uniform scaling of atomic radii. It is presumed that reducing the van der Waals radii implicitly accounts for the side chain and backbone relaxation that occurs to relieve rotamer clashes. The most common scale factor reported in the literature is 90% [4, 68, 100], although 95% is also used [37, 96]. The historic reason for this choice is a study by Dahiyat and Mayo, in which four different radius scale factors between 1.07 and 0.7 were used to design variants of protein G β 1 domain. The peptide resulting from the design with scale factor 0.9 was experimentally shown to be the most well-ordered and to have the highest stability [36].

Another flavor of commonly used van der Waals modifications involves subjecting rotamers to side-chain minimization before calculating their interactions. Vsquez performed native-sequence repacking with rotamers minimized in the presence of

the template, such that each side-chain position had a custom rotamer library [182]. Havranek and Harbury took the same approach in a protein design study [68]. Wodak and coworkers additionally minimized all rotamer pair interactions as well as single-rotamer interactions with the template [186].

Finally, as an alternative way of accurately predicting relevant van der Waals energies from rotameric structures, some investigators have expanded rotamer libraries. Xiang and Honig [192] significantly expanded their rotamer libraries for native side-chain repacking and showed that this led to considerable improvements in χ -angle recovery. Mayo and co-workers [34, 100, 156], Havranek and Harbury [68] and Baker and co-workers [10] have used less expanded rotamer libraries in protein design.

The approaches listed above all share the property that vdW energies can be computed as sums of single-rotamer and rotamer-pair interaction terms. This is necessary for some search algorithms. If this condition is lifted, other classes of modifications can be used that involve, for example, minimizing residues “on the fly” during the search procedure. For example, Baker and co-workers have used a related approach for side-chain placement in docking [184]. The advantage of pairwise decomposability, however, is that globally optimal solutions in RCE can be found.

Clearly, many different van der Waals modifications have been attempted over the years, often with scant justification. It would be difficult to test them all. Our approach was to pick examples that represent each class of commonly used modifications. Because the results by Dahiyat and Mayo [36] turned out to be quite seminal for the field, we tested a range of radius scaling modifications to uncover trends and to see if 90% is indeed the optimum. We considered scale factors from 60% to 100%, referred to here as R60 - R95 and L-J. We also analyzed a modification where both rotamer self and pair energies were evaluated after short side-chain minimization; this is referred to as PRM (pairwise rotamer minimization). Finally, as a representative of modifications in which the shape of the repulsive part of the potential is explicitly altered, we considered a function with a linear ramp-up to 10 kcal/mol after the 6-12 Lennard-Jones potential crosses zero. However, for many atom type pairs this made the initial slope of the van der Waals repulsion steeper than that of the 6-12

Lennard-Jones potential, resulting in a modified potential that was more restrictive than the original in a distance range where many interactions are expected to lie. To avoid this problem, we applied the linear repulsion modification in conjunction with 90% van der Waals radii. This modification is referred to as LR_{90} (linearly repulsive van der Waals with 90% radii). Using LR_{90} , we quickly discovered that setting a limit for van der Waals repulsion but not for other non-bonded terms gave very unrealistic structures in repacking and design. For this reason, we considered an additional modification, where atomic pairwise desolvation and attractive electrostatic interactions were capped as well (see Materials and Methods and below). This modification is referred to as LR_{90}^A (linearly repulsive van der Waals using 90% radii with all non-bonded terms capped). Finally, to explore the effect of rotamer libraries, we considered those proposed by Dunbrack and Cohen [44] and by Lovell *et al.* [113], as well as variants thereof.

4.4.2 Modified van der Waals energies versus NCE

A common approach to computational protein design involves using side-chain packing to score candidate sequences. Given such a framework, two important considerations arise regarding the van der Waals potential used. For a given sequence, the energy of the optimal conformation obtained using a modified potential should be a good estimate of the NCE of that conformation, as the quality of the sequence is judged based on this energy. We refer to this as cross-sequence agreement. Also, to ensure that the lowest-RCE conformation obtained for a sequence is relevant, energies assigned to other conformations must also be good estimates of their NCEs. We refer to this as cross-conformational agreement. We compared the performance of different van der Waals modifications at providing both types of agreement by performing a large number of sequence repacking and design calculations and evaluating the results using a panel of metrics. It was neither practical nor representative of common design applications to consider entire proteins at once. Therefore, these calculations were run on the same 208 compact structural regions defined above.

RCE-to-NCE agreement across different regions

Two types of calculations were performed on each structural region and for each van der Waals modification tested. In the first, the native-sequence was repacked. In the second, we used Monte Carlo sampling to generate 100 low-energy sequences. The NCE was then computed by relaxing the resulting structures and re-evaluating their van der Waals energies with the L-J potential (Figure 4-2). Cross-sequence agreement is quantified as the correlation between the RCE and NCE of these configurations, and results for three vdW modifications are shown in Figure 4-3. The overall agreement is poor in almost all cases, with many significant outliers making it difficult to observe trends. Some standard ways to assess correlation, such as root mean square deviation and the correlation coefficient, are very sensitive to outliers and fail to capture trends in the bulk of the data. In this study, we adopted a measure of agreement that is less sensitive to the presence of outliers – the median of absolute prediction error (median absolute deviation, or MAD). Additionally, we recognized that disagreement between predicted energies and NCE by a constant factor is tolerable, because this can be corrected by scaling. Therefore, the relevant measure of agreement is the MAD between s ·RCE and NCE, where s is chosen in such a way as to minimize the MAD, effectively defining the slope of the minimum-MAD line.

Insets in Figures 4-3a-b show the most relevant portions of the plots in the main figure. Also shown in Figure 4-3 are the least-MAD and least-squares lines. As expected, the least squares fit focuses almost entirely on the outliers, whereas the MAD is able to capture the main trend of the correspondence. Figures 4-4a-b show the MADs between RCE and NCE of either native repacked structures (Figure 4-4a) or designed structures (Figure 4-4b) for all of the tested van der Waals modifications. Potentials R85 and R90 give the best cross-sequence agreement out of the radius-scaling modifications. Capping vdW energies only (modification LR_{90}) does not work at all. In fact, given such poor performance in native-sequence repacking (an “easy” test, because the native sequence is clearly compatible with its backbone), LR_{90} was excluded from further analysis. On the other hand, LR_{90}^A , a modification that caps

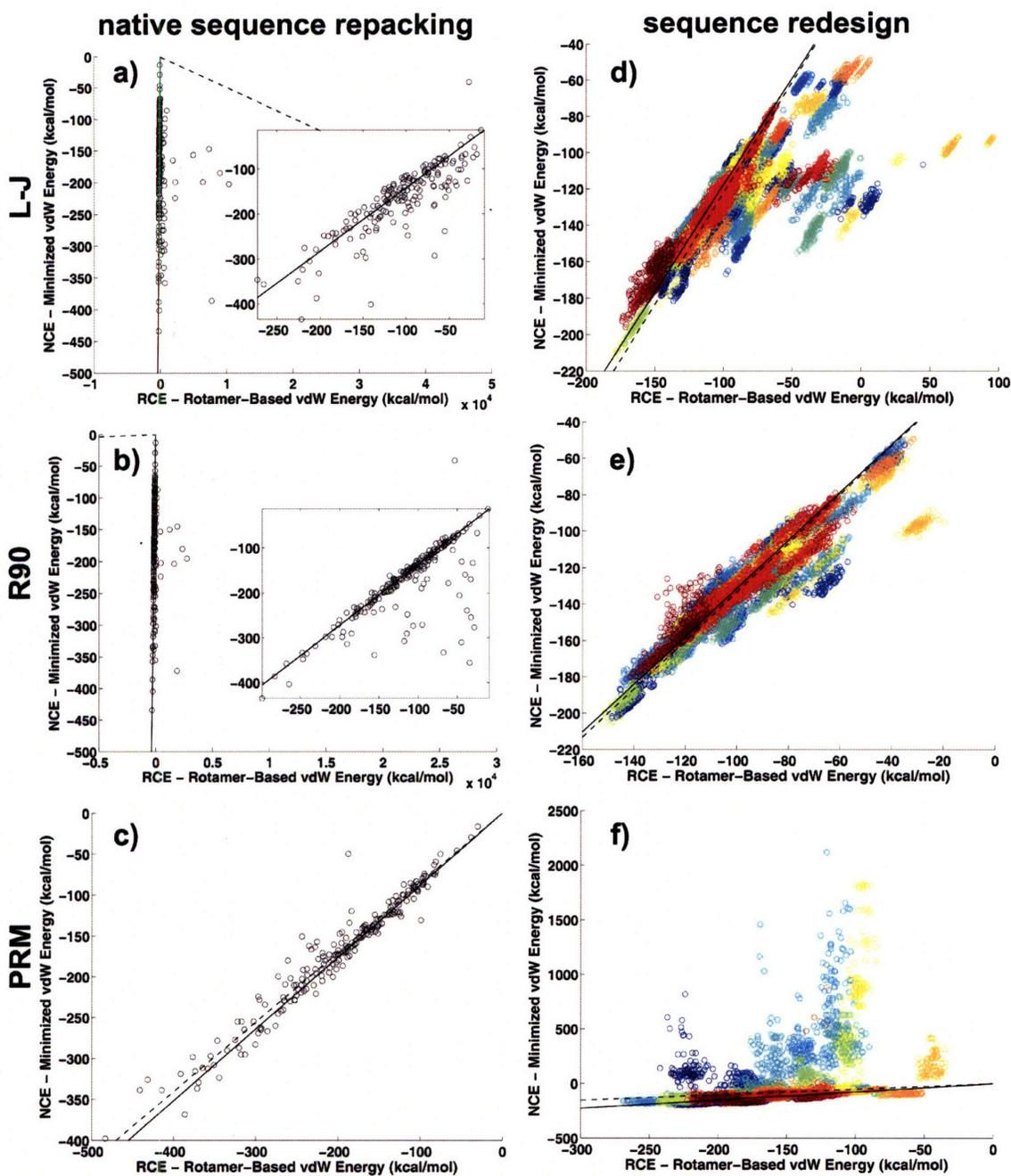


Figure 4-3: Scatter plots of RCE vs. NCE for three vdW modifications - L-J (in a and d), R90 (in b and e) and PRM (in c and f). Structures generated by optimally repacking the native sequence were used for a), b), and c), whereas d), e), and f) show the performance on low-RCE structures from sequence design. Each color corresponds to one of 208 regions. The solid and dashed lines are the least-MAD and RMSD lines, respectively. Insets in a) and b) are enlarged sub-regions of the main plots, where 23% and 9% of the points respectively are outside of the plotted region.

all non-bonded interactions, performs best overall. PRM works well in repacking, but it performs very poorly in sequence design (Figure 4-4b). We also report average absolute deviations from the least-MAD line (AAD), shown in circles in Figures 4-4a-b. The extent to which MAD and AAD are different is an indication of the presence of significant outliers. Although scaling of vdW radii by 90 or 85% does improve performance over L-J on the bulk of the data, significant outliers are still present, particularly in the case of repacking.

Within-region RCE-to-NCE agreement

In Figure 4-3, panels d-f, the color of the data points is used to indicate regions. For some regions, many of the points representing designed sequences lie off of the main diagonal by a roughly constant amount. These points decrease performance when we compare NCE with *s*-RCE. Agreement can be significantly improved if we introduce a region-specific intercept parameter, *b*, such that $s \cdot \text{RCE} + b$ is as close to NCE as possible. Because for many applications it is only necessary to compare structures and sequences within the same structural region (and thus the value of *b* does not matter), we further tested different vdW modifications by looking at within-region RCE-to-NCE agreement. For native-sequence repacking this amounts to looking at cross-conformational agreement. The set of conformations considered for each region consisted of the global RCE optimum structure along with 100 non-optimal conformations generated by Monte Carlo sampling. Due to the sampling procedure, most of the non-optimal solutions had reasonably low RCE and therefore effectively represented the energy funnel around the rotameric global optimum solution. Correct estimation of van der Waals energy for solutions in this funnel is important for identifying the relevant lowest energy conformation. Using these structures, Figure 4-4c shows the within-region RCE-to-NCE MAD and AAD averaged across regions. Modification LR_{90}^A gives the lowest AAD and MAD. Figure 4-4d shows within-region MAD and AAD for sequence design, where the 100 lowest-RCE sequences from the design procedure were considered for each region. In this case R95 performs best in MAD, whereas LR_{90}^A has the lowest AAD. The conclusions for within-region comparison are

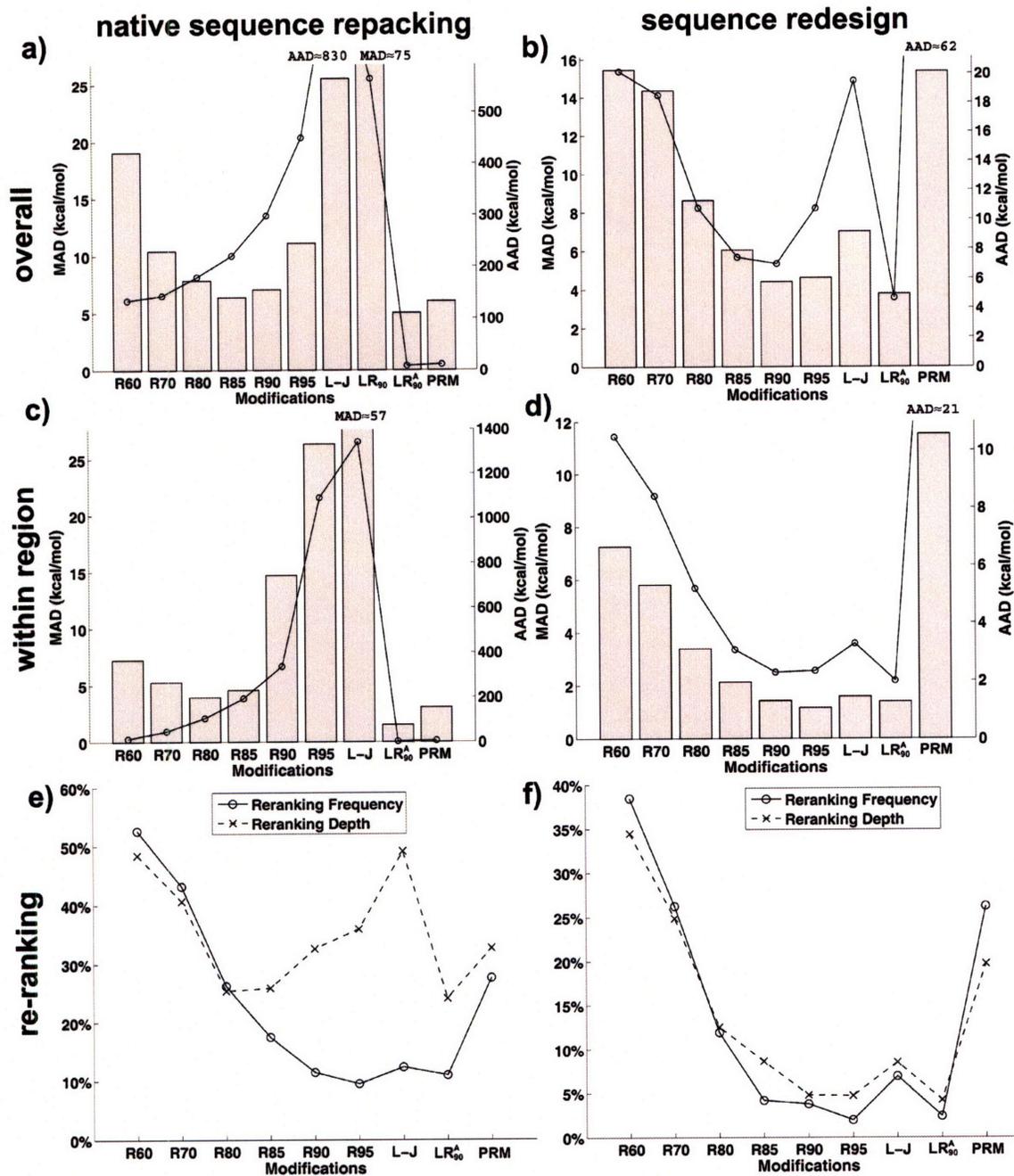


Figure 4-4: Performance of different vdW modifications on predicting the NCE of low-RCE structures in native sequence repacking (left panels) or sequence design (right panels). In (a) and (b), RCE-to-NCE agreement across different structural regions is considered and (c) and (d) report within-region averages. In (a), the lowest-RCE structure from native sequence repacking was used for each region, whereas an additional 100 non-optimal structures for each region were considered in (c). In (b) and (d) 100 low-RCE solutions from sequence design were used for each region. Bars and circles represent MAD and AAD (left and right y-axes) respectively. For data points outside of the limits of the graphs, values are shown. (e) and (f) show average within-region re-ranking frequencies (circles, averaged over all regions) and re-ranking depths (crosses, averaged over regions where re-ranking occurred).

thus very similar to the conclusions for the cross-region case.

As another metric of within region RCE-to-NCE agreement, we looked at the frequency with which structures re-rank upon switching to the more accurate estimate of energy. We defined frequency of re-ranking as the fraction of solutions that have a lower NCE than that of the lowest-RCE solution. Strictly speaking, re-ranking should be defined in terms of the total energy. However, as Table 4.3 shows, van der Waals energy is the term that changes the most upon structural relaxation, so a significant re-ranking in this term will give rise to similar re-ranking in the total energy. In Figure 4-4, panels e-f show the average within-region re-ranking frequencies in native-sequence repacking and sequence design using the same set of structures as in Figures 4-4c-d. R95 shows the lowest frequency of re-ranking in both tests.

In addition to knowing how often structures re-rank when switching from RCE to NCE, it is also useful to know how many rotamer-based solutions around the RCE minimum one has to sample to be certain that the right local NCE minimum is captured. Figure 4-4 panels e-f also show the average RCE rank of the solution that ends up with the lowest NCE (dotted lines with crosses). We refer to this measure as the depth of re-ranking. Modification LR_{90}^A performs best in this test for both native-sequence repacking and sequence design, whereas the optimal radius scale factor lies somewhere between 80 and 95%.

Global properties of the RCE energy landscape

In the tests above we examined the RCE-to-NCE agreement of either the native sequence or sequences judged to be reasonable for the backbone by the particular energy function used. If this agreement is poor, then the low-RCE solutions discovered by protein design are selected based on an incorrect estimate of energy. However, for high-RCE sequences it is also important that their RCE be a reasonable estimate of their NCE. To test this, we repacked 100 random sequences in each region using each of the van der Waals modifications and measured the frequency with which these had lower NCE than that of the lowest-RCE sequence from design (Figure 4-5a). Because they are random, these 100 sequences are inappropriate for the corresponding

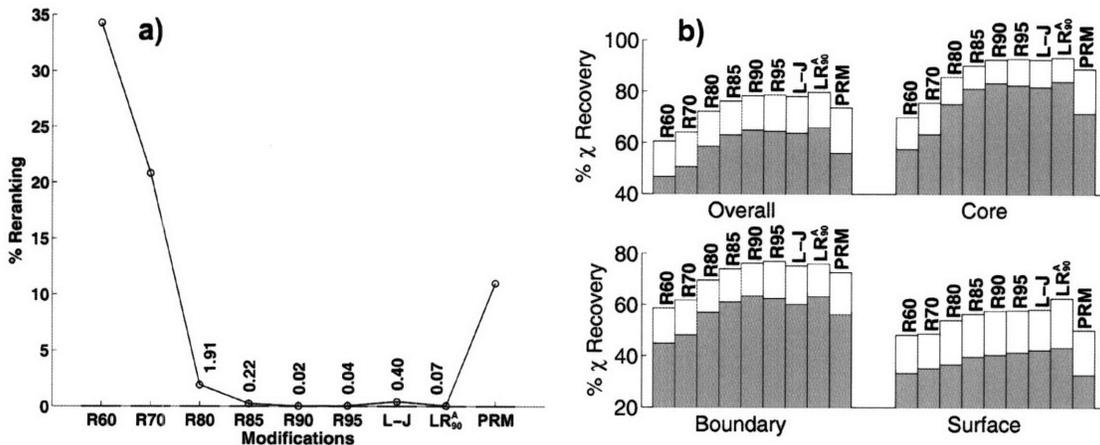


Figure 4-5: Global tests of the RCE energy landscape. (a) NCE re-ranking of random sequences with the lowest-RCE sequences from design. Re-ranking is calculated for each region and shown are cross-region averages for each modification. Re-ranking here indicates global differences between the RCE and NCE-based energy landscapes. (b) χ -angle recovery in native-sequence repacking using different van der Waals modifications. Full bar heights represent the fraction of χ_1 angles predicted correctly and the shorter grey bars correspond to the fraction of χ_1 and χ_2 angle combinations predicted correctly. χ -angle recovery indicates how close the global RCE optimum is to the optimum of the true energy landscape.

backbones and have much higher RCE than the optimized sequences. However, for all modifications the average frequency of re-ranking is non-zero, and it is significant for all but R90, R95 and LR_{90}^A .

As a final test of the ability of different modifications to predict reasonable low-energy structures, we looked at native χ -angle recovery rates. Figure 4-5b shows these data for all of the tested modifications. Consistent with previous observations, R95 and R90 are the best radius-scaling modifications, and LR_{90}^A is best overall.

Rotamer libraries as alternatives to van der Waals modifications

Van der Waals modifications are used to compensate for the rotamer approximation. In the limit of very many rotamers, the van der Waals potential need not be modified at all. Further, some rotamer libraries may be better than others at appropriately sampling the rugged energy landscape. In the calculations above we used the 1999 release of Dunbrack’s rotamer library (DB99) [44]. We repeated all the calculations for

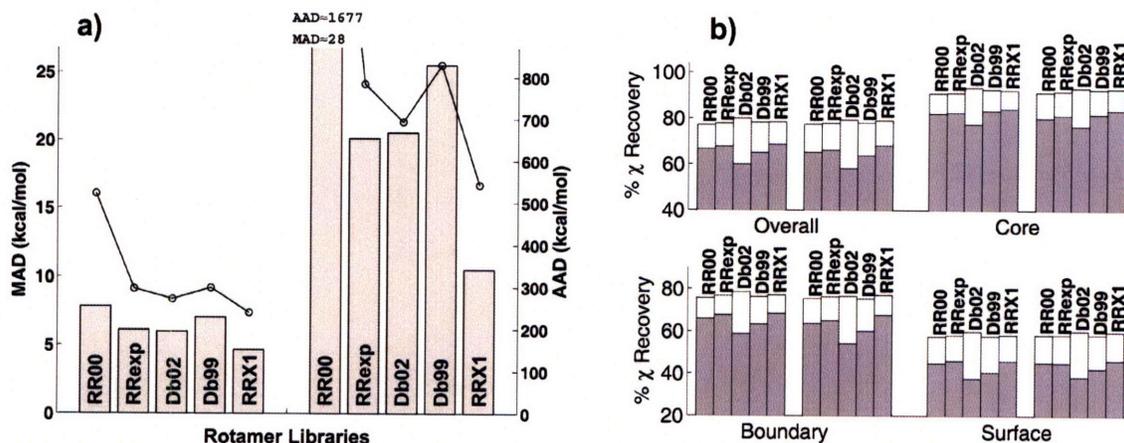


Figure 4-6: Comparison of different rotamer libraries using either the R90 or L-J potentials (first and second sets of bars respectively). **a)** MAD (bars, left y-axis) and AAD (lines, right y-axis) of RCE-to-NCE agreement in native sequence repacking. **b)** χ_1 (full bars) and χ_{1-2} (shorter grey bars) angle recovery.

the 2002 release of the Dunbrack library (DB02) [43], the Richardson and Richardson penultimate rotamer library [113] (RR00), the Richardson library with χ_1 angles expanded for aromatic residues (F, Y and W χ_1 angles expanded by $\pm 5^\circ$ and $\pm 10^\circ$; RRexp), and the Richardson library with all χ_1 angles expanded by \pm the standard deviation (obtained by dividing the half width at half maximum by $\sqrt{(2 \cdot \ln 2)}$; RRX1). The libraries varied in size, consisting of 391 (DB99), 370 (DB02), 175 (RR00), 243 (RRexp), and 521 (RRX1) rotamers. Figure 4-6a compares the MAD and AAD of different rotamer libraries in native-sequence repacking using either the unmodified Lennard-Jones potential or the 90% radius modification. In both cases the ranking of rotamer library performances is the same, with the largest (RRX1 with 521 rotamers) and the smallest (RR00 with 175 rotamers) performing best and worst, respectively. The performance of intermediate-sized libraries does not exactly follow library size. The effects of increasing the rotamer library size by a factor of ~ 3 are minimal compared to the significant improvement in MAD and AAD upon reducing the size of the atomic radii to 90%. No dramatic change was found from either type of modification in native χ -angle recovery (Figures 4 – 5b and 4 – 6b).

Why the problem is hard - the price of pairwise decomposability

Most protein repacking and design algorithms require a pairwise decomposable energy function, and typically all interactions between rotamer pairs are pre-calculated [141]. For a fixed structure, the L-J potential meets this criterion as it can be calculated as a sum of contributions from rotamers and pairs of rotamers. However, once a rotameric structure is allowed to undergo molecular-mechanics minimization, the rotamer-level pairwise-decomposability is lost. This is because the exact manner in which a side chain relaxes depends on its entire structural environment. However, many inherently non-pairwise decomposable measures can be accurately modeled in a pairwise manner. For example, Mayo and co-workers have shown that solvent-accessible and buried surface areas of proteins, even though not strictly decomposable into residue pair contributions, can be effectively approximated in this way [165]. Similarly, pairwise decomposable solvation models have been developed that approximate the exact continuum dielectric results [67]. So the relevant question is not whether predicting NCE from rotamer-based structures is pairwise decomposable – it is not – but rather what the limits of a pairwise approximation are.

One way to analyze this is to look at the mapping between atom-to-atom distances of a rotameric structure (\vec{R}_{ij}) and the corresponding atom-pair interaction energies that result upon minimization (\vec{E}_{ij}^{min}). If this mapping is close to functional, it will be possible to derive a good pairwise expression. Figure 4-7 shows this mapping for a set of rotameric structures and their minimized versions. Even though the overall distribution of data does resemble a Lennard-Jones-like shape, there is significant fuzziness, i.e. the same distance can map to very different energies, depending on the structure.

A pairwise potential that predicts individual atomic interactions in minimized structures, given rotameric structures, is sufficient but not necessary for protein design. All that we need is for the total NCE to be approximately decomposable into contributions from pairs of rotamers (i.e. there can be some cancellation between the errors of predicting atom-pair interaction energies). To analyze the degree of

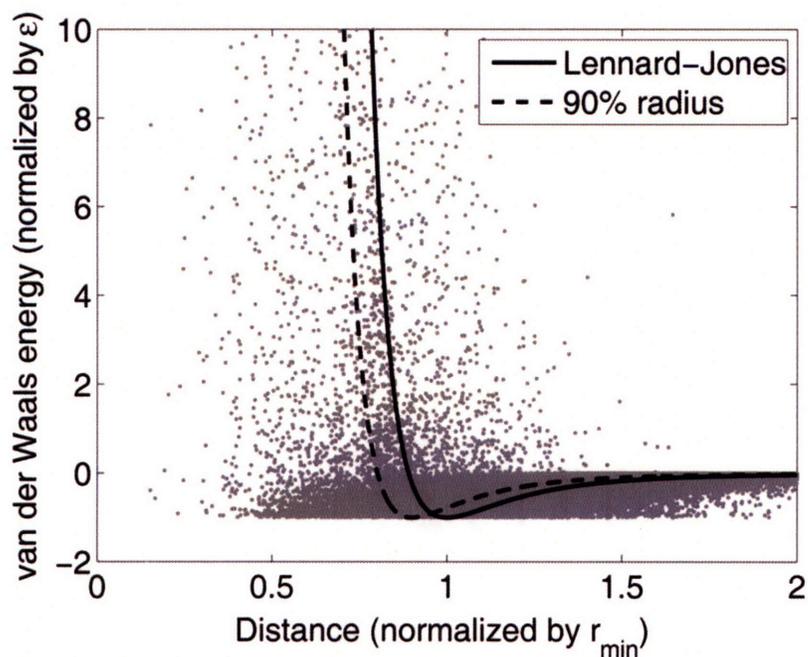


Figure 4-7: Mapping from pre-minimization atom-to-atom distances onto post-minimization atomic interaction energies. Axes are unit-free; for each interaction, the distance is normalized by the equilibrium distance ($r_{min} = r_i + r_j$, where r_i and r_j are the van der Waals radii of interacting atoms) and energies are divided by the well-depth ($\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$, where ϵ_i and ϵ_j are the well-depth parameters of interacting atoms). The full and the dashed lines correspond to the 6-12 Lennard-Jones potential and R90, respectively. Approximately 0.7 % of the data points have scale-free interaction energies above 10 and are not shown. Each structure used for this analysis was generated by perturbing one of the side chains of the native structure of 1AMM (an entry from the dataset in Table 4.1) to a rotamer of the same amino acid selected from the RR00 library [113]. All native amino-acid rotamers were considered in all of the 28 sites of the first region defined on 1AMM, giving rise to 278 structures.

non-pairwise decomposability of NCE at the rotamer level, we looked at the NCE contribution of a given pair of rotamers at a pair of sites as a function of the rotameric states of surrounding sites. Figure 8 shows the results for a set of rotamer pairs that have van der Waals overlap in their rotameric states. The contribution of a pair of rotamers can vary over several orders of magnitude depending on its structural environment. Notably, the strong contextual dependence makes it hard to identify rotamer pairs that should be eliminated due to unfavorable interaction. Figure 4-8b shows the fraction of rotamer pairs that have their lowest NCE-contribution, out of the ten environments sampled, below a cutoff. For over 30% of the rotamer pairs considered, there are structures where the NCE contribution of the pair is negative. On the other hand, the range of interaction energies among these same 30% of rotamer pairs is consistently over 20 kcal/mol.

Reasons for the success of radius scaling: rotamer-level interaction analysis

In most of the tests performed in this study, modifications that scaled the van der Waals radii by 85-95% (LR_{90}^A , R85, R90 and R95) showed best performance. To examine the reason for this (and to isolate the effect of radius scaling), we compared modification R90 with the original Lennard-Jones potential. We considered the set of structures from native-sequence repacking generated by the L-J potential and found that R90 scores these with a lower MAD and AAD than L-J (MAD and AAD, respectively, were 6.7 and 301 kcal/mol for R90 and 25.5 and 829 kcal/mol for L-J). To determine whether this improvement arises from changes in the repulsive part of the potential, the attractive part, or both, we scored the same structures with a hybrid potential (L-J90) in which the L-J value for an atom-pair interaction energy was used if it was below +1 kcal/mol and the energy from R90 was used otherwise. Thus, L-J90 and L-J are almost identical, since atomic interactions above 1 kcal/mol constitute only $\sim 7\%$ of all significant atomic interactions (defined as those with energy magnitudes above 0.1 kcal/mol). Surprisingly, L-J90 gave rise to MAD and AAD values comparable to and even slightly lower than those of R90 (5.8 and 246 kcal/mol respectively). Thus, the improvement in performance offered by R90 mainly comes from

the adjustment to the repulsive part of the potential. However, we have shown previously that accurately predicting the contribution to the NCE of a clashing interaction in a pairwise manner is essentially impossible, so R90 must improve performance in some other way. Figure 4-9 shows the correlation between NCE and RCE contributions, as predicted by either R90 or L-J, of all repulsive atomic interactions with a Lennard-Jones energy above 1 kcal/mol. Neither R90 nor L-J show any appreciable correlation. In fact, the RCE-to-NCE correlation coefficient is 0.05 for R90 and 0.07 for L-J, whereas it is 0.99 for the correlation between R90 and L-J. The difference between the two potentials, however, is that R90 predicts all energies to be lower, which brings its estimates closer by value to the NCE contributions. Therefore, R90 treats repulsive interactions better not because it can recognize when clashes resolve upon minimization and when not, but because it indiscriminately reduces the energy of all clashes, which on average brings it closer to the right answer.

4.5 Discussion

Computational structure prediction and design rely heavily on the concept of side-chain rotamers and the formulation of rotamer libraries [17, 30, 78, 145]. An overwhelming majority of side chains in crystal structures exist in very nearly rotameric conformations [113, 153]. However, we have shown that sampling conformational energy in rotamer space can lead to an apparent energy landscape that is very different from the true energy landscape. Most of this difference comes from van der Waals conformational energies, which change much more than any other term upon relaxation of rotamer-based structures (see Table 4.3). We tested several types of modified potentials that have been used in the literature to compensate for the use of rotamers. The ideal vdW modification would be one for which the energy landscape it describes in rotamer space resembles closely the NCE-based energy landscape (see Figure 4-1). However, it is not practical to analyze the entire energy landscape for a design problem of any reasonable size. For this reason, we defined several simpler properties and used them to compare different methods.

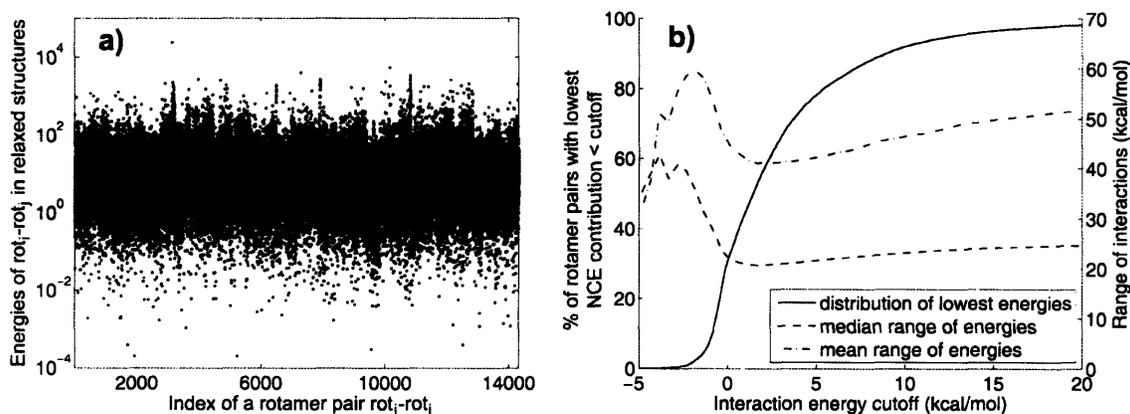


Figure 4-8: The contribution of a given rotamer pair to the NCE of a structure strongly depends on the surrounding structural context. Rotamer pairs (rot_i-rot_j) were considered in 10 different structural contexts, where the rotameric states of surrounding side chains were randomized. Each of these 10 structures was subjected to side-chain minimization (see Methods) and the van der Waals interaction energy between rot_i-rot_j in the minimized structure was recorded. In 4-8a each column (with a particular interaction index) corresponds to a given rotamer pair and the y-axis denotes interaction energies of this pair in the various considered structures. For each rotamer pair, the lowest encountered interaction energy as well as the range of encountered energies (highest minus lowest) was recorded. In 4-8b, for any given interaction cutoff denoted on the x-axis, the solid line (left y-axis) shows the fraction of rotamer pairs with the lowest encountered interaction below this cutoff. For this set of rotamer pairs, the dashed and the dashed-dotted lines (right y-axis) represent the median and the mean of the interaction ranges, respectively. Rotamer pairs for this analysis were picked from the native-sequence repacked structure set based on a criterion for the existence of a clash in their rotameric conformations (van der Waals energy above 10 kcal/mol). At most 10 rotamer pairs were considered per structural region.

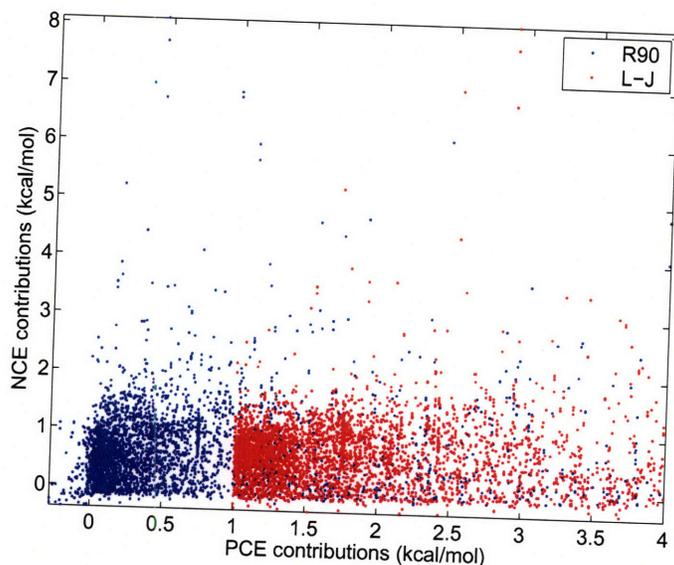


Figure 4-9: Agreement between RCE and NCE contributions for clashing atomic interactions using either R90 or L-J. Clashing interactions here are defined as those that are scored above 1 kcal/mol by L-J in the rotameric representation. Structures with native sequences repacked with L-J were considered for this analysis. 12% of the points have RCE contributions above 4 kcal/mol and are not shown.

One desirable property of a modification is that structures obtained by optimizing the RCE have an RCE in close agreement with their NCE. We tested this in the context of both native-sequence repacking (Figure 4-4a) and sequence design (Figure 4-4b). In both cases, modification LR_{90}^A performs best on average, as well as in the median sense. Striking differences between performance in repacking vs. design are highlighted by modification PRM, which performs very well in repacking but poorly in design. Because rotamers and rotamer pairs are pre-minimized in the presence of only the template, PRM has the potential to over-pack the core of a protein. Although there is little opportunity to over-pack the native sequence of a protein on its own backbone, this problem does show up in sequence design. Figure 4-3f illustrates an example where low-RCE structures have high NCE. Notably, this is the worst type of failure in protein design, because of the time and resource commitment associated with testing sequences experimentally. The over-packing problem for sequence design is also observed with the radius-scaling modifications; AAD decreases with decreasing radius size for repacking, whereas it has a minimum at R90 in sequence design. In-

terestingly, even though LR_{90}^A also has the potential to be too soft, apparently the 10 kcal/mol penalty is sufficient to avoid the over-packing problem and this modification performs well on both tests.

Quantitative agreement of RCE with NCE across sequences and structures, as reported in Figure 4-4 panels a and b, is necessary for some applications. For instance, when several folds are explicitly considered in protein design, e.g. to introduce specificity for one of them, the energies on the different backbones should be on the same scale. On the other hand, such strict agreement across structural environments is not always required. For example, it is irrelevant when predicting structures by side-chain repacking or when evaluating the relative stabilities of different sequences on the same backbone. The within-region RCE-to-NCE correlation, which is the key metric for this application, is shown in Figure 4-4, panels c and d. In each case, the within-region performance trends are the same as the cross-region trends. For sequence design, however, the AAD and MAD values are lower when the structures being compared share the same backbone (compare 4-4d to 4-4b). This is likely due to the presence of hard-to-resolve clashes in certain structures, i.e. groups of sites for which there are no or few clash-free rotamer combinations. Such region-specific clashes would contribute a roughly constant offset to RCE values that would not be penalized in within-region tests.

For yet another set of applications, the correct ordering of sequences by energy within a given structural region, rather than quantitative RCE-to-NCE agreement, may be sufficient. This is the relevant requirement for a hierarchical approach to protein design in which a large pool of candidate sequences is generated using a fast energy function (in our case RCE), and potentials of increasing accuracy and complexity (here NCE) are applied to filter the results. We used two metrics to interrogate the degree of RCE-to-NCE re-ranking. Figure 4-4, panels e-f, show the frequency and depth of re-ranking for RCE-optimized structures in native-sequence repacking and for low-RCE sequences in design. Frequency of re-ranking behaves similarly in the two cases, with R95 and LR_{90}^A showing the best performance. The behavior of re-ranking depth, however, is strikingly different between design and repacking, and

it roughly correlates with the respective within-region MAD values. Starting at R80, re-ranking depth monotonically increases with radius size for native sequence repacking. Larger radii cause more clashes and make it more likely for a high-RCE solution to end up with the lowest NCE. This trend is not seen in sequence design, where there are fewer clashes because there is more opportunity to resolve them by changing the sequence. Interestingly, capping van der Waals interactions (modification LR_{90}^A) gives a dramatic improvement over R90 for repacking, causing LR_{90}^A to have the lowest re-ranking depth for repacking and sequence design. This indicates that in addition to getting rid of structures with unrealistically large RCE, LR_{90}^A also improves the agreement between the RCE and the NCE orderings of structures - not necessarily an expected effect.

The metrics discussed so far test whether RCE and NCE-based energy landscapes are close around RCE minima. Good agreement at these minima does not necessarily mean that the two landscapes are close at other points, however, or that there is rough overall agreement between them. Indeed, differences in RCE-to-NCE agreement for native sequences, which should presumably be scored well by a reasonable design procedure (Figure 4-4a), and sequences that are selected in design (Figure 4-4b) indicate that there is a qualitative difference between these. For this reason, it is important to test RCE-to-NCE agreement not only for low-RCE sequences, but also for high-RCE ones, as it is possible that most of appropriate sequences for the given fold (such as those with well-packed cores after minimization) have high RCE. We tested for a more global RCE-to-NCE agreement in two ways. Figure 4-5a shows the average frequency with which one of 100 random sequences had a lower NCE than the best RCE-designed sequence. Strikingly, the value is close to zero only for R90, R95 and LR_{90}^A . For modifications like PRM or R80 it is essentially meaningless to search in RCE space, because it does not take very many attempts to randomly find a better sequence directly in NCE space. Even for R80, ~ 2 sequences out of 100 random ones are as good or better than the best sequence found through extensive optimization of RCE.

Another way to evaluate the global appropriateness of the RCE-based energy

landscape is to look at χ -angle recovery rates in native-sequence repacking. This provides a test of whether the RCE global minimum is similar to the global minimum of the true energy landscape (see Figure 4-1). Although energy terms other than vdW determine the accuracy of structural prediction, when these are constant it is reasonable to compare χ -angle recovery across different vdW modifications. Figure 4-5b shows the results of this analysis. LR_{90}^A has the highest χ -angle recovery rates, closely followed by R95 and R90. Interestingly, most of the improvement of LR_{90}^A comes from better prediction of surface positions, which almost certainly has more to do with modifications of electrostatics than van der Waals. Indeed, in core and boundary positions, where changes in repulsive van der Waals are expected to play the most important role, the performance of LR_{90}^A is roughly the same as that of R90. Therefore, for χ -angle prediction it is best to scale vdW radii by 90-95%, and most of the improvement is not due to the specifics of the repulsive portion of the potential.

An alternative to introducing vdW modifications in protein design is to use larger rotamer libraries. We tested libraries ranging in size from 175 to 521 rotamers in native-sequence repacking, using the L-J and R90 potentials (Figure 4-6a). Notably, even the smallest library used with R90 outperforms the largest library with L-J. This indicates that appropriate vdW energy modifications can be a far more effective way of addressing the problem of RCE-to-NCE disagreement than rotamer library expansion, especially given the computational cost of the latter. Honig and co-workers have shown that χ -angle recovery can be significantly improved by using a very large expanded rotamer library (7,562 rotamers) [192], presumably due to an improved sampling of the true energy landscape. However, here we have shown that expanding rotamer libraries in a size range more practical for protein design does not effectively address the problems associated with the ruggedness of the energy landscape (Figures 4-6a-b).

The problem of predicting NCE from rotameric structures is non-pairwise decomposable, and this imposes a limit on how well a pairwise, rotamer-based approximation can perform. However, it is not clear whether this limit is close to being achieved by the different modifications we considered, or whether further significant improvement

can be expected. To explore this, we looked at the sensitivity of atomic pair and rotamer pair interactions to their surrounding structural environment (Figures 4-7 and 4-8, respectively). Both tests indicate that the problem is severely non-pairwise decomposable. This is due to the extremely important influence of structural context on the extent to which an atom or a rotamer pair interaction can relax. The data in Figure 4-8 indicate that in the absence of contextual information, rotamer interaction energies can only be predicted with an error of > 20 kcal/mol. This suggests that no pairwise-decomposable modification can be expected to “fix” this problem.

Out of all the modifications we tested, those based on scaling the van der Waals radius by 90 or 95% emerge as the clear winners. This is fortunate, as R90 is also the modification that has been used most frequently in protein design. The choice has been justified using a limited set of experimental data [36], and it is interesting to see it borne out in more extensive computational tests. In this work we investigated the basis for R90’s superior performance. As expected, this mostly has to do with the softer treatment of vdW repulsion by R90 compared to the original Lennard-Jones potential. Somewhat surprisingly, however, R90 is no better than L-J (in the sense of correlation) at predicting the eventual NCE contributions of initially clashing interactions. The difference is that all repulsive interactions are scored uniformly lower by R90, which allows for fewer unrealistically large interactions and better agreement with NCE. Indeed, our analysis of rotamer pair interactions in structures repacked with R90 or L-J shows that 60% of repulsive interactions between rotamers resolve upon minimization to yield neutral (0 kcal/mol) or favorable contributions to the NCE, and 82% yield NCE contributions below 0.5 kcal/mol. Thus, it would seem that potentials that treat repulsion softly should perform very well. However, this is not strictly true because eventually, if vdW repulsion is treated too permissively (as is the case with PRM), sequences with unresolvable clashes are selected in protein design. Scaling van der Waals radii by $\sim 90\%$ seems to be the optimum between these two competing extremes. In many tests, modification LR_{90}^A gives an additional improvement in performance over R90, which is due to the further reduction in the number and magnitude of outliers with unrealistically high RCE.

In the end, the performance of even LR_{90}^A , R90 or R95 is far from perfect. Differences between RCE and NCE in repacking and design are large in comparison with the magnitude of effects normally considered in protein design. Note that to obtain even these deviations in practice, one must find the optimal scale factor for the modified potential. We found that for modifications R90, R95, L-J, LR_{90}^A and PRM the scale factors were close to unity (roughly 0.8 – 1.4), but they were significantly different from unity for smaller radii. Additionally, the reordering between random solutions and RCE-optimized solutions is non-zero even for the best modifications (see Figure 4-5a). We have shown that the problem of predicting NCE from rotameric structures is inherently strongly non-pairwise decomposable. All of these results together raise the question of whether the use of approximate potentials in conjunction with global optimization (e.g. Dead End Elimination) is justified, relative to non-optimal searching using the correct energy function. In fact, some developments in the latter direction have already been made. Baker and co-workers have used non-optimal searching to adjust the choice of side-chain rotamers after initial repacking, by allowing each rotamer at each position to relax independently with the rest of the structure held constant [184]. Additionally, we have recently completed a successful design study in which the sequence search was driven by NCE rather than RCE (unpublished results). Thus, although the extraordinary utility of the rotamer approximation for describing protein structure cannot be disputed, exactly how this approximation should best be incorporated into fixed-backbone protein design calculations remains to be determined.

4.6 Acknowledgements

This chapter is based on a manuscript whose authors are Gevorg Grigoryan, Alejandro Ochoa and Amy E. Keating, currently in press with *Proteins: Structure, Function, and Bioinformatics*. We thank the CSBi high-performance computing platform for computer time and support; MIT's Undergraduate Research Opportunities Program (UROP) for facilitating the involvement of A.O. in this study; J. Apgar and B.

Joughin for comments on the manuscript; the Whitaker Health Sciences Fund Fellowship for funding to G.G. This material is based upon work supported by the National Institutes of Health (GM67681) and by equipment purchased under National Science Foundation Grant No. 0216437.

Chapter 5

A Novel Framework for Specificity Design

Computational design of protein-protein interactions has emerged as a promising approach for engineering new cellular reagents and pharmaceuticals. Several studies have successfully designed new protein interfaces and a few have succeeded in engineering proteins that bind native targets. However, to design practically useful reagents, one must pay attention not only to the intended interaction between the design and the target, but also to the specificity of this interaction, which can be important for function. Here we introduce a novel protein design framework, which allows for the incorporation of an arbitrary number of undesired states. This approach produces a map of provably optimal tradeoffs between stability and specificity and leaves it up to the user to select sequences with satisfactory levels of both. We applied this novel framework to the design of specific partners against the leucine zipper domains of human transcription factors from the bZIP family. Dimerization specificity within this family is known to be functionally determining, so avoiding off-pathway interactions is of great practical utility. We have characterized the space of specificity/stability tradeoffs for designs against all human bZIPs and have shown that often designing solely against the target sequence does not produce the desired levels of specificity. We have also found that some bZIP coiled-coil sequences are inherently easier targets for specificity design than others. Finally, we have designed

specific binding partners against a number of human bZIPs, considering all of the other bZIP sequences as undesired competitors, and proposed them for experimental verification.

5.1 Introduction

Over the last decade, computational protein design has emerged as a promising technique for engineering biologically useful reagents, pharmaceuticals and new materials. Among the successes of the field are the stabilization of existing protein scaffolds [117, 35], solubilization of membrane proteins [161], incorporation of new enzymatic activity into a scaffold [46] and the design of a novel fold [95]. Design of specific protein-protein interactions is of particular practical interest, as it potentially allows one to engineer partners against existing cellular players. To be practical, a method for designing binding partners for cellular proteins must not only take into account the strength of the target complex (stability), but also the possible off-target interactions of the designed protein (specificity), as the latter can have significant functional effects in the cell.

Several studies have reengineered protein-protein interactions [68, 90, 20, 4, 156, 147, 135], although few have considered the problem of designing partners against a fixed target [156, 147], and even fewer have done this by explicitly considering off-target interactions. Havranek and Harbury computationally selected dimeric coiled-coil sequences that preferentially formed either homo- or hetero-dimers by varying both monomers and explicitly considering competing states [68]. Mayo and co-workers computationally redesigned calmodulin to improve its binding to one of its native targets and in doing so also increased its specificity for that target [156]. Similarly, Reina *et. al* reengineered the specificity of a PDZ domain by considering only the strength of the target interaction [147]. Kortemme *et. al* used a computational second-site suppressor strategy to engineer a new variant of an existing protein-protein interaction that is orthogonal towards the native pair [90].

The fact that design of specificity or multi-state design is relatively less studied

compared to single-state stability design is probably partly due to the fact that earlier is a more challenging task. Current methods for computational protein design are still very much in development and are not yet at the point where most proposed designs work experimentally. The most successful single-state design studies have had yields around one out of four to ten designs. If success is defined in terms of more than one state (e.g. the design should bind to protein A, but not a related protein B), the probability of success is expected to drop exponentially with the number of alternative states. In practice that means that many proposed designs would have to be tested experimentally, before a successful one is found.

Other reasons for why specificity design is challenging have to do with the nature of the computational methods. A key component in computational protein design is the scoring function that is used to describe the compatibility of any particular sequence with the fold or interaction in question. A variety of scoring functions are used, and although they vary in terms of their physical realism, a significant empirical component is present in all of them [122, 104]. Given this empirical nature of scoring functions, it is difficult to know how to properly weight stability relative to specificity and derive a single quantity to optimize. Additionally, even if a single quantity is formulated, the global optimization techniques that have worked so well for single-state design [42, 56, 58] are usually no longer applicable, making it necessary to employ non-optimal searching techniques. Given that the sequence search space to deal with is roughly 20^n where n is the number of designed positions, this means that one never knows whether the obtained designs are even close to being optimally specific.

In this study we introduce a new framework for specificity design that allows for the treatment of an arbitrary number of negative states and does not rely on formulating specificity as a single expression. Instead, it systematically explores the space of tradeoffs between specificity and stability in a manner that is easy to analyze. This allows the user to make the final choice of a sequence that likely to be both stable and specific enough from a relatively short list of candidates. The idea behind the framework is diagrammatically shown in Figure 5-1. Here the target state is

designated T and there are four competing states N_1 through N_4 . Initially, if we select a sequence to only optimize the stability of the target state, it may or may not have favorable energies in undesired states. If not, the problem is solved and the specificity was obtained “for free”. However, as is the case with N_1 in figure 5-1 (left panel), some undesired states may turn out to be significant competitors. In this case, some stability in the target state has to be traded to obtain more specificity. In subsequent panels of figure 5-1 the stability in the target state is optimized under a progressively increasing constraint on the gap between the target state and the most stable of the negative states. Eventually this leads to solutions where the target state is relatively more stable than any of the undesired states, albeit some stability of T is lost. Finally, a situation arises where no sequence exist that produces a larger gap between the target and undesired states. We call this procedure a specificity sweep.

Several theoretical insights were necessary to make such a procedure possible. First, we drew from our earlier work on cluster expansions in protein design (see chapter 3 and ref [61]) to express a structure-based energy function as a simple function of sequence, thereby tremendously simplifying the sequence optimization tasks. Also, we formulated the problem of optimizing the energy of the target state under a set of gap constraints as an integer linear program in a way similar to the one used by Singh and co-workers [85].

We have applied this novel framework to design specific partners against coiled-coil regions of human transcription factors from the bZIP family. Dimerization specificity among bZIP proteins is known to determine function in many cases [174, 63], and so avoiding possible off-target interactions is particularly important. The problem is exacerbated further by the significant sequence conservation within the bZIP family, making it difficult to discriminate between competitors. We have designed specific partners against a number of bZIP coiled coils by considering interaction with all non-target bZIPs, as well as homodimerization of the design itself, as undesired states. We have also performed a global computational analysis of the bZIP interactome showing that some sequences are inherently easier targets for specificity design than others.

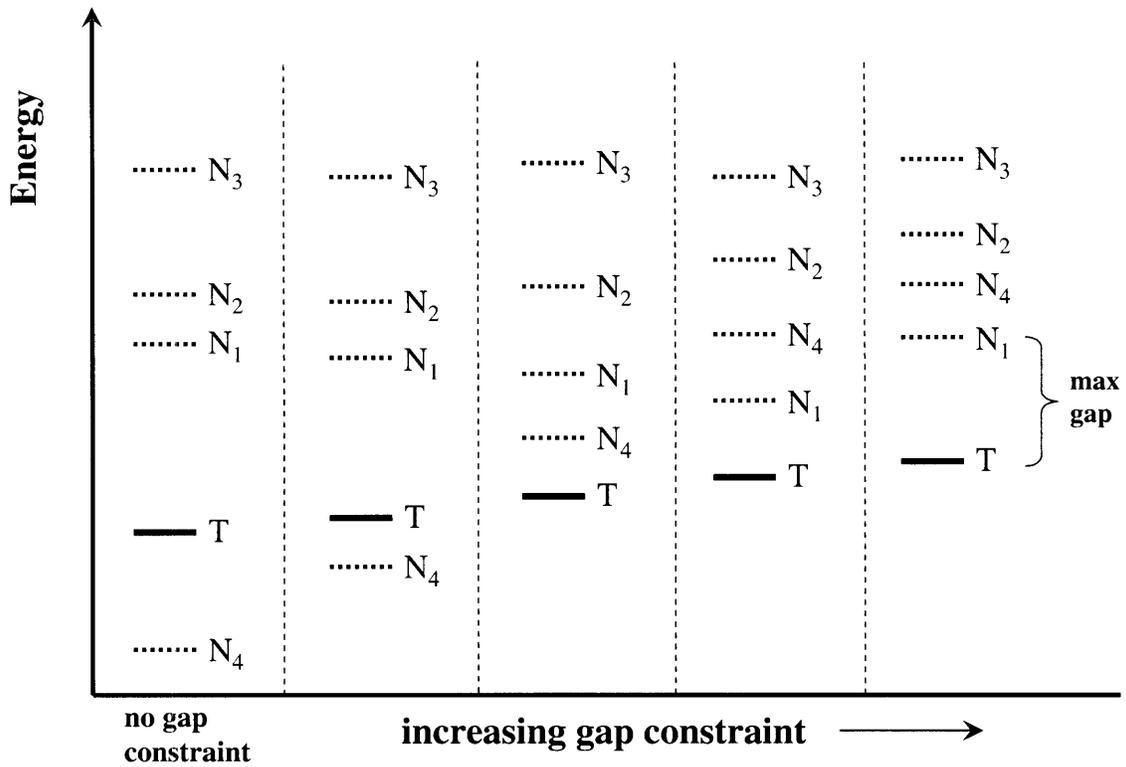


Figure 5-1: Illustration of the specificity sweep procedure. T designates the target state and N_1 through N_4 are undesired states. First, the energy of the target state alone is minimized (left panel), which results in N_4 being more favorable than the target state. Subsequently, an increasing constraint is placed on the gap between T and the most stable undesired state (middle panels). Eventually, a situation arises where the gap can no longer be increased (right-most panel).

5.2 Results and Discussion

The goal of our framework is to allow for optimization of the target-state energy while arbitrarily constraining the energies of undesired states. In computational protein design, the energy of a sequence in a given structural state is normally defined as the conformational energy of that sequence minimized over side-chain rotamer degrees of freedom: $E^{\min}(\vec{\sigma}) = \min_{\vec{r}} \{E^c(\vec{\sigma}, \vec{r})\}$, where $E^c(\vec{\sigma}, \vec{r})$ is the conformational energy for of sequence $\vec{\sigma} = \{\sigma^1, \dots, \sigma^N\}$ with rotamer configurations \vec{r} . Efficient algorithms exist for finding this optimal set of side-chain rotamers and the corresponding minimal energy for a given sequence [42, 56, 58, 101, 105, 143]. However, using these algorithms we can only define the energy of a sequence numerically, which makes it very difficult to perform constrained optimizations in sequence space mentioned above. One way to circumvent this problem is to express sequence energy analytically, rather than numerically. We have previously shown how excellent analytical approximations to $E^{\min}(\vec{\sigma})$ can be obtained using the approach of cluster expansion (CE) [61]. Figure 5-2 shows the agreement between structure-based energies according to model HP/S/C (developed in chapter 2) and corresponding sequence-based approximations (see 5.4.1 for details). The agreement between the two is within 2.2 kcal/mol, which is quite good given the range of energies predicted by model HP/S/C for natural bZIP dimers [60]. Once $E^{\min}(\vec{\sigma})$ for each state is analytically expressed as a function of sequence, the problem of optimizing the target state energy while constraining the energies of undesired states can be formulated as an integer linear program and solved exactly (see section 5.4.2).

5.2.1 Designs that Optimize Stability Hit Off-target Partners

In some instances in the literature, interaction specificity has been obtained without explicitly considering undesired states and instead just improving the stability of the target complex [156, 147, 194]. However, clearly such a strategy can not be expected to work in all situations. In particular, its success depends on the degree of similarity between the target and the competing partners as well as whether the design against

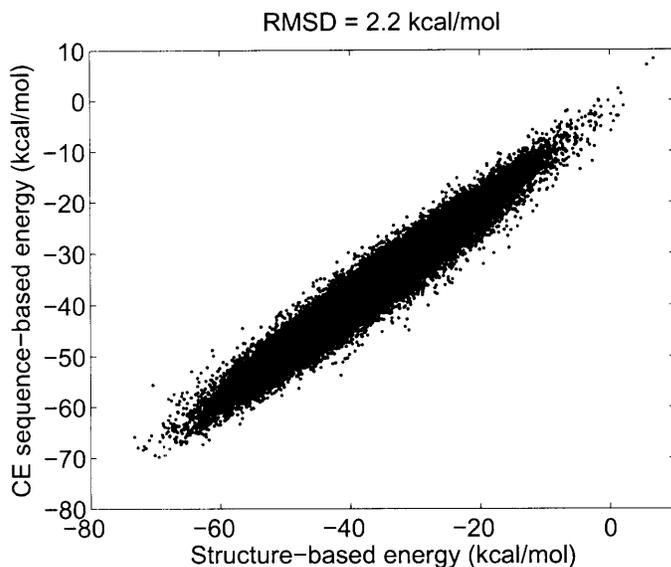
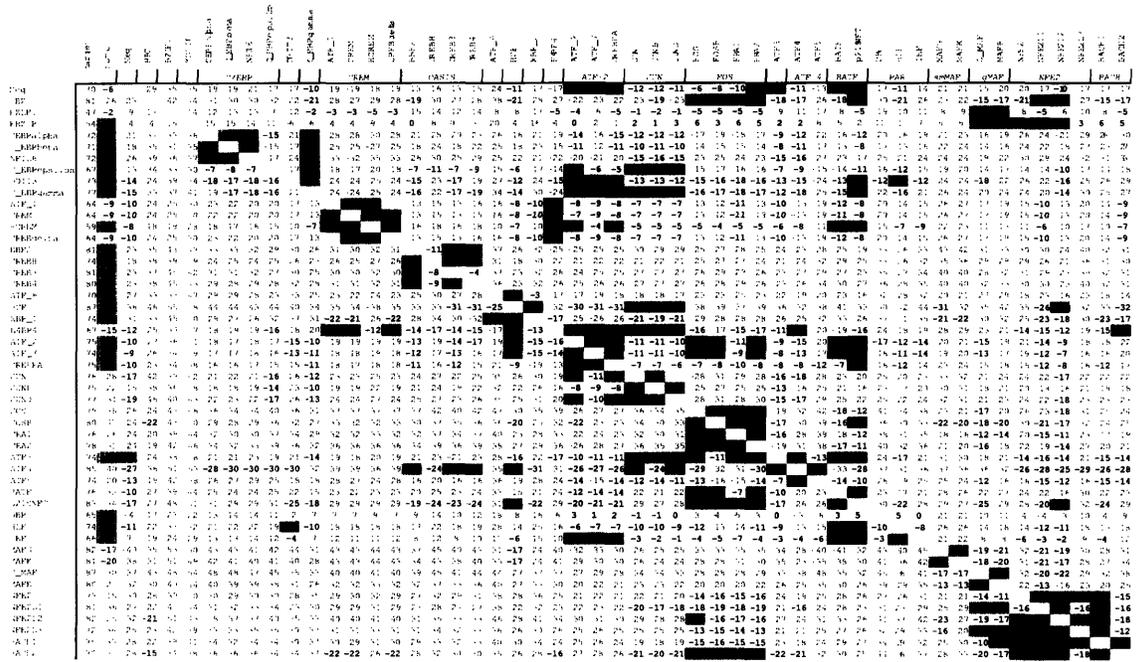
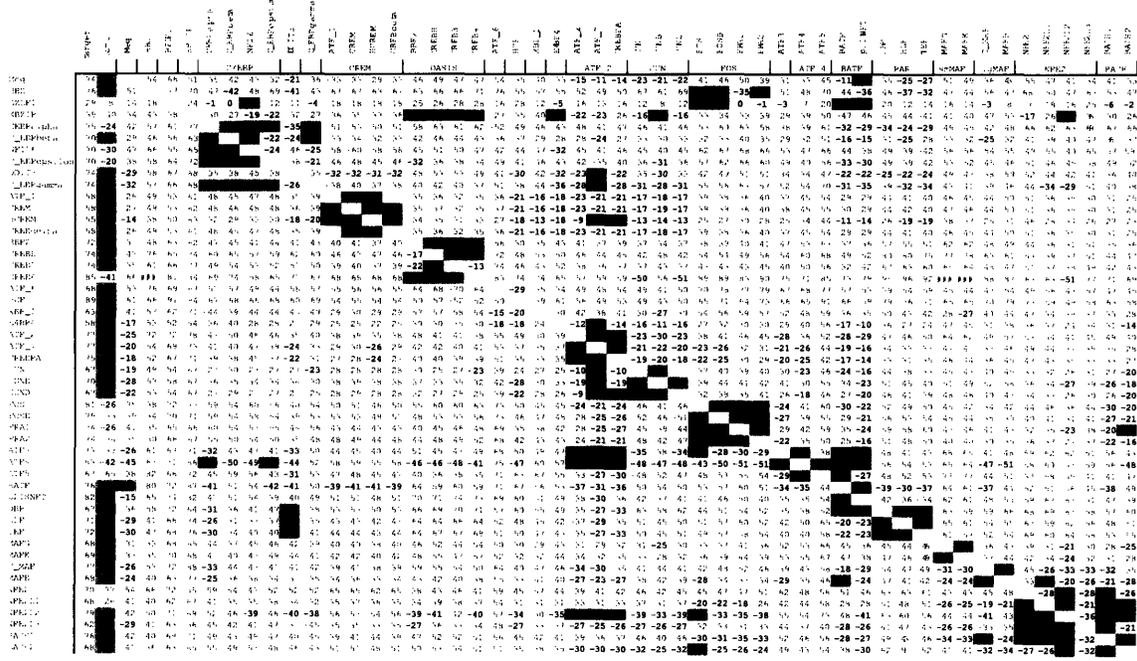


Figure 5-2: Agreement between structure-based energies explicitly calculated with HP/S/C and a CE sequence-based approximation.

the target alone happens to exploit features that are similar between the target and the undesired competitors. We are interested in designing specific coiled-coil probes against members of the human bZIP transcription factor family. Because there is significant sequence conservation within the coiled-coil region of these proteins, we expected that specific designs would be unlikely to originate from optimizing solely against the target protein. We tested this hypothesis computationally. For each human bZIP coiled-coil, we found the sequence of the optimal binding partner according to two different models, and asked whether that sequence scored well against other bZIPs. One of the models, referred to as HP/S/C (described in chapter 2), was shown to reproduce well experimentally observed bZIP coiled-coil interaction preferences [60]. We also considered the sequence-based scoring method for predicting parallel dimeric coiled-coil interactions developed by Singh and co-workers [52]. This model, referred herein as model ML, is based on summing pairwise contributions from amino acid located at seven different pairs of positions ($\mathbf{a} - \mathbf{a}'$, $\mathbf{d} - \mathbf{d}'$, $\mathbf{g} - \mathbf{e}'$, $\mathbf{g} - \mathbf{a}'$, $\mathbf{d} - \mathbf{e}'$, $\mathbf{a} - \mathbf{d}'$, $\mathbf{d} - \mathbf{a}'$) with these contributions derived via a machine learning method and a database of coiled-coil as well as non coiled-coil partners [52].



(a) HP/S/C



(b) ML

Figure 5-3: Results of optimizing binding partners against each human bZIP coiled coil (left panel in figure 5-1). The first column shows the score against the target sequence. Subsequent columns contain gaps between the target state heterodimer and dimers between the design and other bZIPs. Negative gaps indicate that the target scores more favorably. Gaps in each row are colored in light gray and dark gray if they are in the top 40% and 20% of the gap range observed for that target, respectively. Panels (a) and (b) to models HP/S/C and ML respectively. For the latter, scores are negated so that lower scores correspond to higher stabilities, in analogy to binding energies.

The results of optimizing only against the target bZIP (left panel of figure 5-1) are shown in figure 5-3(a) for model HP/S/C. In most instances, just by virtue of optimizing binding with the target bZIP, interactions with other bZIPs are scored weaker. However, for almost all bZIP targets, there is at least one competitor that either produces a positive gap (i.e. the score of the design against the competitor is more favorable than against the target) or a small negative gap. In particular, members of the same sequence family (diagonal blocks) are almost always problematic competitors. Additionally, the homodimer of the design is often stabilized when only the target heterodimer state is considered (darker squares in figure 5-3(a) show the most problematic competitors). There can also be off-family competitors, i.e. sequences not in the same family as the target bZIP that nevertheless are expected to interact well with the design. For instance, according to model HP/S/C, when designing against targets in the CREM family, interactions are also likely with proteins from Jun family and designs against C/EBP ϵ are predicted to interact with sequences from both ATF-2 and Jun families. Interestingly, these combinations are almost never reciprocal (e.g. CREM family members are not expected to compete with design against the Jun family).

Figure 5-3(b) shows the result of optimizing binding with model ML. In this case, the problem of designing specificity does not seem to be nearly as difficult. For the possible exception of competitors from the same sequence family as the target, the magnitudes of the observed gaps are higher with ML, relative to the scores of the target state, than for model HP/S/C. This, however, can be an artifact of the scoring function. When tested against experimentally observed dimerization preferences of human bZIP coiled coils, scores above ~ 35 for the ML model corresponded to strong interactions [52]. So it may be that scores of 70 and 40 correspond to roughly the same interaction strength, meaning that many of the gaps observed in figure 5-3(b) are not actually as large as they appear.

5.2.2 Designing Stability and Specificity

As indicated in figure 5-3 specificity is likely to be a problem if designs are only optimized against their target partners. Using the methodology we have developed (see section 5.4.2), we can optimize binding with the target sequence under arbitrary constraints on the gap between the target dimer and the undesired dimers (see figure 5-1). Here we perform specificity sweeps of all human bZIP coiled coils, while considering the design homo-dimer as well as hetero-dimers with all other bZIPs, except members of the target's family, as undesired states. We have seen that binding to members of the same sequence family as the target is difficult to avoid due to high sequence conservation. Indeed, experimental binding profiles of family members are usually very similar [134]. At the same time, members of a family are expected to have similar functions, so from the practical standpoint, absence of specificity within a family may not be a problem. For this reason, in our further analysis we exclude members of the sequence family of the target bZIP from the list of undesired competitors.

Shown in figure 5-4 is a series of graphs summarizing the results of the specificity sweeps using model HP/S/C. Each specificity sweep (one for each bZIP target) results in a list of optimized sequences of decreasing stability and increasing minimal gap with any of the undesired states. Figure 5-4(a) shows the gaps for the sequences optimized for binding to their targets without any constraints. In figure 5-4(b) gaps are shown for the sequence out of the specificity sweep list that loses at most 5% of the score relative to the optimal sequence. In figure 5-4(c) up to 20% of stability is allowed to be lost. Finally, figure 5-4(d) shows the sequence with the largest possible minimum gap. Clearly, significant specificity can be gained by allowing for the loss of some stability (dark and light grey boxes indicate gaps above -6 and -13 kcal/mol). In fact, the most dramatic difference is between the most stabilizing sequence (figure 5-4(a)) and one that is allowed to be at most 5% less stable (figure 5-4(b)). In this interval of stability, many of the designs gain gaps of ~ 10 kcal/mol against most competitors. Based on previous tests of the energy model, a gap of this size indicates a high degree of confidence in the relative order of stability. The marginal improvement is less

between loosing up to 5% and up to 20% of the original stability (compare figures 5-4(b) and 5-4(c)) and by the time 20% of stability is lost, most designs already have the largest possible minimal gaps (compare figures 5-4(c) and 5-4(d)).

5.2.3 Proposed Designs for Experimental Testing

To test our specificity design framework, as well to further test model HP/S/C, we selected designs against several targets for experimental characterization. We selected targets from a variety of bZIP families that together span a large portion of native bZIP sequence space [8], while at the same time considering reagent availability. Members of families ATF-2, ATF-3 (a family with one sequence), Fos, Jun, L-Maf and NFE2 were considered for target selection. Additionally, we selected a viral protein, Meq, as one of the targets. Meq is a bZIP from Marek's disease virus (MDV) – a very oncogenic herpesvirus that induces T cell lymphomas in chickens [110]. Because Meq is believed to be the oncoprotein of MDV [108], peptides that target Meq in a specific manner may be of practical use.

Figure 5-5 shows examples of specificity sweeps against several targets. Design against ATF-2 is shown in figure 5-5(a). As we can see, simply optimizing binding with ATF-2 causes p21SNFT (from the BATF family) to score nearly as well against the design as against ATF-2 itself, producing a gap of only -2.6 kcal/mol. From earlier characterization of model HP/S/C, we know that a score difference of ~ 13 kcal/mol corresponds to a very high level of confidence in the order of interaction strength, so ideally we would like to see gaps of around that magnitude. The design homodimer, HCF and C/EBP $_{\gamma}$ are also close competitors. As a progressively larger gap constraint is placed on the design, there is initially very little change in core residue identities, and specificity is mostly addressed with **e** and **g** position mutations (amino-acid changes are indicated with blue squares). Eventually, however, when that strategy saturates, core residues begin to change and this is also the point where stability drops more sharply. Two designs were selected against ATF-2 (marked with asterisks in figure 5-5(a)), both of which are above the point where stability is significantly compromised. The first design is expected to be slightly more promiscuous than the

		f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g														
ATF-2		V	W	V	Q	S	L	E	K	K	A	E	D	L	S	S	L	N	G	Q	L	Q	S	E	V	T	L	L	R	N	E	V	A	Q	L	K	Q	L	L	L	A	H	K	D	C
p21SNFT	-2.6	A	K	L	A	A	L	R	A	L	L	A	A	L	E	A	R	N	A	A	L	K	A	L	R	A	A	L	K	A	L	L	A	A	L	R	A	E	L	A	A	L	K	A	E
homo	-3.9	A	K	L	A	A	L	R	A	L	L	A	A	L	E	A	R	N	A	A	L	K	A	L	R	A	A	L	K	A	L	L	A	A	L	E	A	E	L	A	A	L	K	A	E
p21SNFT	-5.1	Q	K	A	D	K	L	H	E	E	Y	E	S	L	E	Q	E	N	T	M	L	R	R	E	I	G	K	L	T	E	E	L	K	H	L	T	E	A	L	K	E	H	E	K	M
*	-74	A	K	L	A	A	L	R	A	L	L	A	A	L	E	A	R	N	A	A	L	K	A	L	R	A	A	L	K	A	L	L	A	A	L	E	A	E	L	A	A	L	E	A	E
HCF	-6.1	E	Y	V	M	G	L	E	S	R	V	R	G	L	A	A	E	N	Q	E	L	R	A	E	N	R	E	L	G	K	R	V	Q	A	L	Q	E	E	S	R	Y	L	R	A	V
C/EBPg	-7.2	Q	K	A	Q	D	T	L	Q	R	V	N	Q	L	K	E	E	N	E	R	L	E	A	K	I	K	L	T	K	E	L	S	V	L	K	D	F	L	E	H	A	H	N		
homo	-8.3	A	K	L	A	A	L	R	A	I	L	A	A	L	E	A	A	N	A	A	L	K	A	L	R	A	A	L	K	A	L	L	A	A	L	E	A	E	L	A	A	M	K	A	E
*	-73	A	K	L	A	A	L	R	A	I	L	A	A	L	E	A	R	N	A	A	L	K	A	L	R	A	A	L	K	A	L	L	A	A	L	E	A	E	L	A	A	H	E	A	E
HCF	-9.4	E	Y	V	M	G	L	E	S	R	V	R	G	L	A	A	E	N	Q	E	L	R	A	E	N	R	E	L	G	K	R	V	Q	A	L	Q	E	E	S	R	Y	L	R	A	V
p21SNFT	-11	Q	K	A	D	K	L	H	E	E	Y	E	S	L	E	Q	E	N	T	M	L	R	R	E	I	G	K	L	T	E	E	L	K	H	L	T	E	A	L	K	E	H	E	K	M
JUND	-71	A	K	L	A	A	L	R	A	I	L	A	A	L	E	A	R	N	A	A	L	K	A	L	R	A	A	L	K	A	L	L	A	A	L	E	A	E	L	A	A	H	K	A	E
Meq	-13	D	Y	V	D	K	L	H	E	A	C	E	E	L	Q	R	A	N	E	H	L	R	K	E	I	R	D	L	R	T	E	C	T	S	L	R	V	Q	L	A	R	H	E		

(a) ATF-2

		f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e														
FOS		E	L	T	D	T	L	Q	A	E	T	D	Q	L	E	D	E	K	S	A	L	Q	T	E	I	A	N	L	L	K	E	K	E	K	L	E	F	I	L	A	A	H	R
p21SNFT	-79	A	E	T	A	A	L	K	A	K	K	A	A	L	R	A	R	I	A	A	L	K	A	K	R	A	A	L	K	A	K	I	A	A	L	R	A	E	I	A	A	Y	K
p21SNFT	-78	A	L	T	A	A	L	K	A	K	K	A	A	L	R	A	I	I	A	A	L	R	A	I	K	A	A	L	K	A	K	I	A	A	L	R	A	E	I	A	A	L	K
*	-78	A	E	K	A	A	L	K	A	K	K	A	A	L	R	A	R	I	A	A	L	K	A	K	R	A	A	L	K	A	K	I	A	A	L	R	A	E	I	A	A	Y	K
NFE2L2	-15	E	N	I	V	E	L	E	Q	D	L	D	H	L	K	D	E	K	E	K	L	L	K	E	K	G	E	N	D	K	S	L	H	L	L	K	K	Q	L	S	T	L	Y
p21SNFT	-77	A	E	T	A	A	L	K	A	I	K	A	A	L	R	A	K	I	A	A	L	K	A	K	R	A	A	L	K	A	K	I	A	A	L	K	A	E	L	A	A	Y	K
NFE2L2	-17	Q	K	A	D	K	L	H	E	E	Y	E	S	L	E	Q	E	N	T	M	L	R	R	E	I	G	K	L	T	E	E	L	K	H	L	T	E	A	L	K	E	H	E
ATF-2	-74	A	Q	K	A	A	L	K	A	K	K	A	A	L	R	A	K	I	A	A	L	K	A	K	R	A	A	L	K	A	K	N	A	A	L	R	A	E	L	A	A	Y	K
C_MAF	-19	Q	Q	R	H	V	L	E	S	E	K	N	Q	L	L	Q	V	D	H	L	K	Q	E	I	S	R	L	V	R	E	R	D	A	Y	K	E	K	Y	E	K	L	V	

(b) Fos

		f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e														
Meq		D	Y	V	D	K	L	H	E	A	C	E	E	L	Q	R	A	N	E	H	L	R	K	E	I	R	D	L	R	T	E	C	T	S	L	R	V	Q	L	A	R	H	E
p21SNFT	1.62	A	L	L	A	A	L	R	A	T	V	A	A	L	E	A	E	N	A	A	L	L	A	E	R	A	A	L	K	A	E	K	A	A	L	R	A	K	L	A	A	Y	K
p21SNFT	-70	A	L	L	A	A	L	R	A	T	K	A	A	L	E	A	E	N	A	A	L	L	A	E	R	A	A	L	K	A	E	K	A	A	L	R	A	K	L	A	A	Y	K
p21SNFT	0.48	Q	K	A	D	K	L	H	E	E	Y	E	S	L	E	Q	E	N	T	M	L	R	R	E	I	G	K	L	T	E	E	L	K	H	L	T	E	A	L	K	E	H	E
p21SNFT	-70	A	L	L	A	A	L	R	A	T	V	A	A	L	E	A	E	N	A	A	L	L	A	E	R	A	A	L	K	A	E	K	A	A	L	R	A	K	L	A	A	Y	K
p21SNFT	-1.2	Q	K	A	D	K	L	H	E	E	Y	E	S	L	E	Q	E	N	T	M	L	R	R	E	I	G	K	L	T	E	E	L	K	H	L	T	E	A	L	K	E	H	E
*	-70	A	L	L	A	A	L	R	A	T	A	A	A	L	E	A	E	N	A	A	L	L	A	E	R	A	A	L	K	A	E	K	A	A	L	L	A	K	L	A	A	Y	K
ATF-2	-2.7	Q	K	A	D	K	L	H	E	E	Y	E	S	L	E	Q	E	N	T	M	L	R	R	E	I	G	K	L	T	E	E	L	K	H	L	T	E	A	L	K	E	H	E
p21SNFT	-3.8	V	W	V	Q	S	L	E	K	K	A	E	D	L	S	S	L	N	G	Q	L	Q	S	E	V	T	L	L	R	N	E	V	A	Q	L	K	Q	L	L	L	A	H	K
p21SNFT	-69	A	L	R	A	A	L	E	A	T	K	A	A	L	E	A	E	N	A	A	L	L	A	E	R	A	A	L	K	A	E	K	A	A	L	L	A	K	I	A	A	F	K
p21SNFT	-4.9	Q	K	A	D	K	L	H	E	E	Y	E	S	L	E	Q	E	N	T	M	L	R	R	E	I	G	K	L	T	E	E	L	K	H	L	T	E	A	L	K	E	H	E
p21SNFT	-69	A	L	R	A	A	L	E	A	T	K	A	A	L	E	A	E	N	A	A	L	E	A	E	R	A	A	L	L	A	D	K	A	A	L	K	A	K	L	A	A	Y	K
CREB3	-5.9	Q	K	A	D	K	L	H	E	E	Y	E	S	L	E	Q	E	N	T	M	L	R	R	E	I	G	K	L	T	E	E	L	K	H	L	T	E	A	L	K	E	H	E
ATF-3	-67	A	L	L	A	A	L	R	A	D	V	A	A	L	E	A	E	N	A	A	L	E	A	E	R	A	A	L	K	A	D	N	A	A	L	L	A	K	I	A	A	Y	K
ATF-3	-7.1	V	Y	V	G	G	L	E	S	R	V	L	K	Y	T	A	Q	N	M	E	L	Q	N	K	V	Q	L	L	E	E	Q	N	L	S	L	L	D	Q	L	R	K	L	Q
ATF-3	-66	A	L	R	A	A	L	L	A	D	V	A	A	L	E	A	E	N	A	A	L	L	A	E	R	A	A	L	R	A	E	N	A	A	L	R	A	K	I	A	A	K	K
ATF-3	-8.1	E	K	T	E	C	L	Q	K	E	S	E	K	L	E	S	V	N	A	E	L	K	A	Q	I	E	E	L	K	N	E	K	Q	H	L	I	Y	M	L	N	L	H	R
ATF-3	-63	A	E	R	A	A	L	E	A	D	T	A	A	L	E	A	E	N	A	A	L	L	A	E	R	A	A	L	R	A	D	N	A	A	L	R	A	K	K	A	A	K	K
ATF-3	-9.1	E	K	T	E	C	L	Q	K	E	S	E	K	L	E	S	V	N	A	E	L	K	A	Q	I	E	E	L	K	N	E	K	Q	H	L	I	Y	M	L	N	L	H	R

(c) Meq

Figure 5-5: Specificity sweeps against several targets with model HP/S/C. The sequence in the top line of each panel is that of the target. Sequences highlighted in color are different designs, ranked from the one with most stability (second line) to the one with the most specificity (next to last line). The interaction score of the target heterodimer is shown to the left of each designed sequence. Coloring of the design sequences indicates heptad position. Orange, yellow, light grey and dark grey correspond to **g**, **e**, **a** and **d**. Amino acids that change in each round of the specificity sweep are indicated with blue squares. Underneath each design sequence, the competing sequence with the smallest gap to the target dimer (the most problematic competitor) is shown and the gap itself is shown to the left (negative gaps indicate that the undesired heterodimer scores less favorably than the target). Sequences marked with an asterisk in the left column have been selected for experimental characterization.

second, with minimal gaps of -6.1 and -9.4 kcal/mol, respectively.

Figure 5-5(b) illustrates the specificity sweep against Fos. In this case, optimization against the target alone produces significant gaps with other bZIPs (minimal gap of -12 kcal/mol) and these gaps can be widened somewhat with little loss in stability. Because Fos has some non-canonical residues at **a** positions (threonines in the first two heptads and lysines in the third and fifth), optimal designs against Fos also do not have the canonical all-hydrophobic selection at **a**. This heterogeneous core is much of the reason why high specificity for Fos is obtained. One design was selected against Fos (marked with an asterisk in 5-5(b)), which had a minimal gap of -15 kcal/mol and was predicted to be nearly as stable as the top design. This design has two lysines and one arginine at **a** positions. The two lysines (the first two **a** positions) are across from threonines at **a'** positions in the other helix. Interestingly, K is the only residue for which a significantly favorable thermodynamic coupling energy is reported with T at this position (-0.45 kcal/mol) [2]. One of the lysines is in the N-terminal **a** position, which will minimize its desolvation upon folding. Finally, the second lysine is poised to make a salt bridge with a glutamate at the opposing **g** position in Fos. The arginine selected in the fourth **a** position can form a similar salt bridging interaction.

Results of the specificity sweep against Meq are shown in figure 5-5(c). This is an example where specificity design is difficult. The top binding partner against Meq scores better against p21SNFT by 1.6 kcal/mol than against Meq itself. In addition to

this, only six out of the 12 **e** and **g** position amino acids in Meq are charged, and four of these six are identical between Meq and p21NSFT. This makes it difficult to impose specificity in the usual manner, i.e. with charge patterning. Instead, design relies on the core to deal with specificity, selecting many of the core position residues to be polar or charged. The most specific design has a minimal gap of -9.1 kcal/mol, but by this point the stability has dropped significantly and all of the **a** position residues are polar. We chose one design against Meq at a point where further improvements in stability caused the core to be much more polar (marker with an asterisk in figure 5-5(c)). This design is expected to be somewhat promiscuous and will probably interact at least with the ATF-2 and BATF families. However, calculations still predict that it should interact with Meq better than any other bZIP.

Upon choosing designs against all targets, **b**, **c** and **f** position amino acids were chosen as described in section 5.4.4. Model HP/S/C does not directly account for **b**, **c** and **f** residues, and in general less is known about the impact of these positions on coiled-coil stability and specificity. Hence, our goal was to choose amino acids at these positions that are most appropriate given what was already chosen at **a**, **d**, **e** and **g**, according to naturally observed distributions. Additionally, through experimental characterization of some initial designs we, discovered that large values of charge or helix propensity can be problematic (high charge causes significant salt effects, and a high helix propensity indiscriminately stabilizes all interactions). Therefore, in our procedure for choosing **b**, **c** and **f** amino acids, we imposed constraints on the values of charge, charge content and helix propensity of the entire final sequence. Table 5.1 shows the list of the final design sequences proposed for experimental characterization. These will be characterized using both protein microarrays to assay global specificity and circular dichroism to investigate select complexes more quantitatively.

5.3 Conclusions

We have presented a new powerful technique for simultaneously designing specificity and stability. This approach is not based on formulating a specific tradeoff function for

Table 5.1: Final sequences for experimental characterization.

register	fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg
antiATF-2	QKLQTLRDLLAVLENRNQELKQLRQHLKDLLKYLEDELATLEKE
antiATF-3-1	NEDLVLENRLAALRNENAALNDLARLEKEIAYLEKEIEREK
antiATF-3-2	ELTDELKNKKEALRKDNAALLNELASLENEIANLEKEIAYFK
antiATF-3-3	NETEQLINKKEQLKNDNAALEKDAASLEKEIANLEKEIAYFK
antiFos	NEKEELKSKKAELRNRIEQLKQKREQLKQKIANLRKEIEAYK
antiJUN	SIAATLENDLARLENENARLEKDIANLERDLAKLEREEAYF
antiMAFG-1	KEIEYLEKEIERLKDRLREHLKQDNAAHRQELNLRLEEAKLEFILAHLLST
antiMAFG-2	KEIERLEKEIKTLINLLTTLRQDNAAHRKEAAALEKEEANLERDIQNLLRY
antiMeq-1	NLLATLRSTAAVLENENHVLEKEKEKLRKEKEQLLNKLEAYK
antiMeq-2	NEVAQLENDVAVIENENAYLEKEIARLRKEIAALRDRLAHKK
antiMeq-3	NEVTTLENDAAFIENENAYLEKEIARLRKEKAALRNRLAHKK
antiNFE2	QKRQQLKQKLAALRRDIENLQDEIAYKEDEIANLKDKIEQLLS

stability and specificity, as this can be difficult in practice given the empirical nature of scoring functions used for computational protein design. The method systematically maps out the space of optimal specificity/stability tradeoffs and leaves the decision of selecting final sequences for synthesis up to the user (although this last step can be easily automated if it is felt that the scoring methods are reliable enough). Although we have chosen to use a specificity sweep strategy (i.e. where the stability of the target state is optimized under an increasing constraint on the smallest gap between the target and the undesired states), many different optimization protocols can be envisioned under the same general ILP approach. In particular, the target state energy can be optimized under a constraint involving any linear combination of gaps with the undesired states. Alternatively, the constraint can be placed on the stability and the gaps can be optimized. Finally, although we have applied our framework to only sequence-based pairwise-decomposable energy functions, it is easy to envision how higher-order terms can be accommodated within the same framework (see section 5.4.2 for a brief elaboration).

Based on our calculations, specificity design is certainly something that needs to be considered explicitly if one hopes to design practical peptide binders against bZIP targets. Sequences that globally optimize target state stability often fail to be predicted to bind better to their intended target than to any of the competitors.

Although our calculations only apply to the bZIP system, intuition dictates that interaction specificity will in general not be obtained “for free” by simply optimizing binding against the target, especially in instances when potential competitors form a family with the target protein. This is because if only information about the target is known during optimization, it is impossible to know whether the design is taking advantage of those features of the target that are conserved within the family or not. Given this, we believe that methods for systematic design of specificity and stability, such as the ILP-based method presented here, will become increasingly important as computational protein design is utilized for practical applications.

5.4 Materials and Methods

5.4.1 Cluster Expansion

The theory behind using cluster expansions to express the fixed-backbone energy of a protein as a function of its sequence is described in chapter 3 and ref [61]. In this study we derived a cluster expansion of model as HP/S/C described in chapter 2. The expansion was truncated at pair contributions. Only amino acids within one heptad of one another (both on the same chain and on opposite chains) were assumed to have significant interactions, so pair ECI only for these pairs were considered. This gave rise to 4 point clusters and 4 homotypic and 36 heterotypic pair clusters. Positions **a**, **d**, **e**, and **g** were allowed to vary among all natural amino acids except proline and glycine and positions **b**, **c**, and **f** were fixed as alanine. This resulted in a total of 9,929 ECI (1 constant, 68 point and 9,860 pair).

Although ECI are chosen to minimize the error of a cluster expansion, it is expected that the error will be larger for sequences very different from those present in the training set. Because the purpose of deriving a cluster expansion in this study was to design specific partners against naturally occurring coiled coils, we built a training set that represented well the sequence space encountered in nature. Amino-acid frequencies specific for each heptad position were derived from 432 native bZIP

sequences from 10 species (dataset obtained through personal communication with Mona Singh). 60,000 six-heptad parallel dimeric coiled-coil sequences were generated by randomly selecting amino acids with probabilities equal to their natural frequencies at the corresponding heptad positions. The training set was then enriched for amino-acid combinations with corresponding ECI that occurred less than five times by augmenting the set with five sequences that contained that particular combination with the rest of the amino acids chosen randomly as before. This brought the final size of the training set to 61,780. Each of these sequences was then repacked and evaluated with model HP/S/C as described before [60] as well as with model HP/S/Cv. Cluster expansions were derived by initially including only constant and point cluster functions and progressively considering all pair cluster functions, keeping only those that decreased the CV RMS error (for each set of cluster functions, ECI were obtained using standard least-squares fitting by the method of pseudo-inverse) [61]. The order in which pair cluster functions were visited was determined by the magnitude of their ECI when all pair cluster functions were included. The final expansion contained a total of 2,544 ECI and had RMS error of 2.2 kcal/mol. Figure 5-2 shows the performance of CE on the training set.

5.4.2 Formulation of the Problem as an Integer Linear Program

Singh and co-workers have shown how the problems of rotameric structure packing and protein design can be expressed as an integer linear program (ILP) [85]. To this end, the sequence/structure space in a protein design problem with p variable sites is represented as a undirected p -partite graph with node set $V = V_1 \cup \dots \cup V_p$. Set V_i contains one node for each rotamer at position i . Each node $u \in V_i$ is assigned a weight E_{uu} corresponding to its self energy. The edges of the graph $D = \{(u, v) : u \in V_i \text{ and } v \in V_j, i \neq j\}$ are assigned weights E_{uv} equal to the pairwise interaction energies between the corresponding rotamer pairs. A particular sequence/rotamer configuration can then be represented by specifying the set of nodes

in V and the set of edges in D that it involves. Given this formulation, the energy of an arbitrary sequence/rotamer configuration becomes $\varepsilon = \sum_{u \in V} E_{uu}x_{uu} + \sum_{u,v \in D} E_{uv}x_{uv}$, where binary decision variables x_{uu} and x_{uv} determine which nodes and edges from the graph are chosen. The problem of optimizing energy can thus be expressed as that of minimizing ε under the constraint that the chosen vertices and edges correspond to one another [85]:

$$\begin{aligned}
& \text{Minimize : } \varepsilon = \sum_{u \in V} E_{uu}x_{uu} + \sum_{u,v \in D} E_{uv}x_{uv} \\
& \text{subject to :} \\
& \sum_{u \in V_j} x_{uu} = 1 \quad \text{for } j = 1, \dots, p \\
& \sum_{u \in V_j} x_{uv} = x_{vv} \quad \text{for } j = 1, \dots, p \text{ and } v \in V \setminus V_j \\
& x_{uu}, x_{uv} \in \{0, 1\}
\end{aligned} \tag{5.1}$$

We extend this formulation to allow for multi-state design. First, to simplify the problem tremendously, we express the energetics of our target as well as negative states as analytical functions of sequence. Note that because these sequence-based expressions are still pairwise-decomposable, the ILP formulated above can still be used to find the optimal sequence in any one state (the difference is that now only amino-acid degrees of freedom remain, which drastically reduces the number of decision variables x_{uu} and x_{uv} and the number of constraints). The energy of a sequence in any state S can be expressed as $\varepsilon^S = \sum_{u \in V} E_{uu}^S x_{uu} + \sum_{u,v \in D} E_{uv}^S x_{uv}$, where weights E_{uu}^S and E_{uv}^S are simply the corresponding ECI from the cluster expansion for state S . Because the same decision variables are involved here as in single-state design, we can build ILPs similar to that in equation 5.1 to optimize any linear combination of state energies as well as to impose arbitrary inequality constraints on state energies. In this study we have chosen to optimize the energy of the target state T under constraints

on the gaps between T and the negative states N_i . An ILP that accomplishes this is:

$$\begin{aligned}
& \text{Minimize : } \varepsilon^T = \sum_{u \in V} E_{uu} x_{uu} + \sum_{u,v \in D} E_{uv} x_{uv} \\
& \text{subject to :} \\
& \sum_{u \in V_j} x_{uu} = 1 \quad \text{for } j = 1, \dots, p \\
& \sum_{u \in V_j} x_{uv} = x_{vv} \quad \text{for } j = 1, \dots, p \text{ and } v \in V \setminus V_j \\
& \varepsilon^{N_1} - \varepsilon^T > gc, \text{ where } \varepsilon^{N_1} = \sum_{u \in V} E_{uu}^{N_1} x_{uu} + \sum_{u,v \in D} E_{uv}^{N_1} x_{uv} \\
& \dots \\
& \varepsilon^{N_k} - \varepsilon^T > gc, \text{ where } \varepsilon^{N_k} = \sum_{u \in V} E_{uu}^{N_k} x_{uu} + \sum_{u,v \in D} E_{uv}^{N_k} x_{uv} \\
& x_{uu}, x_{uv} \in \{0, 1\}
\end{aligned} \tag{5.2}$$

where gc is the particular gap constraint imposed and k is the number of competing states. In this study we solved such ILPs with the *glpsol* tool from the GNU Linear Programming Kit. Because of the simplicity of sequence-dependent energy functions obtained through cluster expansion, solutions to these ILPs with as many as 49 negative states were normally obtained within 1-5 minutes on a single 2.7 GHz CPU. Whereas if full rotamer-level energy functions have to be considered, such optimizations become intractable by any current method.

Note that although here everything was formulated for sequence pairwise-decomposable energy functions, in principle this approach can be easily generalized for higher-order terms. Clearly, the CE methodology is already capable of taking higher-order interactions into account, should there be a need for that [61]. As far as the ILP problem formulation, it can be expanded to handle higher-order terms by introducing additional decision variables. For example, x_{uvw} would be 1 if there is a triplet interaction between rotamers u , v , and w at the corresponding sites. Additionally, constraints for these new decision variables would also have to be imposed to make sure that higher order interactions only occur between those rotamers that are “chosen” (e.g. in this case x_u , x_v and x_w are 1). Note that these higher-order decision variables would have to be introduced only for those clusters of sites that do, in fact, have higher-order interactions. This allows the complexity of the ILP problem to grow naturally with the size of the system (i.e. the number of variables and constraints grows linearly with

the number of interactions in the system). Additionally, clever pruning techniques, such as those proposed by Singh and co-workers [85], may be applied here as well to simplify the ILP problem.

5.4.3 Design Specifications

Each of the design calculations performed in this study sought to find a sequence that bound to a particular natural human bZIP coiled coil in a parallel dimeric manner. The set of explicitly considered negative states consisted of the parallel homo-dimer of the designed sequence as well parallel hetero-dimers with the remaining human bZIP coiled coils, except the sequences in the same family as the target (unless otherwise specified). Positions **a**, **d**, **e**, and **g** were allowed to vary over the 10 most frequently occurring amino acids at each position as follows: {V, L, N, I, K, A, R, T, Y, E} for **a**, {L, V, I, M, H, Y, T, A, K, F} for **d**, {E, K, R, Q, L, S, T, A, V, I} for **e** and {E, K, Q, R, L, Y, T, D, A, I} for **g**. **b**, **c** and **f** positions were fixed as alanine. The energy models used in this study approximate the effect of amino-acid substitutions on the stability of the parallel dimeric coiled-coil structure, but know nothing about alternative structural states such as aggregated states, the stabilization of which can lead to problems with solubility. Therefore, additional efforts were necessary to ensure that designed sequences had a hydrophobic/hydrophylic pattern favoring the coiled-coil state. The most common way of addressing this problem is to restrict the amino-acid library at each position based on the degree of burial, favoring hydrophobic amino acids in the core and polar amino acids on the surface. However, in coiled coils, charged and polar amino acids are frequently found in core positions, especially position **a**, and hydrophobic amino acids are often found on the surface. Therefore, we imposed a restriction at the level of the entire sequence, rather than requiring that particular positions be hydrophobic or hydrophilic. To this end, a position-specific scoring matrix (PSSM) was constructed for each heptad position based on a dataset of 432 native bZIP sequences. A constraint was incorporated into the ILP that required designed sequences to score above a certain cutoff using this PSSM. The cutoff was chosen such that 15% of natural bZIP sequences scored above

it.

The procedure for design consisted of a specificity sweep, where the stability of the target state was optimized under a progressively increasing constraint on the gap between the target state and the competing states (see figure 5-1). The first optimization was run without any gap constraints at all (i.e. $gc = \text{inf}$ in equation 5.2), meaning the sequence that optimized target state energy was found. Gaps between that sequence in the target state and all the negative states were then calculated and the smallest gap g_{min} (the most negative) identified. The next optimization was run with a gap constraint $gc = g_{min} - 1$ kcal/mol. This procedure was repeated until no sequences existed that could satisfy the imposed constraints. This chain of optimizations resulted in a list of sequences of decreasing stability and increasing specificity, which can be viewed as the limiting line in the specificity/stability phase space.

5.4.4 Choosing **b**, **c** and **f** positions

Identities of the **b**, **c** and **f** positions were chosen to be most appropriate for the already selected **a**, **d**, **e**, and **g** positions given what is observed in the dataset of 432 natural bZIP sequences. Thus, for each **b**, **c** or **f** position b_i we sought to optimize $P(b_i|a_1, \dots, a_n)$, where $a_1 \dots a_n$ are the identities of the selected **a**, **d**, **e**, and **g** positions. To this end we expressed this quantity in terms of probabilities we could measure from our dataset:

$$\begin{aligned}
 P(b_i|a_1, \dots, a_n) &= \frac{P(b_i, a_1 \dots a_n)}{P(a_1, \dots, a_n)} = \frac{P(a_1|b_i, a_2 \dots a_n) \cdot P(b_i, a_2 \dots a_n)}{P(a_1, \dots, a_n)} \quad (5.3) \\
 &= \frac{P(a_1|b_i, a_2 \dots a_n) \cdot P(a_2|b_i, a_3 \dots a_n) \dots P(a_n|b_i) \cdot P(b_i)}{P(a_1, \dots, a_n)} \\
 &\approx \frac{P(a_1|b_i) \cdot P(a_2|b_i) \dots P(a_n|b_i) \cdot P(b_i)}{P(a_1, \dots, a_n)}
 \end{aligned}$$

The last step assumes that the pre-selected amino-acid decoration at positions **a**, **d**, **e**, and **g** represents well the natively observed decorations at these positions (i.e. probability $P(a_k|b_i)$ measured in the particular given **adeg** context and that probability averaged over all native contexts are the same). Quantity $P(a_1, \dots, a_n)$ is hard to

estimate, but it is constant with respect to **b**, **c**, **f** and is therefore not important. Conditional probabilities $P(a_k|b_i)$ can be easily measured from the dataset and for each **b**, **c** and **f** position the amino acid that optimizes the probability in equation 5.3 can be found. Using this approach we were able to obtain **b**, **c**, **f** decorations of natural content and distribution. However, we found that infrequently this procedure resulted in sequences with large charge and/or helix propensity (mostly due to the fact that the pre-selected **a**, **d**, **e**, and **g** amino acids already had high values of charge or helix propensity). Some of our initial experimental testing indicated that extreme values of these properties may be undesirable. Large amounts of charge give rise to very strong salt effects, and high helix propensities make it difficult to discriminate between monomeric and dimeric states by circular dichroism. Thus, we modified the procedure for selecting **b**, **c** and **f** to guarantee that sequences with physical properties in a reasonable range were selected. The goal was still to optimize the probability in equation 5.3, but constraints on total charge and charge content (number of charged residues) as well as on the helix propensity of the entire sequence were imposed. The optimization problem was expressed as a integer linear program as for the optimization of energy in section 5.4.2. For each property, the range of acceptable values was defined as $\mu \pm \sigma$, where μ and σ are the mean and standard deviation of the corresponding property in the native bZIP dataset. In a few instances this resulted in no solutions (i.e. the selected **a**, **d**, **e** or **g** were already outside of the range for one of the properties) and for these cases more liberal intervals were allowed (either $\mu \pm 1.5\sigma$ or $\mu \pm 2\sigma$). Finally, since we wanted to rely on UV absorbance for determining concentration of our peptides in experimental characterization, we placed an additional constraint on the sequence of **b**, **c**, **f** to contain at least one Y or W residue (unless there was one already present at **a**, **d**, **e** or **g**).

Chapter 6

Conclusions

DSF-based of approaches have lead to significant successes over the past few years. Novel protein structures as well as enzymatic activity have been computationally designed [46, 95], structures of many proteins have been predicted with atomic accuracy [23] and interactions between proteins have been predicted using structural models [60, 84]. However, as I describe throughout my thesis, there are clearly still many limitations and challenges. Now is an interesting time in the evolution of protein modeling as computing technology has become more available and affordable than ever before. Thus, it is interesting to speculate on future prospects for DSF-based modeling.

6.1 Coarseness of Structural Sampling

In general, there is no principal difference between modeling proteins as having a discrete set of conformations and treating them continuously, as all modeling *in silico* is discrete. Even molecular dynamics simulations have to have a finite time step of integration, which means that conformational changes occurring on timescales beyond this interval can not be modeled. So the real difference then between DSF and continuous models is in the fineness of structural sampling. In fact, as computing technology advances, the boundaries between DSF and continuous modeling blur. This is particularly apparent in the work by Baker and co-workers [148]. Their ROSETTA

approach to structural modeling involves the assembly of protein backbones from a pre-compiled list of three- and nine-residue structural fragments obtained from the PDB. Although strictly a discrete sampling method, through its successful applications to structure prediction and docking, Baker and colleagues have demonstrated that the degree of flexibility obtained with ROSETTA is sufficient to describe much of the space of low-energy protein conformations. Of course, rigorous application of this methodology is very computationally complex as the search space of possible protein conformations is immense. Certainly, some of the recent success of this method can be attributed to the use of a distributed computing platform Rosetta@Home. Others have shown alternative methods, by which protein conformations can be systematically explored. Harbury *et. al* have used Crick parameterization of coiled-coil backbones [32] to explore flexibility in computational protein design [65]. DeGrado and co-workers have also used parameterization approaches to model helical bundles [136]. Dihedral angle perturbations, NMR ensembles and normal modes have also been used to generate collections of protein conformations [99, 88, 193].

The work of Pande and co-workers has been blurring the lines between continuous and discrete modeling from the other extreme – explicit atomic-level molecular dynamics simulation. Pande and colleagues represent the space of possible protein conformations on a folding pathway as a graph connecting discrete conformational neighborhoods, and they compute the transition probabilities between the neighborhoods close in structure [160]. Combining these data allows them to simulate transition probabilities of much larger conformational rearrangements extending simulation timescales far beyond those associated with traditional molecular dynamics. Although this approach employs explicit molecular dynamics simulation, the manner in which protein conformation space is represented shares resemblance with DSF-based models. Pande and colleagues have also benefitted tremendously from distributed computing technology through the Folding@Home platform.

The idea behind the DSF framework – breaking down the space of protein conformations into discrete bins, is sound and is a very promising direction. However, it is hard to know *a priori* what level of structural discretization will be appropriate for

different applications. That is why I think it is important to develop methods that can be generalized for arbitrary fineness of structural sampling. For example, the rotamer approximation used in protein design often appears to be sufficient to explain the relevant side-chain flexibility. However, it is possible that in order to model certain phenomena, it will be necessary to account for flexibility on a finer level. Some of the methods currently used to treat side-chain flexibility (e.g. self consistent mean field approaches) can in principle incorporate any number of off-rotameric structural states for each amino acid. On the other hand, modeling backbone flexibility by considering a finite set of variant backbones and performing separate calculations on each, may not scale quite as well.

6.2 Adjustable Energy Functions

Because it is not known what level of structural sampling will be necessary for different applications, it is also important to develop energy models that can be adjusted for different levels of coarseness. For example, as I show in chapter 4, although van der Waals interaction energy modeled with the Lennard-Johns potential is completely decomposable in terms of atom pair contributions, when structural degrees of freedom are discretized, such strict decomposability is lost. The severity of this problem is directly related to the fineness of structural sampling. Thus, an approach must be developed that can systematically adjust energetic models for an arbitrary level of structural sampling. In the case of van der Waals energy, this may involve introducing triplet or higher-order interactions between side-chain rotamers. One method that can potentially fill this need is the cluster expansion approach I describe in chapter 3. However, other approaches are also possible. For example, in the field of reduced protein models (beads-on-a-string or lattice representations of proteins), one is often concerned with choosing an appropriate energy function for the reduced representation so as to optimally recapitulate the properties observed in real proteins [38]. Approaches of similar nature may also prove to be useful for dealing with structural discretization.

6.3 Unfolded States

Finally, the ability of DSF models to treat unstructured states of proteins is quite limited. In general, the idea of structural discretization is probably less natural for the unfolded state given that it is an heterogeneous ensemble of a large number of structural conformations. Modeling the unfolded state by only accounting for side chain-to-backbone interactions with one backbone conformation is unrealistic and I do not expect this approach to work well in the future. However, in principle, it is possible to improve upon this model while still remaining in the DSF framework. One could explicitly consider a large enough ensemble of discrete backbone conformations such that averages over this ensemble would approximate well thermodynamic properties of the unfolded state. Approaches akin to those employed by Pande and co-workers or Baker and colleagues may make this possible. However, it is not clear that this is the best approach. Because of its heterogeneous nature, it may be better to model the unfolded state with the help of more classical thermodynamic methods. In this respect, previous work on lattice models [155, 1] and Ising-like models [26] may prove useful. It is also important to realize that part of the reason that unfolded state models are currently very limited is the small amount of experimental evidence isolating the effects of protein behavior to the unfolded state. Marti *et. al* have demonstrated that electrostatic repulsion in the unfolded state can stabilize a leucine zipper [118]. There have also been attempts to structurally characterize unfolded state ensembles [14]. More studies of this sort should aid greatly in the development of appropriate models.

6.4 Summary

In summary, discrete structural flexibility models have been useful for a large range of applications over the past decade. Nowadays, as high-performance computing technology becomes more available, the boundaries between discrete and continuous modeling begin to disappear. To address this convergence, new methods for systematically dealing with varying degrees of coarseness of structural sampling need to be

developed. Some promising new directions towards addressing this need already exist and over the next decade I think we will see a qualitative improvement in the accuracy and applicability of DSF-based models.

Chapter 7

Possible Future Directions

7.1 Specificity Design Framework

We have so far limited the application of our specificity design framework to the parallel dimeric coiled coil system. Other flavors of coiled coils, such as antiparallel dimers or mixed higher-order oligomers, come to mind as obvious candidates for future applications. For coiled coils the barrier between different orientations and oligomerization states can often be low and a few mutations can easily tip the balance for one state relative another. So the ability to account for various possible orientation and oligomerization states can be very useful in design. Additionally, it may be practically useful to be able to specifically design anti-parallel dimers or higher-order oligomers. The biggest limitation for this problem is currently the lack of reliable energy functions that account for these alternative coiled-coil states. Deriving such energy functions is therefore an important future direction (see section 7.2 for on this).

In principle, the specificity design framework, as formulated in chapter 5, is generalizable and can be applied to systems other than the coiled coil. Some technical augmentations, however, may make this generalization easier. Given the current formulation of the framework, only up to pairwise interactions between amino acid at various sites in any given state can be treated. For the dimeric coiled-coil system, up to pairwise interactions capture the majority of energetic effects and are sufficient for

reasonable accuracy (see ref [61] and figure 5-2 in chapter 5). However, it is possible that for other systems higher-order interactions will be necessary [61]. The cluster expansion formalism readily allows for the incorporation of such higher-order contributions, should they be necessary to increase the accuracy of the expansion. However, the Integer Linear Program used by the specificity sweep procedure currently only allows for up to pairwise interactions. In section 5.4.2 of chapter 5 I briefly outline how this limitation can be broken and implementing this functionality is probably key to applying the framework more widely.

7.2 Structure-based Modeling of Coiled-coil Interactions

Our structure-based model for parallel dimeric coiled-coil interactions (model HP/S/C – see chapter 2) has proven reasonably accurate in prediction as well as shown good potential in design. However, there are several limitations to the model that we are aware of, and there are many potential approaches to addressing these limitations that we have not yet been pursued. One of the shortfalls of our structure-based approach is that it is not able to correctly predict values of experimentally measured coupling energies for $\mathbf{a} - \mathbf{a}'$ interactions, especially those involving asparagine. In HP/S/C we have temporarily addressed this limitation by replacing computed $\mathbf{a} - \mathbf{a}'$ and $\mathbf{d} - \mathbf{d}'$ interactions with corresponding empirical weights from a machine learning model (see chapter 2). Although this has worked well so far, it is not the most satisfying solution to the problem. Additionally, we have already noticed some biases arising in design that are most likely due to the less precise manner by which interaction weights are assigned in the machine learning approach, especially for those amino-acid pairs that occur rarely in the training set.

One possibility for why our DSF-based approaches have failed to reproduce correct $\mathbf{a} - \mathbf{a}'$ coupling energies may be the lack of backbone flexibility in our modeling. In deriving model HP/S/C, we attempted to crudely account for backbone flexibility

by giving each potential dimer a choice of eight different ideal backbones. Although this did not give a significant improvement in result, there is much more to be done before we can rule out backbone flexibility as a major source of error. First, it is likely that eight backbones are not enough and more sampling is required. Also, it is possible that rather than performing grid-based sampling (i.e. using the same predefined backbones for all sequences), it may be more efficient and appropriate to search in backbone space separately for each sequence. There is precedent for this type of approach in the design and structure prediction and it may work very well in our case, since the space of backbone variations in parallel dimeric coiled coils is quite limited. The search for an appropriate backbone can be done either with stochastic Monte Carlo-like techniques or with dynamics (either explicit molecular dynamics or reduced complexity dynamics).

I am currently pursuing a molecular dynamics-based approach for calculating $\mathbf{a} - \mathbf{a}'$ coupling energies. If this approach is successful, it may provide insight into why the DSF-based models we have applied to the task have failed. Whether it comes from molecular dynamics, another modeling approach or from experimentation, I think it is important to gain a deeper physical understanding for why the measured $\mathbf{a} - \mathbf{a}'$ coupling energies are what they are and how dependent on context they are. This will potentially allow us to adjust our reduced models to capture the necessary effects.

When deriving model HP/S/C, we did not systematically analyze the effect of explicitly modeling positions \mathbf{b} , \mathbf{c} and \mathbf{f} . The fact that we get quite reasonable performance by ignoring these positions says that much of the coiled-coil interaction specificity in natural sequences is independent of amino acids in \mathbf{b} , \mathbf{c} and \mathbf{f} . However, through our design work we have discovered that inappropriate choice of sequence at \mathbf{b} , \mathbf{c} and \mathbf{f} can lead to significantly weakened interactions. Therefore, I think that looking for possible improvements in performance due to explicitly modeling \mathbf{b} , \mathbf{c} and \mathbf{f} positions, is a potentially fruitful future direction. Besides being applicable to the coiled-coil system, findings of such a study may help understand the contributions of non-interfacial amino acids in other helix-mediated interactions (such as helix-to-

grove).

Something that is currently at a very simplistic level in model HP/S/C is the treatment of the reference (unfolded) state. Our finding that intra-helix pairwise interaction contribute much less to stability than inter-helix ones indicates that pairwise contributions to the unfolded state energy are important. Unfortunately, it is not trivial how these contributions can be accounted for in a rigorous manner, so in HP/S/C they are treated implicitly and crudely by scaling intra- versus inter-chain interactions differently. An alternative way to approaching this problem is to use a statistical representation of the unfolded state, where all interactions have some probability of occurring, and tune the parameters of the statistical ensemble as well as the parameters of residue-residue interactions, to optimize performance. A possible drawback of such an approach is that too many adjustable parameters may need to be used to make the model physically reasonable, making it difficult to obtain a statistically-meaningful fit. Another potential approach would be to explicitly model a representative structural ensemble of unfolded structures. Sosnick and co-workers have developed a method for generating explicit random coil ensembles for arbitrary protein sequences, and have shown that their ensembles reproduce experimentally-measured unfolded state characteristics such as radius of gyration, while retaining a significant amount of locally native structure – a feature of the unfolded state often noted in spectroscopic studies [79]. It would be very interesting to know whether such explicit ensembles can be used to improve modeling of the unfolded state. One problem with such an approach is that the amount of computational time necessary to treat a reasonable unfolded ensemble even for one sequence can be quite large, especially if used in conjunction with a sophisticated energy function. However, initially one can try to do this using a very simple energy function (one with a fast treatment of solvation, such as EEF1), such that the evaluation of hundreds of structures can be done per second. One then would simply test whether the presence of such an explicit unfolded state model improves prediction results relative to using the same simple energy function but without an unfolded state (i.e. all sequences have the same free energy in the unfolded state). Using such a test, we were able to eliminate the

penta-peptide model for the unfolded state as inappropriate for modeling coiled-coil association. I would be very curious to know if any improvement can be obtained with an explicit unfolded state model and, if yes, how this improvement changes with the number of structures considered in the unfolded ensemble.

Finally, it would be interesting and useful to try to extend model HP/S/C to treat other coiled-coil orientations and oligomerization states. The biggest limitation I see with this is the lack of a uniform and clear experimental dataset that can be used for model validation and training. Microarray technology applied to the bZIP system provided us with a semi-quantitative dataset of relative interaction strengths for over 1,000 potential parallel dimeric coiled-coil interactions, which was integral in deriving a reasonable model. A similar dataset does not exist for other orientation and oligomerization states. Of course, one can compile a dataset of coiled-coils sequences with verified orientation and oligomeric states, such that the ability of different methods to discriminate between these can be ascertained. Perhaps the easiest way to do this is to look for coiled coils with available structures and sequences with homology to those with known structure can also be considered. However, the problem of discriminating between different orientation and oligomerization states is different from the problem of capturing the relative stability of different sequences in the same state, although an ideal energy function could do both. For example, the unfolded state is unimportant for the simple folded state discrimination problem. In order to be able to derive a reasonable unfolded state model, a dataset of relative stabilities is necessary. It would be nice, for example, to perform similar microarray experiments to the ones done with human bZIPs on a set of anti-parallel coiled coils. Gathering relative stability data for higher-order coiled-coil oligomers may be more difficult. However, if we derive a unified structure-based model that works well for predicting relative stabilities of both parallel and anti-parallel dimeric coiled coils, then it may be reasonable to expect that the model is general enough that its extension to arbitrary oligomerization states may also work well. Perhaps then such a model does not need to be quite as extensively verified for higher-order oligomers in prediction mode and can be directly applied in design mode.

7.3 Cluster Expansion

One of the significant advantages of the cluster expansion approach as I present it in chapter 3 is that computationally very expensive models can potentially be expressed as very simple functions of sequence. The only limitation is that the original models have to be fast enough such that energies for a training set of sequences (usually several tens of thousands) can be computed. Optimally, one would like to expand an energy function that is based on explicit molecular dynamics (MD) simulations, but obtaining free energies from explicit MD is currently difficult for protein-sized systems. Hybrid MD-based models, such as MM/PBSA methods, have recently shown some promising results [167]. They are computationally fast enough that cluster expansions based on MM/PBSA energies are probably within reach (at least for small systems). It would be interesting to apply this approach in the context of design.

A potential improvement to the cluster expansion method itself would be finding better ways to identify potentially contributing higher-order interactions. Because the number of possible interactions grows exponentially with cluster size, for most systems it is impossible to enumerate over all interactions beyond the pairwise ones. Thus, one has to have an idea which high-order interactions are likely to contribute. Clearly, physical intuition dictates that combinations of residues very far apart in structure should generally not have a significant energetic contribution. However, this does not restrict the number of potential clusters to a small enough set and, further, this is just a trend rather than a strict condition. It would be nice to have some set of criteria, by which potentially contributing high-order interactions can be identified. An interesting project may be to consider one or a few systems that are small enough such that all triplets can be enumerated and important ones identified, and see if there are any conditions or structural properties that correlate with a high-order cluster having large contributions. If such properties are identified, one may try to move to a larger system and see how much of an improvement in expansion accuracy (in the sense of cross-validated error) can be obtained by considering high-order interactions identified *a priori* with the above properties.

Bibliography

- [1] Kolinski A. and Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins*, 18(4):338–352, 1994.
- [2] A. Acharya, V. Rishi, and C. Vinson. Stability of 100 Homo and Heterotypic Coiled-Coil α - α' Pairs for Ten Amino Acids (A, L, I, V, N, K, S, T, E, and R). *Biochemistry*, 45(38):11324 –11332, 2006.
- [3] A. Acharya, S. B. Ruvinov, J. Gal, J. R. Moll, and C. Vinson. A heterodimerizing leucine zipper coiled coil system for examining the specificity of a position interactions: amino acids I, V, L, N, A, and K. *Biochemistry*, 41(48):14122–31, 2002.
- [4] M. H. Ali, C. M. Taylor, G. Grigoryan, K. N. Allen, B. Imperiali, and A. E. Keating. Design of a heterospecific, tetrameric, 21-residue miniprotein with mixed α / β structure. *Structure (Camb)*, 13(2):225–34, 2005.
- [5] P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R. B. Russell. Structure-based assembly of protein complexes in yeast. *Science*, 303(5666):2026–9, 2004.
- [6] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, 99(9):5896–901, 2002.
- [7] Fogolari F. and Brigo A. and Molinari H. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.*, 15(6):377–392, 2002.

- [8] G. D. Amoutzias, A. S. Veron, J. III Weiner, M. Robinson-Rechavi, E. Bornberg-Bauer, S. G. Oliver, and D. L. Robertson. One Billion Years of bZIP Transcription Factor Evolution: Conservation and Change in Dimerization and DNA-Binding Site Specificity. *Mol. Biol. Evol.*, 24(3):827–835, 2006.
- [9] P. Angel and M. Karin. The role of Jun, Fos and the AP-1 complex in cell-proliferation and transformation. *Biochim Biophys. Acta.*, 1072(2-3):129–57, 1991.
- [10] J. Ashworth, J. J. Havranek, C. M. Duarte, D. Sussman, Jr. Monnat, R. J., B. L. Stoddard, and D. Baker. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441(7093):656–9, 2006.
- [11] M. Asta, V. Ozolins, and C. Woodward. A First-Principles Approach to Modeling Alloy Phase Equilibria. *JOM*, pages 16–19, 2001.
- [12] P. V. Benos, A. S. Lapedes, and G. D. Stormo. Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol*, 323(4):701–27, 2002.
- [13] B. Berger, D. B. Wilson, E. Wolf, T. Tonchev, M. Milla, and P. S. Kim. Predicting coiled coils by use of pairwise residue correlations. *Proc Natl Acad Sci U S A*, 92(18):8259–63, 1995.
- [14] P. Bernado, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci. USA*, 102(47):1700217007, 2005.
- [15] Paul Beroza and D. R. Fredkin. Calculation of amino acid pKaS in a protein from a continuum electrostatic model: Method and sensitivity analysis. *Journal of Computational Chemistry*, 17(10):1229–1244, 1996.
- [16] Paul Beroza and D. R. Fredkin. Calculation of amino acid pKaS in a protein from a continuum electrostatic model: Method and sensitivity analysis. *Journal of Computational Chemistry*, 17(10):1229–1244, 1996.

- [17] T. N. Bhat, V. Sasisekharan, and M. Vijayan. An analysis of side-chain conformation in proteins. *Int J Pept Protein Res*, 13(2):170–84, 1979.
- [18] M. Blaber, X. J. Zhang, and B. W. Matthews. Structural basis of amino acid alpha helix propensity. *Science*, 260(5114):1637–40, 1993.
- [19] V. Blank and N. C. Andrews. The Maf transcription factors: regulators of differentiation. *Trends Biochem Sci*, 22(11):437–41, 1997.
- [20] D. N. Bolon, D. A. Wah, G. L. Hersch, T. A. Baker, and R. T. Sauer. Bivalent Tethering of SspB to ClpXP Is Required for Efficient Substrate Delivery: A Protein-Design Study. *Mol. Cell*, 13:443–449, February 2004.
- [21] P. Bork, L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee, and E. M. Marcotte. Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, 14(3):292–9, 2004.
- [22] P. Bradley, L. Cowen, M. Menke, J. King, and B. Berger. BETAWRAP: successful prediction of parallel beta -helices from primary sequence reveals an association with many microbial pathogens. *Proc Natl Acad Sci U S A*, 98(26):14819–24, 2001.
- [23] P. Bradley, K. M. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309:1868–1871, 2005.
- [24] B. Brannetti, A. Via, G. Cestra, G. Cesareni, and M. Helmer-Citterich. SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. *J Mol Biol*, 298(2):313–28, 2000.
- [25] 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 Comp Chem*, 4(2):187–217, 1983.
- [26] J. D. Bryngelson and P. G. Wolynes. Spin Glasses and the Statistical Mechanics of Protein Folding. *Proc. Natl. Acad. Sci. USA*, 84:7524–7528, 1987.

- [27] Jr. Carter, C. W., B. C. LeFebvre, S. A. Cammer, A. Tropsha, and M. H. Edgell. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J Mol Biol*, 311(4):625–38, 2001.
- [28] G Ceder. Predicting Properties from Scratch. *Science*, 280(15 May):1099–1100, 1998.
- [29] A. Chakrabartty, T. Kortemme, and R. L. Baldwin. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci*, 3(5):843–52, 1994.
- [30] R. Chandrasekaran and G. N. Ramachandran. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformations in proteins. *Int. J. Protein. Res.*, 2:223–233, 1970.
- [31] T. P. Creamer and G. D. Rose. Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc Natl Acad Sci U S A*, 89(13):5937–41, 1992.
- [32] F. H. C Crick. The Packing of alpha-Helices: Simple Coiled-Coils. *Acta Crystallogr*, 6:689 – 697, 1953.
- [33] F. H. C Crick. The Packing of alpha-Helices: Simple Coiled-Coils. *Acta Crystallogr*, 6:689 – 697, 1953.
- [34] B. I. Dahiyat and S. L. Mayo. Protein design automation. *Protein Sci*, 5(5):895–903, 1996.
- [35] B. I. Dahiyat and S. L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–7, 1997.
- [36] B. I. Dahiyat and S. L. Mayo. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A*, 94(19):10172–7, 1997.

- [37] G. Dantas, B. Kuhlman, D. Callender, M. Wong, and D. Baker. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J Mol Biol*, 332(2):449–60, 2003.
- [38] P. Das, S. Matysiak, and C. Clementi. Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc. Natl. Acad. Sci. USA*, 102(29):10141–10146, 2005.
- [39] D. de Fontaine. Cluster Approach to Order-Disorder Transformations in Alloys. *Solid State Phys*, 47:33, 1994.
- [40] C. D. Deppmann, A. Acharya, V. Rishi, B. Wobbes, S. Smeeckens, E. J. Taparowsky, and C. Vinson. Dimerization specificity of all 67 B-ZIP motifs in *Arabidopsis thaliana*: a comparison to *Homo sapiens* B-ZIP motifs. *Nucleic Acids Res*, 32(11):3435–45, 2004.
- [41] J. R. Desjarlais and T. M. Handel. De novo design of the hydrophobic cores of proteins. *Protein Sci*, 4(10):2006–18, 1995.
- [42] J. Desmet, M. De Maeyer, B. Hazes, and I Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.
- [43] Jr. Dunbrack, R. L. Rotamer libraries in the 21st century. *Curr Opin Struct Biol*, 12(4):431–40, 2002.
- [44] Jr. Dunbrack, R. L. and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*, 6(8):1661–81, 1997.
- [45] Jr. Dunbrack, R. L. and M. Karplus. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol*, 230(2):543–74, 1993.
- [46] M. A. Dwyer, L. L. Looger, and H. W. Hellinga. Computational design of a biologically active enzyme. *Science*, 304(5679):1967–71, 2004.

- [47] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.
- [48] T. E. Ellenberger, C. J. Brandl, K. Struhl, and S. C. Harrison. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell*, 71(7):1223–37, 1992.
- [49] 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(4):451–64, 1998.
- [50] H. Eyring. Steric hindrance and collision diameters. *J. Am. Chem. Soc.*, 54:3191–3203, 1932.
- [51] N. C. Fitzkee and G. D. Rose. Reassessing random-coil statistics in unfolded proteins. *Proc Natl Acad Sci U S A*, 101(34):12497–502, 2004.
- [52] J. H. Fong, A. E. Keating, and M. Singh. Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol*, 5(2):R11, 2004.
- [53] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–79, 1995.
- [54] D. Gilis and M. Rooman. PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng*, 13(12):849–56, 2000.
- [55] J. N. Glover and S. C. Harrison. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, 373(6511):257–61, 1995.
- [56] R. F. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys J*, 66(5):1335–40, 1994.
- [57] D. B. Gordon, S. A. Marshall, and S. L. Mayo. Energy functions for protein design. *Curr Opin Struct Biol*, 9(4):509–13, 1999.

- [58] D. B. Gordon and S. L. Mayo. Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure Fold Des*, 7(9):1089–98, 1999.
- [59] D. F Green, A. T. Dennis, P. S. Fam, B. Tidor, and A. Jasanoff. Rational Design of New Binding Specificity by Simultaneous Mutagenesis of Calmodulin and a Target Peptide. *Biochemistry*, 45:12547–12559, 2006.
- [60] G. Grigoryan and A. E. Keating. Structure-based prediction of bZIP partnering specificity. *J. Mol. Biol.*, 355:1125–1142, 2006.
- [61] G. Grigoryan, F. Zhou, S. R. Lustig, G. Ceder, D. Morgan, and A. E. Keating. Ultra-Fast Evaluation of Protein Energies Directly from Sequence. *PLoS Comp Biol*, 2(6):e63, June 2006.
- [62] R. Guerois, J. E. Nielsen, and L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*, 320(2):369–87, 2002.
- [63] T. Hai and T. Curran. Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *Proc Natl Acad Sci U S A*, 88(9):3720–4, 1991.
- [64] T. Hai and M. G. Hartman. The molecular biology and nomenclature of the activating transcription factor/cAMP responsive element binding family of transcription factors: activating transcription factor proteins and homeostasis. *Gene*, 273(1):1–11, 2001.
- [65] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, 282(5393):1462–7, 1998.
- [66] 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(18):8408–12, 1995.

- [67] J. J. Havranek and P. B. Harbury. Tanford-Kirkwood electrostatics for protein modeling. *Proc Natl Acad Sci U S A*, 96(20):11145–50, 1999.
- [68] James J. Havranek and Pehr B. Harbury. Automated design of specificity in molecular recognition. *Nature Structural Biology*, 10(1):45–52, 2003.
- [69] A. HEIFETZ, E KATCHALSKI-KATZIR, and M EISENSTEIN. Electrostatics in proteinprotein docking. *Prot. Sci.*, 11:571–587, 2002.
- [70] 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(7):1381–92, 1999.
- [71] T. Herdegen and J. D. Leah. Inducible and constitutive transcription factors in the mammalian nervous system: control of gene expression by Jun, Fos and Krox, and CREB/ATF proteins. *Brain Res Brain Res Rev*, 28(3):370–490, 1998.
- [72] V. J. Hilser, J. Gomez, and E. Freire. The enthalpy change in protein folding and binding: refinement of parameters for structure-based calculations. *Proteins*, 26(2):123–33, 1996.
- [73] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–9, 1995.
- [74] A. Horovitz, J. M. Matthews, and A. R. Fersht. Alpha-helix stability in proteins. II. Factors that influence stability at an internal position. *J Mol Biol*, 227(2):560–8, 1992.
- [75] S. J. Hubbard and J. M. Thornton. 'NACCESS' Computer Program. *Department of Biochemistry and Molecular Biology, University College, London*, 1993.
- [76] H. C. Hurst. Transcription factors 1: bZIP proteins. *Protein Profile*, 2(2):101–68, 1995.

- [77] J. L. Ilesley, M. Sudol, and S. J. Winder. The WW domain: linking cell signalling to the membrane cytoskeleton. *Cell Signal*, 14(3):183–9, 2002.
- [78] J. Janin, S. Wodak, M. Levitt, and B. Maigret. Conformation of amino acid side-chains in proteins. *J Mol Biol*, 125(3):357–86, 1978.
- [79] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. USA*, 102(37):13099–13104, 2005.
- [80] L. Jiang, B. Kuhlman, T. Kortemme, and D. Baker. A “solvated rotamer” approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins*, 58:893–904, 2005.
- [81] R. A. Kammerer, T. Schulthess, R. Landwehr, A. Lustig, J. Engel, U. Aebi, and M. O. Steinmetz. An autonomous folding unit mediates the assembly of two-stranded coiled coils. *Proc Natl Acad Sci U S A*, 95(23):13419–24, 1998.
- [82] T. Kaplan, N. Friedman, and H. Margalit. Ab Initio Prediction of Transcription Factor Targets Using Structural Knowledge. *PLoS Computational Biology*, 1(1):5–13, 2005.
- [83] 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 Natl Acad Sci U S A*, 98(26):14825–30, 2001.
- [84] C. Kiel, L. Serrano, and C. Herrmann. A detailed thermodynamic analysis of ras/effector complex interfaces. *J Mol Biol*, 340(5):1039–58, 2004.
- [85] C. L. Kingsford, B. Chazelle, and M. Singh. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, 21(7):1028–1036, June 2005.
- [86] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol*, 239(2):249–75, 1994.

- [87] P. Koehl and M. Levitt. De novo protein design. I. In search of stability and specificity. *J Mol Biol*, 293(5):1161–81, 1999.
- [88] H. Kono and J. G. Saven. Statistical Theory for Protein Combinatorial Libraries. Packing Interactions, Backbone Flexibility, and the Sequence Variability of a Main-chain Structure. *J. Mol. Biol.*, 306:607–628, 2001.
- [89] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, 99(22):14116–21, 2002.
- [90] T Kortemme, L. A. Joachimiak, A. N. Bullock, A. D. Schuler, B. L. Stoddard, and D. Baker. Computational redesign of protein-protein interaction specificity. *Nat. Struct. Mol. Biol.*, 11(4):371–379, 2004.
- [91] T. Kouzarides and E. Ziff. Leucine zippers of fos, jun and GCN4 dictate dimerization specificity and thereby control DNA binding. *Nature*, 340(6234):568–71, 1989.
- [92] C. M. Kraemer-Pecore, J. T. Lecomte, and J. R. Desjarlais. A de novo redesign of the WW domain. *Protein Sci*, 12(10):2194–205, 2003.
- [93] C. M. Kraemer-Pecore, A. M. Wollacott, and J. R. Desjarlais. Computational protein design. *Curr Opin Chem Biol*, 5(6):690–5, 2001.
- [94] D. Krylov, J. Barchi, and C. Vinson. Inter-helical interactions in the leucine zipper coiled coil dimer: pH and salt dependence of coupling energy between charged amino acids. *J Mol Biol*, 279(4):959–72, 1998.
- [95] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–8, 2003.
- [96] Brian Kuhlman and David Baker. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, 97(19):10383–10388, 2000.

- [97] E. Lacroix, A. R. Viguera, and L. Serrano. Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol*, 284(1):173–91, 1998.
- [98] J. H. Laity, B. M. Lee, and P. E. Wright. Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol*, 11(1):39–46, 2001.
- [99] S. M. Larson, J. L. England, J. R. Desjarlais, and V. S. Pande. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. *Protein Sci.*, 11:2804–2813, 2002.
- [100] J. K. Lassila, J. R. Keeffe, P. Oelschlaeger, and S. L. Mayo. Computationally designed variants of Escherichia coli chorismate mutase show altered catalytic activity. *Protein Eng Des Sel*, 18(4):161–3, 2005.
- [101] I. Lasters, M. De Maeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng*, 8(8):815–22, 1995.
- [102] G. A. Lazar, S. A. Marshall, J. J. Plecs, S. L. Mayo, and J. R. Desjarlais. Designing proteins for therapeutic applications. *Curr Opin Struct Biol*, 13(4):513–8, 2003.
- [103] T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins*, 35(2):133–52, 1999.
- [104] T. Lazaridis and M. Karplus. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol*, 10(2):139–45, 2000.
- [105] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33(2):227–39, 1998.
- [106] J. Lekstrom-Himes and K. G. Xanthopoulos. Biological role of the CCAAT/enhancer-binding protein family of transcription factors. *J Biol Chem*, 273(44):28545–8, 1998.

- [107] M. Levitt. Protein Folding by Restrained Energy Minimization and Molecular Dynamics. *J. Mol. Biol.*, 170:723–764, 1983.
- [108] A. M. Levy, O. Gilad, L. Xia, Y. Izumiya, J. Choi, A. Tsalenko, Z. Yakhini, R. Witter, L. Lee, C. J. Cardona, and H. Kung. Mareks disease virus Meq transforms chicken cells via the v-Jun transcriptional cascade: a converging transforming pathway for avian oncoviruses. *Proc. Natl. Acad. Sci. USA*, 102(41):1483114836, 2005.
- [109] T. D. Littlewood and G. I. Evan. Transcription factors 2: Helix-Loop-Helix Proteins. *Protein Profile*, 2(6):621–702, 1995.
- [110] J. L. Liu, L. F. Lee, Y. Ye, Z. Qian, and H. J. Kung. Nucleolar and nuclear localization properties of a herpesvirus bZIP oncoprotein, MEQ. *J. Virol.*, 71(4):31883196, 1997.
- [111] S. Liu, C. Zhang, H. Zhou, and Y. Zhou. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, 56(1):93–101, 2004.
- [112] A. Liwo, M. Khalili, , and H. A. Scheraga. *Ab initio* simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. USA*, 102(7):23622367, 2005.
- [113] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins*, 40(3):389–408, 2000.
- [114] L. Lu, H. Lu, and J. Skolnick. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):350–64, 2002.
- [115] A. Lupas, M. Van Dyke, and J. Stock. Predicting coiled coils from protein sequences. *Science*, 252(5010):1162–4, 1991.

- [116] P. C. Lyu, M. I. Liff, L. A. Marky, and N. R. Kallenbach. Side chain contributions to the stability of alpha-helical structure in peptides. *Science*, 250(4981):669–73, 1990.
- [117] S. M. Malakauskas and S. L. Mayo. Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol*, 5(6):470–5, 1998.
- [118] Daniel N. Marti and Hans Rudolf Bosshard. Inverse Electrostatic Effect: Electrostatic Repulsion in the Unfolded State Stabilizes a Leucine Zipper. *Biochemistry*, 43:12436 – 12447, 2004.
- [119] J. M. Mason and K. M. Arndt. Coiled coil domains: stability, specificity, and biological implications. *ChemBiochem*, 5(2):170–6, 2004.
- [120] A. V. McDonnell, T. Jiang, A. E. Keating, and B. Berger. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*, 2005.
- [121] M. J. McGregor, S. A. Islam, and M. J. Sternberg. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol*, 198(2):295–310, 1987.
- [122] J. Mendes, R. Guerois, and L. Serrano. Energy estimation in protein design. *Curr Opin Struct Biol*, 12(4):441–6, 2002.
- [123] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091, 1953.
- [124] L. A. Mirny and E. I. Shakhnovich. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol*, 264(5):1164–79, 1996.
- [125] K. M. Misura and D. Baker. Progress and challenges in high-resolution refinement of protein structure models. *Proteins*, 59(1):15–29, 2005.

- [126] K. M. Misura, D. Chivian, C. A. Rohl, D. E. Kim, and D. Baker. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc. Natl. Acad. Sci. USA*, 103:5361–5366, 2006.
- [127] R. Mndez, R. Leplae, M. F. Lensink, and S. J. Wodak. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins: Structure, Function, and Bioinformatics*, 60(2):150–169, 2005.
- [128] J. Moitra, L. Szilak, D. Krylov, and C. Vinson. Leucine is the most stabilizing aliphatic amino acid in the d position of a dimeric leucine zipper coiled coil. *Biochemistry*, 36(41):12567–73, 1997.
- [129] Y. K. Mok, E. L. Elisseeva, A. R. Davidson, and J. D. Forman-Kay. Dramatic stabilization of an SH3 domain by a single substitution: roles of the folded and unfolded states. *J Mol Biol*, 307(3):913–28, 2001.
- [130] V. Munoz and L. Serrano. Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J Mol Biol*, 245(3):275–96, 1995.
- [131] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, 1995.
- [132] J. K. Myers and T. G. Oas. Reinterpretation of GCN4-p1 folding kinetics: partial helix formation precedes dimerization in coiled coil folding. *J Mol Biol*, 289(2):205–9, 1999.
- [133] S. Nauli, B. Kuhlman, and D. Baker. Computer-based redesign of a protein folding pathway. *Nat Struct Biol*, 8(7):602–5, 2001.
- [134] J. R. Newman and A. E. Keating. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science*, 300(5628):2097–101, 2003.

- [135] M. J. Nohaile, Z. S. Hendsch, B. Tidor, and R. T. Sauer. Altering dimerization specificity by changes in surface electrostatics. *Proc. Natl. Acad. Sci. USA*, 98(6):31093114, 2001.
- [136] B. North, C. M. Summa, G. Ghirlanda, and W. F. DeGrado. D_n -symmetrical tertiary templates for the design of tubular proteins. *J. Mol. Biol.*, 311(5):1081–1090, 2001.
- [137] A. J. Oakley, M. Lo Bello, G. Ricci, G. Federici, and M. W. Parker. Evidence for an induced-fit mechanism operating in pi class glutathione transferases. *Biochemistry*, 37(28):9912–7, 1998.
- [138] M. G. Oakley and P. S. Kim. A buried polar interaction can direct the relative orientation of helices in a coiled coil. *Biochemistry*, 37(36):12603–10, 1998.
- [139] K. T. O’Neil and W. F. DeGrado. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science*, 250(4981):646–51, 1990.
- [140] A. Onufriev, D. A. Case, and D. Bashford. Effective Born radii in the generalized Born approximation: the importance of being perfect. *J Comput Chem*, 23(14):1297–304, 2002.
- [141] S. Park, X. Yang, and J. G. Saven. Advances in computational protein design. *Curr Opin Struct Biol*, 14(4):487–94, 2004.
- [142] N. P. Pavletich and C. O. Pabo. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*, 252(5007):809–17, 1991.
- [143] N. A. Pierce, J. A. Spriet, J. Desmet, and S. L. Mayo. Conformational splitting: A more powerful criterion for dead-end elimination. *Journal of Computational Chemistry*, 21(11):999–1009, 2000.
- [144] N. Pokala and T. M. Handel. Review: protein design—where we were, where we are, where we’re going. *J Struct Biol*, 134(2-3):269–81, 2001.

- [145] 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(4):775–91, 1987.
- [146] D. P. Ramji and P. Foka. CCAAT/enhancer-binding proteins: structure, function and regulation. *Biochem J*, 365(Pt 3):561–75, 2002.
- [147] J. Reina, E. Lacroix, S. D. Hobson, G. Fernandez-Ballester, V. Rybin, M. S. Schwab, L. Serrano, and C. Gonzalez. Computer-aided design of a PDZ domain to recognize new target sequences. *Nat. Struct. Biol.*, 9(8):621–627, 2002.
- [148] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using Rosetta. *Methods Enzymol.*, 383:69–93, 2004.
- [149] A. Rossi, C. Micheletti, F. Seno, and A. Maritan. A self-consistent knowledge-based approach to protein design. *Biophys J*, 80(1):480–90, 2001.
- [150] W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437(7058):579–83, 2005.
- [151] W. P. Russ and R. Ranganathan. Knowledge-based potential functions in protein design. *Curr Opin Struct Biol*, 12(4):447–52, 2002.
- [152] J. M. Sanchez, F. Ducastelle, and D. Gratias. Generalized cluster description of multicomponent systems. *Physica A*, 128(1-2):334–350, 1984.
- [153] H. Schrauber, F. Eisenhaber, and P. Argos. Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol*, 230(2):592–612, 1993.
- [154] 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(40):9343–52, 1990.

- [155] E. I. Shakhnovich. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.*, 7:29–40, 1997.
- [156] J. M. Shifman and S. L. Mayo. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc Natl Acad Sci U S A*, 100(23):13274–9, 2003.
- [157] S. S. Sidhu, G. D. Bader, and C. Boone. Functional genomics of intracellular peptide recognition domains with combinatorial biology methods. *Curr Opin Chem Biol*, 7(1):97–102, 2003.
- [158] K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34(1):82–95, 1999.
- [159] M. Singh and P. S. Kim. Towards predicting coiled-coil protein interactions. *Proceedings of the fifth annual international conference on Computational biology*, pages 279–286, 2001.
- [160] N. Singhal, C. D. Snow, and V. S. Pande. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.*, 121(1):415–425, 2004.
- [161] A. M. Slovic, H. Kono, J. D. Lear, J. G. Saven, and W. F. DeGrado. Computational design of water-soluble analogues of the potassium channel KcsA. *Proc. Natl. Acad. Sci. USA*, 101(7):1828–1833, 2004.
- [162] G. R. Smith and M. J. Sternberg. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, 12(1):28–35, 2002.
- [163] C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande. How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.*, 34:43–69, 2005.

- [164] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–8, 2005.
- [165] A. G. Street and S. L. Mayo. Pairwise calculation of protein solvent-accessible surface areas. *Fold Des*, 3(4):253–8, 1998.
- [166] M. D. Struthers, R. P. Cheng, and B. Imperiali. Economy in Protein Design: Evolution of a Metal-Independent bba Motif Based on the Zinc Finger Domains. *J. Am. Chem. Soc.*, 118(13):3073–3081, 1996.
- [167] A. Suenaga, M. Hatakeyama, M. Ichikawa, X. Yu, N. Futatsugi, T. Narumi, K. Fukui, T. Terada, M. Taiji, M. Shirouzu, S. Yokoyama, and A. Konagaya. Molecular Dynamics, Free Energy, and SPR Analyses of the Interactions between the SH2 Domain of Grb2 and ErbB Phosphotyrosyl Peptides. *Biochemistry*, 42(18):5195–5200, 2003.
- [168] C. M. Summa, M. M. Rosenblatt, J. K. Hong, J. D. Lear, and W. F. DeGrado. Computational de novo design, and characterization of an A(2)B(2) diiron protein. *J Mol Biol*, 321(5):923–38, 2002.
- [169] A. Szilgyi, V. Grimm, A. K. Arakaki, and J. Skolnick. Prediction of physical protein-protein interactions. *Phys Biol*, 2:S1–S16, 2005.
- [170] A. H. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–4, 2002.
- [171] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *J Biomol Struct Dyn*, 8(6):1267–89, 1991.

- [172] R. Tupler, G. Perini, and M. R. Green. Expressing the human genome. *Nature*, 409(6822):832–3, 2001.
- [173] Bowie J. U., Luthy R., and Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, 1991.
- [174] H. van Dam and M. Castellazzi. Distinct roles of Jun:Fos and Jun:ATF dimers in oncogenesis. *Oncogene*, 20:2453–2464, 2001.
- [175] A. van de Walle, M. Asta, and G. Ceder. The Alloy Theoretic Automated Toolkit: A user guide. *Calphad-Computer Coupling of Phase Diagrams and Thermochemistry*, 26(4):539–553, 2002.
- [176] A Van der Ven, M K Aydinol, and G Ceder. First-Principles Evidence for Stage Ordering in Li_xCoO_2 . *Journal of the Electrochemical Society*, 145(6):2149–2155, 1998.
- [177] A Van der Ven, M.K. Aydinol, G Ceder, G Kresse, and J Hafner. First principles investigation of phase stability in Li_xCoO_2 . *Phys. Rev. B*, 58(6):2975–2987, 1998.
- [178] M. Vendruscolo and E. Domany. Pairwise contact potentials are unsuitable for protein folding. *J Chem Phys*, 109(24):11101–11108, 1998.
- [179] C. Vinson, M. Myakishev, A. Acharya, A. A. Mir, J. R. Moll, and M. Bonovich. Classification of human B-ZIP proteins based on dimerization properties. *Mol Cell Biol*, 22(18):6321–35, 2002.
- [180] C. R. Vinson, P. B. Sigler, and S. L. McKnight. Scissors-grip model for DNA recognition by a family of leucine zipper proteins. *Science*, 246(4932):911–6, 1989.
- [181] C. L. Vizcarra and S. L. Mayo. Electrostatics in computational protein design. *Curr. Opin. Struct. Biol.*, 9:622–626, 2005.

- [182] M. Vsquez. An evaluation of discrete and continuum search techniques for conformational analysis of side chains in proteins. *Biopolymers*, 36(1):53–70, 1995.
- [183] J. Walshaw and D. N. Woolfson. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J Mol Biol*, 307(5):1427–50, 2001.
- [184] C. Wang, O. Schueler-Furman, and D. Baker. Improved side-chain modeling for protein-protein docking. *Protein Sci*, 14(5):1328–39, 2005.
- [185] L. Wernisch, S. Hery, and S. J. Wodak. Automatic protein design with all atom force-fields by exact and heuristic optimization. *J Mol Biol*, 301(3):713–36, 2000.
- [186] L. Wernisch, S. Hery, and S. J. Wodak. Automatic protein design with all atom force-fields by exact and heuristic optimization. *J Mol Biol*, 301(3):713–36, 2000.
- [187] U. Wiedemann, P. Boisguerin, R. Leben, D. Leitner, G. Krause, K. Moelling, R. Volkmer-Engert, and H. Oschkinat. Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J Mol Biol*, 343(3):703–18, 2004.
- [188] G. Williams. Least-Squares Curves. In S. Solomon, editor, *Linear Algebra with Applications*, pages 417–428. Jones and Bartlett Publishers, Inc., Boston, 5th edition, 2005.
- [189] E. Wolf, P. S. Kim, and B. Berger. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci*, 6(6):1179–89, 1997.
- [190] S. A. Wolfe, H. A. Greisman, E. I. Ramm, and C. O. Pabo. Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J Mol Biol*, 285(5):1917–34, 1999.

- [191] A. M. Wollacott and J. R. Desjarlais. Virtual interaction profiles of proteins. *J Mol Biol*, 313(2):317–42, 2001.
- [192] Z. Xiang and B. Honig. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol*, 311(2):421–30, 2001.
- [193] F. Xiaoran, J. Apgar, and A. Keating. Modeling backbone flexibility to achieve sequence diversity: The design of novel alpha-helical ligands for Bcl-xL. *J. Mol. Biol.*, 2007. submitted.
- [194] H. Yin, J. S. Slusky, B. W. Berger, R. S. Walters, G. Vilaire, R. I. Litvinov, J. D. Lear, G. A. Caputo, J. S. Bennett, and W. F. DeGrado. Computational Design of Peptides That Target Transmembrane Helices. *Science*, 315:1817–1822, 2007.
- [195] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTeraction database. *FEBS Lett*, 513(1):135–40, 2002.
- [196] X. Zeng, A. M. Herndon, and J. C. Hu. Buried asparagines determine the dimerization specificities of leucine zipper mutants. *Proc Natl Acad Sci U S A*, 94(8):3673–8, 1997.
- [197] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J Med Chem*, 48(7):2325–35, 2005.
- [198] F. Zhou, G. Grigoryan, S. R. Lustig, A. E. Keating, G. Ceder, and D. Morgan. Coarse-Graining Protein Energetics in Sequence Variables. *Phys. Rev. Lett.*, 95:148103, 2005.
- [199] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, and M. Snyder. Global analysis of protein activities using proteome chips. *Science*, 293(5537):2101–5, 2001.