Modeling the flexibility of alpha helices in protein interfaces: Structure-based design and prediction of helix-mediated protein-protein interactions

by

James R. Apgar

B.A. Chemistry
Williams College, 2001

SUBMITTED TO THE DEPARTMENT OF CHEMISTRY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2008

© 2008 Massachusetts Institute of Technology
All rights reserved

Signature of Author.....

James R. Apgar
Department of Chemistry
May 5, 2008

Certified by .....

Amy E. Keating
Department of Biology
Thesis Supervisor

Accepted by .....

Robert W. Field
Chairman, Departmental Committee on Graduate Students

ARCHIVES

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUN 10 2008
LIBRARIES

This doctoral thesis has been examined by a committee of the Department of Chemistry that includes:

Troy Van Voorhis .....
(Thesis Committee Chair)

Robert G. Griffin.....

Stuart S. Licht .....

Amy E. Keating .....
(Thesis Supervisor)

Modeling the flexibility of alpha helices in protein interfaces: Structure based design and prediction of helix-mediated protein-protein interactions

by

James R. Apgar

Submitted to the Department of Chemistry
on May 23, 2008 in Partial Fulfillment of the
Requirements for the Degree of Doctorate of Philosophy
in Chemistry

ABSTRACT

Protein-protein interactions play an essential role in many biological functions. Prediction and design of these interactions using computational methods requires models that can be used to efficiently sample structural variation. This thesis identifies methods that can be used to sample an important sub-space of protein structure: alpha helices that participate in protein interfaces. Helices, the global structural properties of which can be described with only a few variables, are particularly well suited for efficient sampling. Two methods for sampling helical backbones are presented: Crick parameterization for coiled coils and normal-mode analysis for all helices. These are shown to capture most of the variation seen in the PDB. In addition, these methods are applied to problems in protein structure prediction and design. Normal-mode analysis is used to design novel nanomolar peptide inhibitors of the apoptosis-related Bcl-2 family member, Bcl-x_L, and a modification of Crick Parameterization is used to predict the binding orientation of dimeric coiled coils with greater than 80% accuracy. Finally, this study addresses the increase in computational time required by flexible-backbone methods and the use of cluster expansion to quickly map structural energies to sequence-based functions for increased efficiency.

Thesis Supervisor: Amy E. Keating
Title: Associate Professor of Biology

For Grandpa and Baby Mack

Acknowledgements

My family has helped guide me through my life as well as my graduate career. My parents have always provided me with encouragement and consistent optimism, and my brother Josh has always given me thoughtful advice no matter how well he knows the subject. I would like to thank my Grandma for always asking me what I was up to, and my Aunt Sandy and Uncle Jackson for encouraging me to go for my goals no matter the challenges.

My graduate work would not have been possible without my advisor Amy Keating. I am very fortunate to have received her guidance and mentoring throughout the past six years. She has always been available for input and assistance, as well as giving me the encouragement to find my own way. I would like to thank my thesis committee members, Robert Griffin, Stuart Licht and Troy Van Voorhis for their helpful comments and suggestions. In particular, the many meetings with my thesis committee chair, Troy Van Voorhis, guided my progression through graduate school and helped me keep my goals in sight.

I am lucky to have worked in such a wonderful place as the Keating lab. Thanks to Orr Ashenberg, Scott Chen, Shaun Deignan, Sanjib Dutta, Kelly Elkins, Jeremy Fisher, Emiko Fire, Xiaoran Fu, Gevorg Grigoryan, Karl Gutwin, Seungsoo Hahn, Taijiao Jiang, Devdoot Majumdar, Aaron Reinke, Christy Taylor, Kevin Weston, and Nora Zizlsperger for your friendships and your help over the past few years. I would especially like to thank Gevorg Grigoryan for his help on almost every aspect of my computational work, as well as, for the many interesting conversations we had both related and unrelated to our research. I would also like to thank Christy Taylor for her friendliness and for encouraging me to join the lab. Finally, thanks to Emiko Fire, Nora Zizlsperger, Shaun Deignan and Jeremy Fisher for all the fun times we have had in and out of lab.

In the Keating lab I have had several productive collaborations involving most areas of my research. These included working with Xiaoran Fu on the Bcl-x_L inhibitor design, Karl Gutwin on the coiled-coil orientation prediction project, and Gevorg Grigoryan and Seungsoo Hanh on cluster expansion. Any successes I have had in my research would not have been possible without them.

Last but not least, I would like to thank my wonderful wife, Elizabeth, for her love, patience and support throughout my graduate career. As Alan Jackson says, "There is no reason to doubt her when she says I wouldn't last 10 minutes without her."

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Protein-protein interactions play a critical role in most biological functions. Investigations into these interactions raise many important questions such as: What are the sequence, structure and physicochemical properties that determine the function, stability or specificity of binding interactions? How do specific interactions fit into the entire cellular network? Tools that allow for the prediction of protein-protein interactions and design methods that can generate novel reagents to change or interfere with native interactions can provide a wealth of information to answer these types of questions. Structural models are often used in both of these pursuits as a way to obtain a detailed view of protein interfaces. They can provide complementary information that is often difficult or sometimes impossible to obtain through experimental techniques. By building accurate structural models, one can potentially extract the physical basis of interactions and understand how sequence and structure relate to biological function.

An important sub-class of protein interactions is alpha-helix mediated interfaces. Alpha helices are one of the two common secondary-structure elements found in proteins, thus understanding their role at protein interfaces can lead to insights into a wide variety of protein complexes. In my thesis work I focused on methods for modeling the structural variability of

alpha helices in protein interfaces, as well as on incorporating these structural models into methods for the design and prediction of protein-protein interactions.

In this chapter, I first review some of the biologically relevant interfaces that are mediated by alpha helices. I then discuss the use of discrete structural models of proteins and protein-protein interactions in design and prediction.

Helix-mediated interfaces

The structure of the alpha helix was first described in 1951 as a basic building block of proteins.[1] Alpha helices and beta sheets pack together to form the majority of the three-dimensional structure of proteins,[2] and also mediate a wide variety of protein-protein interactions. These structures include complexes of helical peptides with receptors, as well as assembly motifs such as coiled coils, helical bundles and helix-loop-helix (Figure 1-1).[3] Many important intermolecular interactions covering a diverse range of biological functions involve helix-mediated interfaces.

Coiled coils are commonly found as structural proteins, like myosin or tropomyosin, as well as oligomerization motifs. [3] These proteins are involved in many biological functions including the transcriptional regulation by bZIPs,[4] transmembrane signaling of the HAMP domain,[5] cellular membrane fusion mediated by the core structure of gp41 from the HIV envelope glycoprotein,[6] and the intercellular membrane fusion role of SNARE proteins.[7] Helical bundles are also a common oligomerization motif that occur as isolated domains like ROP,[8] LacR, tumor supresson,[9] p53,[10] and bacterial luciferase,[11] or as part of larger domains including histidine kinases,[12] apoliprotein E3,[13] granulocyte-macrophage colony-stimulation factor,[14] human growth hormone,[15] interleukin-4,[16] T4 lysoyzme[17] and 3-isopropylamatelate

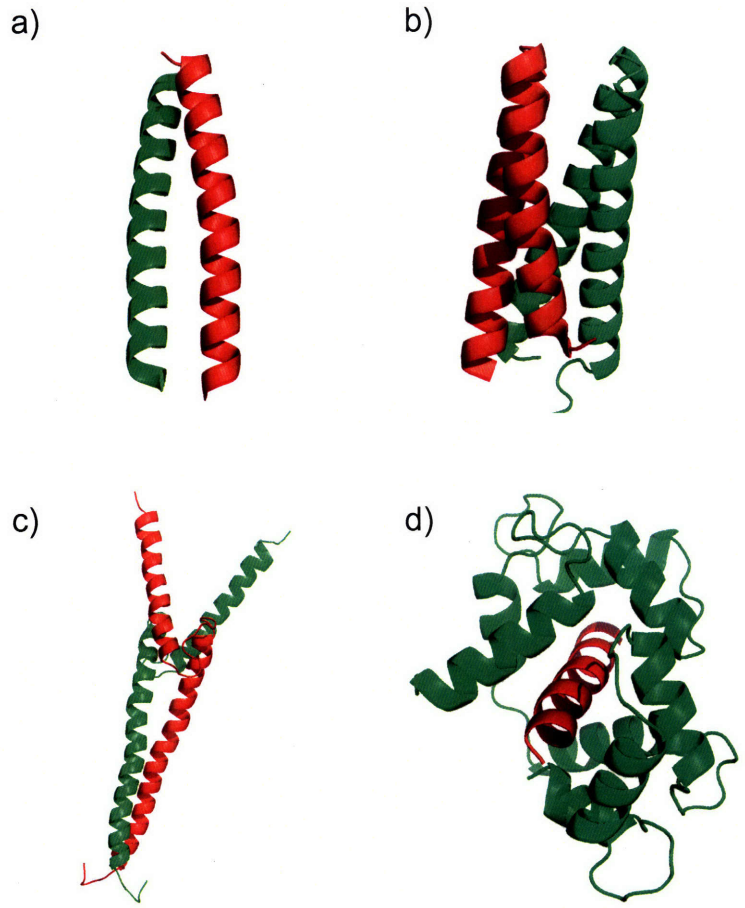


Figure 1-1: Structures of helix-mediate interfaces. (a) coiled coil, (b) helical bundle, (c) helix-loop-helix, and (d) helical-peptide interface.

dehydrogenase.[3,18] Helix-loop-helix domains are typically present as dimers, and bind with DNA as part of transcriptional regulation.[19]

There are also a number of interactions in which a helical peptide or a helical region of a larger protein mediates the interaction with a receptor. This type of interaction is involved in many areas of cellular regulation including apoptosis,[20-23] cell-cycle[24], inter-membrane trafficking, [25] proteolysis, [26,27] and the Ca^{2+}-dependent regulation of muscle contraction, inflammation, metabolism, intracellular movement, memory, nerve growth and immune response.[28,29] The receptor-binding component of many hormones is also an alpha-helical

element.[30-32] The activities of surface bound alpha helices are found in diverse categories of biological functions including the antimicrobial function of manganin[33] and cecropin[34], the channel-forming peptide alamethicin,[35,36] the channel-forming toxin colicin[37] and the membrane surface binding motif of prostaglandin H2 synthase-1.[38] Finally, helical elements are also important for membrane and vesicle fusions,[39] cell signaling,[40] recruitment of cellular components[41], and oligomerization.[42-44] In view of the myriad of interfaces that are mediated through alpha helices, accurate models of these interactions could offer a better understanding of a wide range of biological processes.

Discrete structural models

Modeling of proteins for computational design and large scale-prediction relies on the use of discrete structural models. The number of degrees of freedom available to a given protein is enormous. The additional complexity of searching over a large sequence space makes continuous models of structure, e.g. molecular dynamics, computationally unfeasible. By limiting the search to a finite set of conformations with discrete differences this problem becomes tractable. This type of modeling has yielded many great successes in computational protein design. Of particular note are experimentally verified designs such as stabilized native folds,[45-49] novel folds,[50-52] proteins with novel functions,[53,54] functionally active enzymes,[55-57] specific interacting protein pairs,[58-62] novel proteins that bind to native targets[63-67] and conformational switches.[68] These examples include several novel alpha-helical proteins designed for a range of biological applications. Novel helical peptides were designed to interact with native transmembrane integrin proteins.[63] Helical bundle proteins were designed as a model system to study conformational specificity.[69] Coiled-coil proteins were designed to bind to the helical portion of

the protein calcineurin.[67] A coiled-coil dimer was generated to be a mimetic of interleukin-4 and bound to its high affinity receptor IL-4Rα.[70] Novel heterodimeric coiled coils were also designed to specifically not homodimerize.[62] Finally, a non-native right-handed coiled coil was designed.[50,71,72] In addition to the design of the helical proteins, receptors have been designed to bind to helices, including the redesign of calmodulin to bind to a native helical target smMLCK.[64,65]

Discrete structural sampling for protein modeling was introduced in 1987 by Ponder and Richards. This work described modeling of protein cores with the simplifying approximation of fixing the protein backbone and sampling side-chain conformation using a discrete set of rotational isomers (rotamers).[73] This method of discrete structural sampling has been used for most computational protein design applications. The basic scheme for the process has been to identify a backbone conformation to use as a starting backbone template, most commonly from an x-ray crystal structure, and then, in the context of this fixed structure, the conformational space of a set side-chain rotamers is searched to find the lowest energy configuration. To facilitate this efficient structural search, the energy function needs to be pairwise decomposable.[74-79] This requires energy calculations to be broken down into component terms that include the interaction between pairs of side chains, the side chains with the backbone, and sometimes the backbone-backbone interactions.

Protein design is typically expressed as an optimization problem with the goal of finding sequences that maximize an objective function. The most common objective in protein design is to find a sequence that stabilizes a particular fold, calculated as the difference in energy between the folded and unfolded state. This involves searching through discrete structural conformations of the target's folded state and evaluating the energy using a pairwise decomposable energy

function. Thus, the energy functions must be able to be used to identify the stabilizing sequences for a particular backbone. In addition to finding a single stabilizing sequence, other design objectives have been explored that require modeling of more than one state. In this case, each state must be modeled with equivalent accuracy. This may require increased structural plasticity, including introducing flexibility into the protein backbone, as well as energy functions that can adequately capture the structural difference of sequences on different backbones. Below, I describe alternative design objectives that benefit from accurate models of multiple states and increased structural flexibility. I then introduce several current techniques that are used to integrate backbone flexibility into the discrete structural modeling framework. Finally, I discuss how different energy function components relate to discrete structural flexibility.

Design objectives

The most common goal in computational protein design experiments is to find a single sequence that stabilizes a particular fold. This is frequently done using a single fixed backbone structure and has had much success.[49,80-85] However, there are many biases associated with discrete structural models using a single fixed-backbone structure. Most notable is that only a small range of native-like sequences can be modeled on the backbone.[86-88] This limits the types of computational experiments that can be conducted because only a subset of sequences will have reasonable energies. However, it is possible to increase the sequence space of design through increases in the structure space. This has been shown in a few examples where the primary goal was to sample the sequence space compatible with a protein's fold.[89-92] Larson et al. included backbone flexibility to maximize the sequence based entropy from a single protein fold.[90] Saunders et al. showed that a flexible-backbone design was better than a fixed-backbone

design for modeling the sequence diversity of a protein family.[89] Finally, Wollacott et al. used flexible backbone design to explore the sequence profiles of peptides in several different protein-peptide complexes.[91] Inclusion of backbone flexibility was essential to these experiments because, for many sequences, a stable conformation of side chains could not be found using a single native backbone. Increasing the sequence space for design is important when stability is not the only design objective. With other objectives, additional constraints will restrict the sequence space. Therefore it is important to remove artificial constraints imposed by using a single fixed backbone.

Specificity is an important example of another design objective. This requires sequences that stabilize a particular state and avoid other states. To do this explicitly, models of the desired state and the undesired, or negative state, must both be modeled accurately. If the correct approximate energy cannot be determined from the negative state, then this will give false information about the specificity of the sequence. There have been several examples where specificity design has been done with explicit models of both the target and negative states.[61,62,93-95] Bolon et al. designed specific SspB heterodimers that avoided the homodimeric state.[93] Their work illustrated the importance of negative design since without it they selected mutants that both homo- and heterodimerized. In addition, these authors demonstrated the trade-off between increased specificity and loss of stability. This trade-off is indicative of the reduced sequence space that results when multiple constraints are imposed. Havranek et al. used negative design to generate coiled-coiled heterodimers that disfavored the homodimeric state as well as the aggregated state.[62] Here they built structural models for the desired heterodimers and the undesired homodimer, and an approximate model of the aggregated state. They achieved good specificity, but their predicted energy differences were not as large as what was determined

experimentally. The authors suspected that this was the result of over-estimation of the energy of the negative state due to the use of a fixed backbone.[62] Kortemme et al. also designed specific novel protein interactions of DNAase colicin E7 with its inhibitor, immunity protein I7, using negative design.[94] Their method introduced destabilizing mutations on each partner and then compensatory mutations to improve the other side of the interaction. This was effective in designing novel interacting pairs and relied on the ability to model both the desired pair as well as the destabilizing mutations. This method, called second-site mutation, was also explored in a follow-up paper that sought to introduce backbone flexibility into this design.[95] In that paper Joachimiak et al. used rigid-body rotation to generate an ensemble of starting structures. They found that this allowed for the design of more specific proteins partners, but not from the sequences that were selected using the second-site mutations. This indicates that they were not modeling the negative state as accurately as the desired target, and thus provided incorrect information to the design procedure. This result exemplifies the utility of including backbone flexibility in design, but also that care must be taken to generate accurate structural models for all states.

Backbone sampling methods

Protein backbones have many degrees of freedom, and sampling these efficiently in protein design is quite challenging.[92] To address this computational complexity, a number of approaches have been introduced that allow for the sampling of the backbone space in a manner that can be incorporated for use in protein design. Since the computational complexity of each fixed-backbone design is already quite large, these structural searches must be focused into high-probability regions of structure space.

The parameterization of several different types of protein structures has been introduced for protein design. This technique has been applied to coiled coils for the modeling of native templates,[96] including the variability of the native structure to better predict mutational energies,[97] the effects of core mutations on structure and oligomerization state[98] and the design of non-native right-handed coiled coils.[50] The design of symmetrical 4-helix di-iron proteins was enabled by sampling the starting template using three parameters for the displacement of the helical monomer and three for the relative orientation.[99,100] The design of a novel beta-barrel structure was generated by reducing the structural space to ten parameters that could be optimized to generate a low energy template.[101] A related approach has been used to vary the orientation of secondary structure elements in the α/β fold of the $\beta 1$ immunoglobulin-binding domain of streptococcal protein G.[48]

All of the methods described above introduced global deformations in the protein structure. Several groups have used discrete structural modeling to sample local deformations. The Baker group has had tremendous success modeling backbones in structure prediction by sampling from peptide fragments in the PDB.[102,103] They demonstrated that this approach is effective in protein design with numerous examples including a novel protein fold[51] and recapitulation of the native sequence space found in a protein familiy.[89] In another method for local structural sampling, Larson et al. used a Monte Carlo procedure to randomly sample the backbone ϕ and ψ angles and generate 'native-like' structure ensembles.[90] Wollacott et al. used the same method to sample the structure of peptides that were part of a protein complex.[91]

In addition to generating *de novo* structural deformations, native structures can also be used to model structural flexibility. Kono and Saven used NMR structure ensembles to represent possible backbone conformations.[104] Ali et al. generated models of a tetrameric BBA proteins

using two crystal structures that differed by a single mutation.[61] Yin et al. used multiple crystal structures of transmembrane proteins to generate an ensemble of protein interfaces for the design of peptides that bind specifically the integrin proteins.[63]

A number of docking methods have been introduced to generate protein-protein complexes.[105] A few of these methods show promise for discrete sampling in protein design. Huang and Mayo have shown that they can describe the relative orientation of monomers in a homodimer using a fast-Fourier transform method.[106] This allows for a rapid search of the relative position of the two structures that can be used as template candidates. This method was used to design a novel dimer from the β1 immunoglobulin-binding domain of streptococcal protein G.[107] Additionally, the Baker group has extended their protein fragment search to allow sampling of the relative orientation of different binding partners and internal structural flexibility simultaneously.[108]

Principal-component (PC) analysis is a common method used to capture structural flexibility in proteins. This method reduces the dimensionality of the structural space into a set of key deformations. This has been used with molecular dynamics to capture the structural changes in short time-scale simulations.[109-114] In addition, the range of structural variation amongst ensembles of conformations of a protein or homologous proteins can also be captured using this method.[115-118] In particular, the structural deviation of beta sheets and alpha helices can be captured using PC analysis.[119,120] Finally, Qian et. al. have shown that PC analysis can be useful for design by limiting the structural search to the space defined by a small set of principal-component vectors.[121]

The motion of proteins in molecular dynamic simulations can also be described in a reduced framework using normal-mode (NM) analysis. As described in a review by Ma,[122] a

small number of low-frequency normal modes can be used to model functionally important conformational transitions in several biomolecules that agree with motions observed in molecular dynamics simulations. It has also been noted that a significant amount of the variation seen among different crystal structures of the same, or closely related, proteins can be described by a small set of NM values.[123,124] Specifically for helices, Emberly et al. have shown that most of the deformation of the C_α–trace can be captured by three low-energy modes.[120]

Energy functions

Protein design relies on efficient methods for evaluating the energy of a structure. The energy functions, similar to those used in molecular mechanics, are often a summation of pairwise-energy terms that include van der Waals, electrostatics, hydrogen bonding and solvation.[125,126] Some terms like the van der Waals energy, can be easily broken down into pairwise contributions. However, other terms like electrostatics and solvation rely on multiple side-chain positions or different dielectric areas, and thus require additional approximations to make them pairwise decomposable. In addition, discrete structural sampling poses a particular problem for energy components like van der Waals and hydrogen bonding which are sensitive to the precise three-dimensional location of atoms. This is a result of the fact that the most favorable conformation may not be one of the discrete states sampled. Additionally, the hydrogen bonding term requires changes in charge distribution specific to each interacting pair to be truly modeled correctly. This is not present in discrete structures searches with fixed charges. A variety of approximations are made to allow these functions to be used for protein design with discrete structural sampling.

Core packing and hydrophobic burial are important parts of protein stability. These are typically captured using the van der Waals interaction energy, calculated with a 6-12 Lennard Jones potential.[126] This function is pairwise decomposable and is therefore well suited for computational design. However, due to the nature of the function, the van der Waals energy is very sensitive to the precise three-dimensional position of interacting atoms. Since the side chains sample discrete rotamer states, this often results in unrealistic clashes. This problem can be somewhat addressed by increasing the structural search space to include a larger number of rotamer conformations[127] or increased backbone space.[98,128] In addition, there are several modifications to the energy function that are made to avoid this problem. The most common of these is to scale the van der Waal radii of the atoms to 90%[61,62,129,130] or 95%[51,131] of their initial value. This prevents many clashes that occur using the rotamer approximation by softening the interaction. Other common modifications are to cap the magnitude of the repulsive term[80] or to replace the repulsive terms with a linear function that has a finite maximum at a radius of 0.[51,86] These modifications allow small clashes between side chains that would likely be relieved with slight relaxation to be included in the design solutions. Grigoryan et al. conducted a systematic examination of these types of approximations and demonstrated their strengths and weaknesses.[132] The authors discussed the fine balance between modifications that allow additional packing in the core and producing structures with unrealistic energies and geometries. Another type of modification is to use a solvent-accessible surface-area based term to describe the van der Waals energy. This is typically done by multiplying the amount of burial or hydrophobic burial by a constant term.[53,133] These values are often experimentally determined by fitting to small molecule data or scaled to reproduce mutational data.[133-135] These terms are less

sensitive to the precise structure and therefore work well with the discrete structure searches. However, these need to be offset by a repulsive term or the cores can become over-packed.

Proteins exist in a highly polarizable environment, therefore modeling the electrostatics correctly requires accurate models of solvation.[136] Protein design relies on implicit solvent models since explicit modeling of water would be too computationally expensive. A common solvent model used for biomolecules is finite difference Poisson Boltzmann (FDPB), which for continuum models is often considered as a standard for accuracy.[137,138] This and other continuum solvent models treat the protein as a low dielectric cavity surrounded by a high dielectric solvent. Generating this dielectric boundary requires the exact position of all atoms. This is a problem for protein design because the exact positions of side-chain atoms are not determined before the calculation is complete. Several models have been generated that break down the effect of the polarizable environment into interactions of side-chain pairs. These include distance dependent dielectric functions,[61,104,126] dielectric constants that are dependent on the protein structure,[133,139] statistical methods based on distributions of charged pairs in the PDB,[86,140] or pairwise screening described using the generalized Born equation.[141] Likewise, the solvation energy of a charged atom is also dependent on the dielectric environment. Approximations to this effect range from simple phenomenological models to detailed physical models. The simplest methods rely solely on the hydrophobic polar patterning seen in the core and surface positions of proteins.[142,143] Other statistical methods parameterize the solvation free energy as function of burial.[104] Several continuum solvent models have also been introduced for design including the approximation of the FDPB as a pairwise function,[136,144] Tanford-Kirkwood[145] and the approximations of the generalized Born method to allow for pairwise determination of Born radii.[146,147] Finally,

volume-based methods like EEF1[61,86,141,148] and surface area based methods[146,149] have also been used to approximate the solvation energy with reasonable results.

Hydrogen bonding is very common in proteins.[150] It is thought to play an important role in stabilizing native proteins[151-153] and provide specificity to proteins and protein-protein interactions.[154,155] However, these have been difficult to model with fixed-charge molecular-mechanics methods. Additionally, discrete structural models may not sample the precise three-dimensional geometry associated with hydrogen bonding pairs. Thus approximate methods have been introduced to assure that hydrogen bonding donors and acceptors are satisfied. These include a statistical orientation-dependent hydrogen bonding function that uses the native distribution of hydrogen bonding geometry in the PDB,[59] a rules-based geometric hydrogen bonding function that requires packed polar residues to form at least one hydrogen bond,[156] and an inventory-based hydrogen bonding function that weights the importance of hydrogen bonding in native interactions.[53] Like for van der Waals energy, increasing the rotamer library and allowing backbone flexibility will also generate more conformations with native-like hydrogen bonding geometry.

In addition to the physics based methods discussed above, experimental data has also been used to improve the energy function. This has been done by optimizing the relative magnitudes of terms to improve the performance of a particular function or replacing part of the calculation with experimentally validated values. Native sequence recovery in single-site design was used to optimize a composition-dependent unfolded state, as well as the relative magnitude of the components of the energy function.[86] In a similar approach, experimental binding data for single mutants has been used to optimize different energy components.[133] Finally, for cases when the structural models fail, it can be useful to replace some parts of the calculation with

experimental data. Grigoryan et al. described this as a way to improve the prediction of bZIP partnering. There they used experimentally determined coupling energies to replace core interactions that were modeled inaccurately.[141]

Many of the approximations of the energy function result from the necessity to be both pairwise decomposable and computationally inexpensive to allow for efficient structural searches. However, this requirement can be removed if the structure determination step and the energy evaluation step are split into two separate calculations.[58] These methods are implemented by using a low-resolution pairwise function to search the structure space and then selecting a subset of candidate conformations to re-evaluate using a more accurate high-resolution function. Thus, any non-pairwise energy functions could be used to determine the final energy solutions for design. Care must be taken that the subset of sequences that are re-evaluated with the high-resolution function contains the global minimum energy conformation.[58] In addition to different energy components, it is also possible to include continuous structure sampling methods at this step, like continuous minimization. This has been shown to be very helpful in relieving steric clashes associated with the van der Waals function.[132]

Summary of work

In the following chapters I discuss my work in developing methods for sampling the backbone structure space of proteins for use in design and prediction. Because the computational complexity associated with the sequence search is large, exploring the structure space in a rigorous manner, such as in molecular dynamics, would be too costly. However, more efficient structural searches that focus on a sub-space highly enriched with native-like conformations are acceptable. In my work, I have sampled alpha-helical structures at protein-protein interfaces to

search for alternate native-like conformations (Chapter 2). Using this approach I explore three applications: designing novel BH3 peptides that bind to Bcl-2 receptors (Chapter 3), predicting the binding orientation of dimeric coiled coils (Chapter 4), and exploring the use of cluster expansion for flexible-backbone energy evaluation (Chapter 5).

References

1. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 1951;37(4):205-211.
2. Chothia C, Levitt M, Richardson D. Structure of proteins: packing of alpha-helices and pleated sheets. Proc Natl Acad Sci U S A 1977;74(10):4130-4134.
3. Kohn WD, Mant CT, Hodges RS. Alpha-helical protein assembly motifs. J Biol Chem 1997;272(5):2583-2586.
4. O'Shea EK, Klemm JD, Kim PS, Alber T. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science 1991;254(5031):539-544.
5. Hulko M, Berndt F, Gruber M, Linder JU, Truffault V, Schultz A, Martin J, Schultz JE, Lupas AN, Coles M. The HAMP domain structure implies helix rotation in transmembrane signaling. Cell 2006;126(5):929-940.
6. Chan DC, Fass D, Berger JM, Kim PS. Core structure of gp41 from the HIV envelope glycoprotein. Cell 1997;89(2):263-273.
7. Antonin W, Fasshauer D, Becker S, Jahn R, Schneider TR. Crystal structure of the endosomal SNARE complex reveals common structural principles of all SNAREs. Nat Struct Biol 2002;9(2):107-111.
8. Predki PF, Agrawal V, Brunger AT, Regan L. Amino-acid substitutions in a surface turn modulate protein stability. Nat Struct Biol 1996;3(1):54-58.
9. Friedman AM, Fischmann TO, Steitz TA. Crystal structure of lac repressor core tetramer and its implications for DNA looping. Science 1995;268(5218):1721-1727.
10. Clore GM, Omichinski JG, Sakaguchi K, Zambrano N, Sakamoto H, Appella E, Gronenborn AM. High-resolution structure of the oligomerization domain of p53 by multidimensional NMR. Science 1994;265(5170):386-391.
11. Fisher AJ, Raushel FM, Baldwin TO, Rayment I. Three-dimensional structure of bacterial luciferase from Vibrio harveyi at 2.4 A resolution. Biochemistry 1995;34(20):6581-6586.
12. Marina A, Waldburger CD, Hendrickson WA. Structure of the entire cytoplasmic portion of a sensor histidine-kinase protein. Embo J 2005;24(24):4247-4259.
13. Wilson C, Wardell MR, Weisgraber KH, Mahley RW, Agard DA. Three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E. Science 1991;252(5014):1817-1822.
14. Diederichs K, Boone T, Karplus PA. Novel fold and putative receptor binding site of granulocyte-macrophage colony-stimulating factor. Science 1991;254(5039):1779-1782.

15. de Vos AM, Ultsch M, Kossiakoff AA. Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. Science 1992;255(5042):306-312.
16. Smith LJ, Redfield C, Smith RA, Dobson CM, Clore GM, Gronenborn AM, Walter MR, Naganbushan TL, Wlodawer A. Comparison of four independently determined structures of human recombinant interleukin-4. Nat Struct Biol 1994;1(5):301-310.
17. Rose DR, Phipps J, Michniewicz J, Birnbaum GI, Ahmed FR, Muir A, Anderson WF, Narang S. Crystal structure of T4-lysozyme generated from synthetic coding DNA expressed in Escherichia coli. Protein Eng 1988;2(4):277-282.
18. Imada K, Sato M, Tanaka N, Katsube Y, Matsuura Y, Oshima T. Three-dimensional structure of a highly thermostable enzyme, 3-isopropylmalate dehydrogenase of Thermus thermophilus at 2.2 A resolution. Journal of molecular biology 1991;222(3):725-738.
19. Murre C, McCaw PS, Baltimore D. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell 1989;56(5):777-783.
20. Sattler M, Liang H, Nettesheim D, Meadows RP, Harlan JE, Eberstadt M, Yoon HS, Shuker SB, Chang BS, Minn AJ, Thompson CB, Fesik SW. Structure of Bcl-x(L)-Bak peptide complex: Recognition between regulators of apoptosis. Science 1997;275(5302):983-986.
21. Petros AM, Nettesheim DG, Wang Y, Olejniczak ET, Meadows RP, Mack J, Swift K, Matayoshi ED, Zhang HC, Thompson CB, Fesik SW. Rationale for Bcl-x(L)/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. Protein Science 2000;9(12):2528-2534.
22. Liu XQ, Dai SD, Zhu YN, Marrack P, Kappler JW. The structure of a Bcl-x(L)/Bim fragment complex: implications for bim function. Immunity 2003;19(3):341-352.
23. Denisov AY, Chen G, Sprules T, Moldoveanu T, Beauparlant P, Gehring K. Structural Model of the BCL-w-BID Peptide Complex and Its Interactions with Phospholipid Micelles. Biochemistry 2006;45(7):2250-2256.
24. Schon O, Friedler A, Bycroft M, Freund SM, Fersht AR. Molecular mechanism of the interaction between MDM2 and p53. Journal of molecular biology 2002;323(3):491-501.
25. Zhu G, Zhai P, Liu J, Terzyan S, Li G, Zhang XC. Structural basis of Rab5-Rabaptin5 interaction in endocytosis. Nat Struct Mol Biol 2004;11(10):975-983.
26. Aguilar CF, Cronin NB, Badasso M, Dreyer T, Newman MP, Cooper JB, Hoover DJ, Wood SP, Johnson MS, Blundell TL. The three-dimensional structure at 2.4 A resolution of glycosylated proteinase A from the lysosome-like vacuole of Saccharomyces cerevisiae. J Mol Biol 1997;267(4):899-915.
27. Li M, Phylip LH, Lees WE, Winther JR, Dunn BM, Wlodawer A, Kay J, Gustchina A. The aspartic proteinase from Saccharomyces cerevisiae folds its own inhibitor into a helix. Nat Struct Biol 2000;7(2):113-117.
28. Vassylyev DG, Takeda S, Wakatsuki S, Maeda K, Maeda Y. Crystal structure of troponin C in complex with troponin I fragment at 2.3-A resolution. Proc Natl Acad Sci U S A 1998;95(9):4847-4852.
29. Meador WE, Means AR, Quiocho FA. Target enzyme recognition by calmodulin: 2.4 A structure of a calmodulin-peptide complex. Science 1992;257(5074):1251-1255.
30. Kaiser ET, Kezdy FJ. Amphiphilic secondary structure: design of peptide hormones. Science 1984;223(4633):249-255.

31. Beck-Sickinger AG, Jung G. Structure-activity relationships of neuropeptide Y analogues with respect to Y1 and Y2 receptors. Biopolymers 1995;37(2):123-142.
32. Moreli MAC, Gons N, Temussi PA. Biochemistry 1989;28:7996-8002.
33. Bechinger B. Structure and functions of channel-forming peptides: magainins, cecropins, melittin and alamethicin. J Membr Biol 1997;156(3):197-211.
34. Gazit E, Miller IR, Biggin PC, Sansom MS, Shai Y. Structure and orientation of the mammalian antibacterial peptide cecropin P1 within phospholipid membranes. Journal of molecular biology 1996;258(5):860-870.
35. Cafiso DS. Alamethicin: a peptide model for voltage gating and protein-membrane interactions. Annu Rev Biophys Biomol Struct 1994;23:141-165.
36. Sansom MS. Structure and function of channel-forming peptaibols. Q Rev Biophys 1993;26(4):365-421.
37. Cramer WA, Heymann JB, Schendel SL, Deriy BN, Cohen FS, Elkins PA, Stauffacher CV. Structure-function of the channel-forming colicins. Annu Rev Biophys Biomol Struct 1995;24:611-641.
38. Picot D, Loll PJ, Garavito RM. The X-ray crystal structure of the membrane protein prostaglandin H2 synthase-1. Nature 1994;367(6460):243-249.
39. Zhu G, Zhai P, He X, Wakeham N, Rodgers K, Li G, Tang J, Zhang XC. Crystal structure of human GGA1 GAT domain complexed with the GAT-binding domain of Rabaptin5. Embo J 2004;23(20):3909-3917.
40. Dajani R, Fraser E, Roe SM, Yeo M, Good VM, Thompson V, Dale TC, Pearl LH. Structural basis for recruitment of glycogen synthase kinase 3beta to the axin-APC scaffold complex. Embo J 2003;22(3):494-501.
41. Kamada K, Roeder RG, Burley SK. Molecular mechanism of recruitment of TFIIF-associating RNA polymerase C-terminal domain phosphatase (FCP1) by transcription factor IIF. Proc Natl Acad Sci U S A 2003;100(5):2296-2299.
42. Sixma TK, Kalk KH, van Zanten BA, Dauter Z, Kingma J, Witholt B, Hol WG. Refined structure of Escherichia coli heat-labile enterotoxin, a close relative of cholera toxin. Journal of molecular biology 1993;230(3):890-918.
43. Sixma TK, Pronk SE, Kalk KH, Wartna ES, van Zanten BA, Witholt B, Hol WG. Crystal structure of a cholera toxin-related heat-labile enterotoxin from E. coli. Nature 1991;351(6325):371-377.
44. Terrak M, Wu G, Stafford WF, Lu RC, Dominguez R. Two distinct myosin light chain structures are induced by specific variations within the bound IQ motifs-functional implications. Embo J 2003;22(3):362-371.
45. Calhoun JR, Kono H, Lahr S, Wang W, DeGrado WF, Saven JG. Computational design and characterization of a monomeric helical dinuclear metalloprotein. Journal of molecular biology 2003;334(5):1101-1115.
46. Slovic AM, Kono H, Lear JD, Saven JG, DeGrado WF. Computational design of water-soluble analogues of the potassium channel KcsA. Proceedings of the National Academy of Sciences of the United States of America 2004;101(7):1828-1833.
47. Ogata K, Jaramillo A, Cohen W, Briand JP, Connan F, Choppin J, Muller S, Wodak SJ. Automatic sequence design of major histocompatibility complex class I binding peptides impairing CD8+ T cell recognition. J Biol Chem 2003;278(2):1281-1290.
48. Su A, Mayo SL. Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Science 1997;6(8):1701-1707.

49. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82-87.
50. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282(5393):1462-1467.
51. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302(5649):1364-1368.
52. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ, DeGrado WF. Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. Journal of the American Chemical Society 2005;127(5):1346-1347.
53. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature 2003;423(6936):185-190.
54. Mena MA, Treynor TP, Mayo SL, Daugherty PS. Blue fluorescent proteins with enhanced brightness and photostability from a structurally targeted library. Nat Biotechnol 2006;24(12):1569-1571.
55. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, Dechancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. Nature 2008.
56. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D. De novo computational design of retro-aldol enzymes. Science 2008;319(5868):1387-1391.
57. Bolon DN, Mayo SL. Enzyme-like proteins by computational design. Proceedings of the National Academy of Sciences of the United States of America 2001;98(25):14274-14279.
58. Green DF, Dennis AT, Fam PS, Tidor B, Jasanoff A. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. Biochemistry 2006;45(41):12547-12559.
59. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. Journal of molecular biology 2003;326(4):1239-1259.
60. Chevalier BS, Kortemme T, Chadsey MS, Baker D, Monnat RJ, Stoddard BL. Design, activity, and structure of a highly specific artificial endonuclease. Molecular Cell 2002;10(4):895-905.
61. Ali MH, Taylor CM, Grigoryan G, Allen KN, Imperiali B, Keating AE. Design of a heterospecific, tetrameric, 21-residue miniprotein with mixed alpha/beta structure. Structure 2005;13(2):225-234.
62. Havranek JJ, Harbury PB. Automated design of specificity in molecular recognition. Nature Structural Biology 2003;10(1):45-52.
63. Yin H, Slusky JS, Berger BW, Walters RS, Vilaire G, Litvinov RI, Lear JD, Caputo GA, Bennett JS, DeGrado WF. Computational design of peptides that target transmembrane helices. Science 2007;315(5820):1817-1822.
64. Shifman JM, Mayo SL. Exploring the origins of binding specificity through the computational redesign of calmodulin. Proceedings of the National Academy of Sciences of the United States of America 2003;100(23):13274-13279.

65. Shifman JM, Mayo SL. Modulating calmodulin binding specificity through computational protein design. Journal of molecular biology 2002;323(3):417-423.
66. Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, Schwab MS, Serrano L, Gonzalez C. Computer-aided design of a PDZ domain to recognize new target sequences. Nature Structural Biology 2002;9(8):621-627.
67. Ghirlanda G, Lear JD, Lombardi A, DeGrado WF. From synthetic coiled coils to functional proteins: automated design of a receptor for the calmodulin-binding domain of calcineurin. Journal of molecular biology 1998;281(2):379-391.
68. Ambroggio XI, Kuhlman B. Computational design of a single amino acid sequence that can switch between two distinct protein folds. J Am Chem Soc 2006;128(4):1154-1161.
69. Hill RB, Raleigh DP, Lombardi A, DeGrado WF. De novo design of helical bundles as models for understanding protein folding and function. Acc Chem Res 2000;33(11):745-754.
70. Domingues H, Cregut D, Sebald W, Oschkinat H, Serrano L. Rational design of a GCN4-derived mimetic of interleukin-4. Nat Struct Biol 1999;6(7):652-656.
71. Sales M. FitCC Personal Communication with Tom Alber; http://ucxray.berkeley.edu/~mark/fitcc.html; 2007.
72. Plecs JJ, Harbury PB, Kim PS, Alber T. Structural test of the parameterized-backbone method for protein design. Journal of molecular biology 2004;342(1):289-297.
73. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of molecular biology 1987;193(4):775-791.
74. Desmet J, Demaeyer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 1992;356:539-542.
75. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophysical Journal 1994;66(5):1335.
76. Gordon DB, Mayo SL. Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure 1999;7(9):1089-1098.
77. Lasters I, De Maeyer M, Desmet J. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. Protein Engineering 1995;8(8):815-822.
78. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins 1998;33(2):227-239.
79. Pierce NA, Spriet JA, Desmet J, Mayo SL. Conformational splitting: A more powerful criterion for dead-end elimination. J Comput Chem 2000;21:999-1009.
80. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. Protein Sci 1995;4(10):2006-2018.
81. Lazar GA, Desjarlais JR, Handel TM. De novo design of the hydrophobic core of ubiquitin. Protein Sci 1997;6(6):1167-1178.
82. Dahiyat BI, Sarisky CA, Mayo SL. De novo protein design: towards fully automated sequence selection. Journal of molecular biology 1997;273(4):789-796.
83. Malakauskas SM, Mayo SL. Design, structure and stability of a hyperthermophilic protein variant. Nat Struct Biol 1998;5(6):470-475.
84. Bryson JW, Desjarlais JR, Handel TM, DeGrado WF. From coiled coils to small globular proteins: design of a native-like three-helix bundle. Protein Sci 1998;7(6):1404-1414.
85. Hellinga HW. Computational protein engineering. Nat Struct Biol 1998;5(7):525-527.

86. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.
87. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. Protein Science 2000;9:1106-1119.
88. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332(2):449-460.
89. Saunders CT, Baker D. Recapitulation of protein family divergence using flexible backbone protein design. Journal of molecular biology 2005;346(2):631-644.
90. Larson SM, England JL, Desjarlais JR, Pande VS. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. Protein Science 2002;11(12):2804-2813.
91. Wollacott AM, Desjarlais JR. Virtual interaction profiles of proteins. Journal of molecular biology 2001;313(2):317-342.
92. Butterfoss GL, Kuhlman B. Computer-based design of novel protein structures. Annual Review of Biophysics and Biomolecular Structure 2006;35:49-65.
93. Bolon DN, Grant RA, Baker TA, Sauer RT. Specificity versus stability in computational protein design. Proc Natl Acad Sci U S A 2005;102(36):12724-12729.
94. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D. Computational redesign of protein-protein interaction specificity. Nature Structural & Molecular Biology 2004;11(4):371-379.
95. Joachimiak LA, Kortemme T, Stoddard BL, Baker D. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. Journal of molecular biology 2006;361(1):195-208.
96. Crick FH. The Fourier Transform of a Coiled-Coil. Acta Cryst 1953;6:685-689.
97. Keating AE, Malashkevich VN, Tidor B, Kim PS. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. Proc Natl Acad Sci U S A 2001;98(26):14825-14830.
98. Harbury PB, Tidor B, Kim PS. Repacking Protein Cores with Backbone Freedom - Structure Prediction for Coiled Coils. Proceedings of the National Academy of Sciences of the United States of America 1995;92(18):8408-8412.
99. Summa CM, Lombardi A, Lewis M, DeGrado WF. Tertiary templates for the design of diiron proteins. Curr Opin Struct Biol 1999;9(4):500-508.
100. North B, Summa CM, Ghirlanda G, DeGrado WF. D(n)-symmetrical tertiary templates for the design of tubular proteins. Journal of molecular biology 2001;311(5):1081-1090.
101. Offredi F, Dubail F, Kischel P, Sarinski K, Stern AS, Van de Weerdt C, Hoch JC, Prosperi C, Francois JM, Mayo SL, Martial JA. De novo backbone and sequence design of an idealized alpha/beta-barrel protein: evidence of stable tertiary structure. Journal of molecular biology 2003;325(1):163-174.
102. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.
103. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309(5742):1868-1871.
104. Kono H, Saven JG. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. Journal of molecular biology 2001;306(3):607-628.

105. Vajda S, Camacho CJ. Protein-protein docking: is the glass half-full or half-empty? Trends Biotechnol 2004;22(3):110-116.
106. Huang PS, Love JJ, Mayo SL. Adaptation of a fast Fourier transform-based docking algorithm for protein design. J Comput Chem 2005;26(12):1222-1232.
107. Czabotar PE, Lee EF, van Delft MF, Day CL, Smith BJ, Huang DC, Fairlie WD, Hinds MG, Colman PM. Structural insights into the degradation of Mcl-1 induced by BH3 domains. Proc Natl Acad Sci U S A 2007;104(15):6217-6222.
108. Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. Journal of molecular biology 2007;373(2):503-519.
109. Amadei A, Linssen ABM, Berendsen HJC. Essential Dynamics of Proteins. Proteins-Structure Function and Genetics 1993;17(4):412-425.
110. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal component analysis and long time protein dynamics. J Phys Chem-Us 1996;100(7):2567-2572.
111. Garcia AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett 1992;68(17):2696-2699.
112. Horiuchi T, Go N. Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. Proteins 1991;10(2):106-116.
113. Kitao A, Hirata F, Go N. The Effects of Solvent on the Conformation and the Collective Motions of Protein - Normal Mode Analysis and Molecular-Dynamics Simulations of Melittin in Water and in Vacuum. Chem Phys 1991;158(2-3):447-472.
114. van Aalten DM, Amadei A, Linssen AB, Eijsink VG, Vriend G, Berendsen HJ. The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. Proteins 1995;22(1):45-54.
115. Theobald DL, Wuttke DS. Accurate Structural Correlations from Maximum Likelihood Superpositions. PLoS Comput Biol 2008;4(2):e43.
116. Velazquez-Muriel JA, Carazo JM. Flexible fitting in 3D-EM with incomplete data on superfamily variability. J Struct Biol 2007;158(2):165-181.
117. Velazquez-Muriel JA, Valle M, Santamaria-Pang A, Kakadiaris IA, Carazo JM. Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies. Structure 2006;14(7):1115-1126.
118. Alber F, Forster F, Korkin D, Topf M, Sali A. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. Annu Rev Biochem 2008.
119. Emberly EG, Mukhopadhyay R, Tang C, Wingreen NS. Flexibility of beta-sheets: Principal component analysis of database protein structures. Proteins-Structure Function and Bioinformatics 2004;55(1):91-98.
120. Emberly EG, Mukhopadhyay R, Wingreen NS, Tang C. Flexibility of alpha-helices: Results of a statistical analysis of database protein structures. Journal of molecular biology 2003;327(1):229-237.
121. Qian B, Ortiz AR, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proceedings of the National Academy of Sciences of the United States of America 2004;101(43):15346-15351.
122. Ma J. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. Structure 2005;13(3):373.
123. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. Biophysical Journal 2005;88(2):1291.

124. Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Engineering 2001;14(1):1-6.
125. Park S, Yang X, Saven JG. Advances in computational protein design. Curr Opin Struct Biol 2004;14(4):487-494.
126. Gordon DB, Marshall SA, Mayo SL. Energy functions for protein design. Curr Opin Struct Biol 1999;9(4):509-513.
127. Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci 2004;13(3):735-751.
128. Desjarlais JR, Handel TM. Side-chain and backbone flexibility in protein core design. Journal of molecular biology 1999;290(1):305-318.
129. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. Proc Natl Acad Sci U S A 1997;94(19):10172-10177.
130. Lassila JK, Keeffe JR, Oelschlaeger P, Mayo SL. Computationally designed variants of Escherichia coli chorismate mutase show altered catalytic activity. Protein Eng Des Sel 2005;18(4):161-163.
131. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. Journal of molecular biology 2003;332(2):449-460.
132. Grigoryan G, Ochoa A, Keating AE. Computing van der Waals energies in the context of the rotamer approximation. Proteins 2007;68(4):863-878.
133. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of molecular biology 2002;320(2):369-387.
134. Radzicka A, Wolfenden R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry 1988;27(5):1664-1670.
135. Hellinga HW, Richards FM. Optimal sequence selection in proteins of known structure by simulated evolution. Proc Natl Acad Sci U S A 1994;91(13):5803-5807.
136. Vizcarra CL, Zhang N, Marshall SA, Wingreen NS, Zeng C, Mayo SL. An improved pairwise decomposable finite-difference Poisson-Boltzmann method for computational protein design. J Comput Chem 2008;29(7):1153-1162.
137. Feig M, Onufriev A, Lee MS, Im W, Case DA, Brooks CL, 3rd. Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem 2004;25(2):265-284.
138. Baker NA. Improving implicit solvent simulations: a Poisson-centric view. Curr Opin Struct Biol 2005;15(2):137-143.
139. Wisz MS, Hellinga HW. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins 2003;51(3):360-377.
140. Liang S, Grishin NV. Effective scoring function for protein sequence design. Proteins 2004;54(2):271-281.
141. Grigoryan G, Keating AE. Structure-based prediction of bZIP partnering specificity. Journal of molecular biology 2006;355(5):1125-1142.
142. Wei Y, Kim S, Fela D, Baum J, Hecht MH. Solution structure of a de novo protein from a designed combinatorial library. Proc Natl Acad Sci U S A 2003;100(23):13270-13273.

143. Walsh ST, Cheng H, Bryson JW, Roder H, DeGrado WF. Solution structure and dynamics of a de novo designed three-helix bundle protein. Proc Natl Acad Sci U S A 1999;96(10):5486-5491.
144. Marshall SA, Vizcarra CL, Mayo SL. One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. Protein Sci 2005;14(5):1293-1304.
145. Havranek JJ, Harbury PB. Tanford-Kirkwood electrostatics for protein modeling. Proc Natl Acad Sci U S A 1999;96(20):11145-11150.
146. Pokala N, Handel TM. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. Protein Sci 2004;13(4):925-936.
147. Qiu D, Shenkin PS, Hollinger FP, Still WC. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximated Born Radii. Journal of Physical Chemistry A 1997;101(16):3005-3014.
148. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35(2):133-152.
149. Street AG, Mayo SL. Pairwise calculation of protein solvent accessible surface areas. Folding and Design 1998;3:253-258.
150. Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 1984;44(2):97-179.
151. Doig AJ, Williams DH. Why water-soluble, compact, globular proteins have similar specific enthalpies of unfolding at 110 degrees C. Biochemistry 1992;31(39):9371-9375.
152. Shirley BA, Stanssens P, Hahn U, Pace CN. Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. Biochemistry 1992;31(3):725-732.
153. Murphy KP, Gill SJ. Solid model compounds and the thermodynamics of protein unfolding. J Mol Biol 1991;222(3):699-709.
154. Petrey D, Honig B. Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci 2000;9(11):2181-2191.
155. Lumb KJ, Kim PS. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. Biochemistry 1995;34(27):8642-8648.
156. Bolon DN, Marcus JS, Ross SA, Mayo SL. Prudent modeling of core polar residues in computational protein design. Journal of molecular biology 2003;329(3):611-622.

Chapter 2

Parameterized methods for modeling alpha-helix flexibility in native proteins

Excerpts included with permission of John Wiley & Sons, Inc. from:

Apgar, J.R., Gutwin, K.N., Keating. A.E. (2008) "Predicting helix orientation for coiled-coil dimers" *Proteins, In Press*

Excerpts included with permission of Elsevier B.V. from:

Fu, X., Apgar, J.R., Keating. A.E. (2007) "Modeling backbone flexibility to achieve sequence diversity: The design of novel alpha-helical ligands for Bcl-xL" *J. Mol. Bio.* **31**, 1099-1117

Collaborators Notes:

Karl Gutwin generated the coiled-coil test set.

The alpha helix is a basic building block of many protein structures.[1] As described in the previous Chapter, helical elements mediate many protein-protein interactions in a wide variety of biological processes. Accurate structural models of alpha helices are therefore an important part of understanding many types of protein interfaces. The backbone conformation of alpha helices is localized in selected regions of the Ramachandran ϕ,ψ space,[2] thus many methods treat these elements as rigid when sampling the structure.[3-9] However, treatment of the helix as a fixed element does not capture all of the deformation that occurs naturally, as evidenced by the distribution in the Protein Databank (PDB) (See below). Modeling the flexibility of these abundant structural elements is an important part of structure prediction, as well as of protein design and prediction of protein interactions. In this chapter, I examine several methods for capturing the structural variability of helices found in the PDB and demonstrates how these methods can be used to generate helices that span the native structure space. These methods are normal-mode analysis and principal component analysis to determine the flexibility of single helices, and expanded Crick parameterization to model flexibility in parallel and antiparallel dimeric coiled coils.

Alpha helices

The stability of an alpha helix (E_H) can be thought of as a sum of internal factors and external factors. The internal factors include the self energies of the backbone (E_{BB}) and side chains (E_{SC}), and the backbone-side chain ($E_{BB\text{-}SC}$) and side chain-side chain ($E_{SC\text{-}SC}$) interaction energies. The external factors include the interaction of the backbone ($E_{BB\text{-}R}$) or side-chains ($E_{SC\text{-}R}$) with other partners, and the interaction of the helix with the solvent ($E_{H\text{-}S}$).

$$E_H = E_{BB} + E_{SC} + E_{BB-SC} + E_{SC-SC} + E_{BB-R} + E_{SC-R} + E_{H-S} \qquad (2\text{-}1)$$

A balance between these terms determines how favorable it is for a peptide to be in the helical state.[10-12] A similar balance of effects is also thought to dictate smaller variations of alpha-helical structure.[13] The specific geometry of the alpha-helix backbone is important, allowing for a network of hydrogen bonds[12] and strong van der Waals contacts[14] to overcome the loss of backbone entropy.[10-12] This, in combination with the stabilizing effects of some side chains[11,15,16] or pairs of side chains,[17,18] could resist deformations. In contrast, any destabilizing side chains[11,15-18] should make the helix more prone to deformations. Finally, stabilizing protein-protein interactions can offset the loss of stability associated with bending. Assuming that one can approximate the low energy backbone conformation with an ideal helix, I hypothesize that the deviations from this structure are dictated by the sequence and surrounding environment.

In the PDB, helices are found with many different sequences in a wide range of environments. By examining all the helices, it is possible to sample the large and somewhat random distributions of environments, and thus approximate the relative distribution of possible deformations. To accomplish this I generated a set of helical structures from the PDB and examined the types and distribution of deformations that occur. I then examined two different methods to capture the structural variability: normal-mode analysis and principal-component analysis.

Helix database

To probe the structural variation of helices in the PDB, I extracted over 45,000 protein fragments of at least 15 consecutive residues with ϕ and ψ angles in the range of -50^{o} +/- 30^{o} from x-ray crystal structures with resolution of 2.5 Å or better. Helices of each length were then aligned to ideal helices permitting the deformation to be captured using the difference between

the N-C_α-C backbone of the target and the ideal helix. The structures in the database were all close to the ideal helix with the average root mean square deviation (rmsd) ranging from 0.45 +/- 0.14 to 1.1 +/- 0.38 Å for helices of length 15 to 30 respectively. As these helices were all part of larger structures, I also evaluated their degree of burial. The mean degree of burial was 67.6% with a standard deviation of 14.5 %. This was relatively similar for helices of all lengths, with the mean ranging from 62.4% to 70.9%, and the range was not correlated to the length of the helix.

I speculated that the helical deformation was dictated by a balance between the stability of the ideal structure and other features including the sequence and the environment. Therefore I investigated whether for alpha helices in the database the sequence composition or degree of burial were significantly correlated with structural deformation, as defined as the rmsd of the each alpha helix with respect to an ideal helix. To determine the degree of structural deformation, I broke the helices into sets by length and calculated the mean rmsd for each set. Using these data, the set of all helices was subdivided into two groups: more bent than the average from its own length set (larger rmsd), or less bent than the average. I then performed a χ^2 test to see if there was a statistically significant difference for each of these properties.

The results for the sequence composition are shown in Figure 2-1, along with experimentally determined helix propensities.[16] Helix propensities, or the intrinsic helix forming tendencies of a side-chain, are thought to strongly determine the stability of a helix.[11,15-17,19] χ^2 analysis reveals that a number of different amino-acid types were enriched in helices that differed significantly from the mean (p-value < 10^{-3}). Ala, Cys, Glu, Lys, Gln and Arg correlated with significantly less deformation while Phe, Gly, Pro, Ser, Thr, Trp and Tyr correlated with significantly more deformation. For some amino acid types the degree of bending correlates with

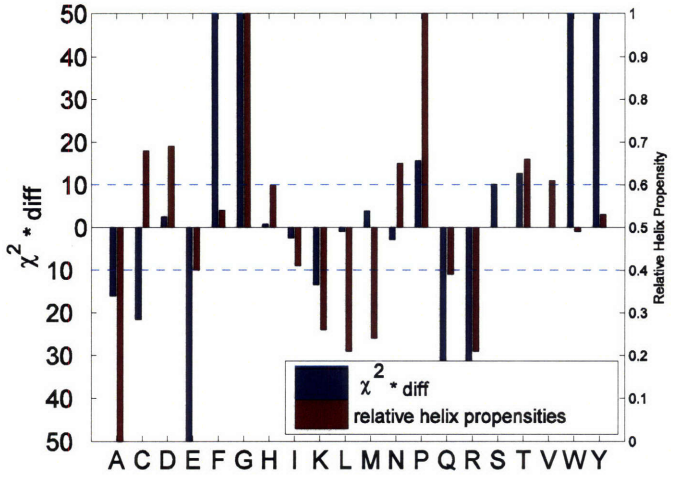


Figure 2-1: Effect of amino-acid type on deformation of alpha helices. Using the database of alpha helices of length 15 to 30, a χ^2 test was conducted to see if helices that were more or less bent than the mean were more enriched with any amino acid type. Plotted here on the left axis and in blue are the χ^2 values multiplied by 1 if there is a higher proportion of that amino-acid in bent helices and -1 if there is a smaller proportion. The dashed lines indicate significant values with p-scores < 10^{-3}. In the right axis and in red are helix propensities.[16]

the helix propensities. For example, Pro and Gly, which strongly disfavor the helical state, and Thr, which weakly disfavors the helical state, are all more prevalent in bent helices. Other amino acids are less bent, like Ala, Glu, Lys, Gln, and Arg, and have helix propensities that suggest they stabilize the helical structure. Trp, Tyr and Phe are all more prevalent in bent helices, but do not follow a helix propensity trends. However they are all large residues and their size could cause disruption to the helix because they are more likely to have interactions with external structures. These results are consistent with my hypothesis that the degree of deformation is a tradeoff between the energy of being in the ideal helical state and outside influences. The observation that residues that stabilize the helix more are more prevalent in straighter helices and those that are bulky or destabilize the helix are found more often in bent helices, supports this view. The two exceptions to this observation are the amino acids Ser and Cys. Ser is more prevalent in bent helices, though it does not favor or disfavor the helical state, and Cys strongly disfavors the helical state but is found more often in straighter helices.

In addition to sequence composition, I also compared the degree of burial to the probability of a helix being more bent. χ^2 analysis revealed that these two variables were not independent (p-value of 3×10^{-7}). This suggests that being more buried correlates with being more bent. If we assume that on average the more buried a helix is, the larger number of interactions it has with other parts of the protein, this result is also consistent with the hypothesis that external forces compensate for the bending of a helix.

The database was also used to investigate the types of helical interfaces present in the PDB. All multi-chain structures in which a helix was part of the interface (at least one residue within 5 Å of a different chain) were isolated. These structures were then filtered to look for cases where the helix made up more than 50% of interface residues for one chain. A total of 170 structures fit these criteria. Figure 2-2 illustrates the types of interfaces found in this subset of structures.

Normal-mode analysis

Normal-mode (NM) analysis has been widely recognized as a way to model functionally important conformational changes in biomolecules.[20-23] It is also speculated that it might provide an effective strategy for modeling the backbone variation in a protein fold as the sequence changes.[20,24] NM analysis can generate basis vectors that allow for sampling all 3N-6 internal degrees of freedom of any structure with N atoms, but the mode space required to accomplish this is prohibitively large. However, if the number of modes that contribute to significant structural deviations is small, NM analysis could provide a highly efficient way of sampling non-local conformational change. Emberly et al.[13] demonstrated this potential for alpha helices in a small set of 399 structures. Their results suggest NM analysis is a promising method to sample

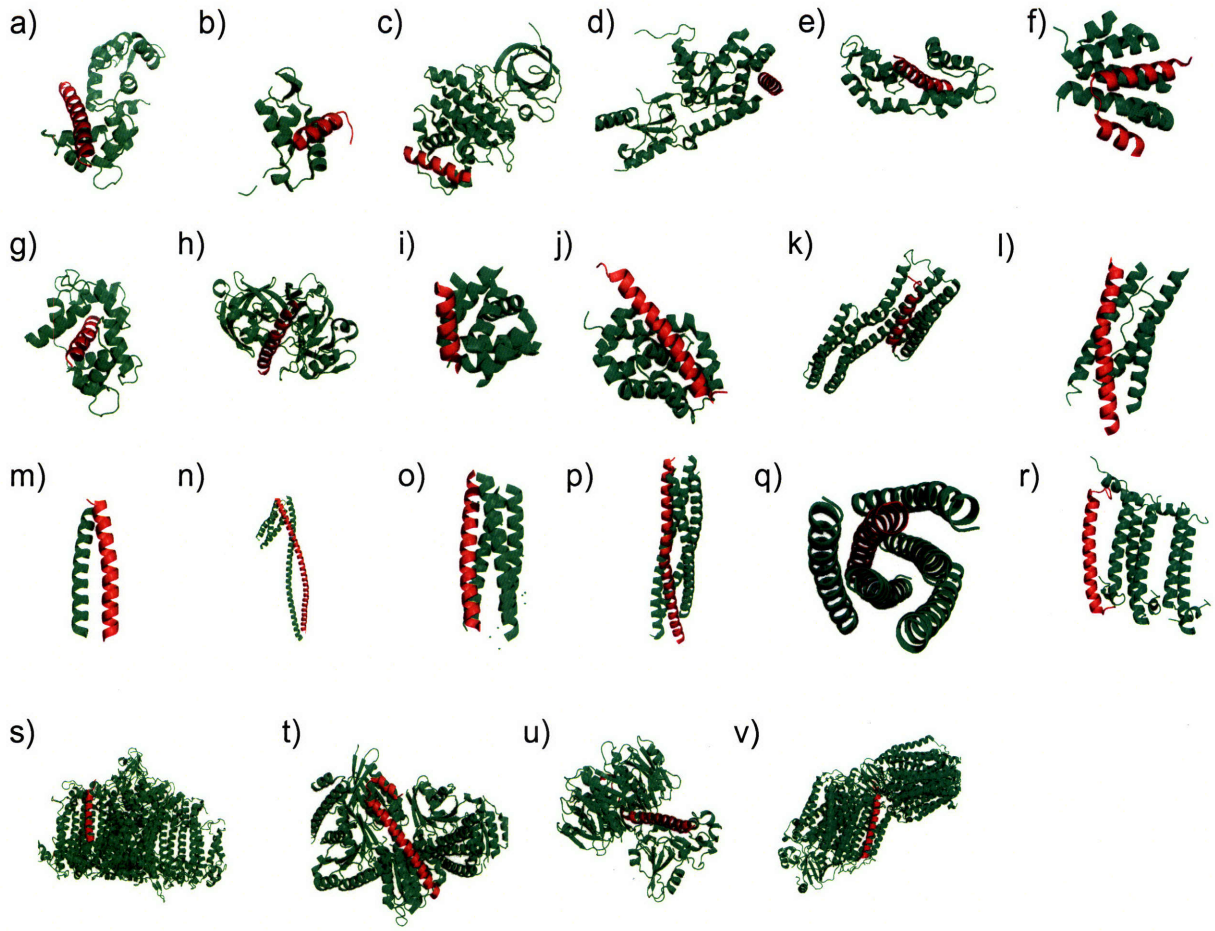


Figure 2-2: Representative helix-mediated protein-protein interfaces in the PDB. All structures were solved by x-ray diffraction with a resolution better than 2.5 Å. These are collected such that the helix, shown in red, contains at least 50% of all interface residues for its chain. The structures represent the types of interfaces found through this search method. Listed are the PDB ids and a description of the interacting partners. (a) 1ax2: Interaction of troponin C with helix fragment from troponin I. (b) 1j2x: Interaction of transcription initiation factor IIF with helix from RAP74 subunit of transcription factor IIF. (c) 1o9u: Interaction of glycogen synthase kinase-3 beta with helical axin peptide. (d) 1t0j: Interaction between voltage-gated calcium subunit beta-2A and the helical peptide of the voltage dependent L-type calcium channel alpha-1C subunit. (e) 1m45: Interaction of myosin light chain with helical MYO2P class V myosin. (f) 1g39: HNF-1alpha dimerization domain. (g) 1cdl: Calmodulin in complex with a helical peptide from smooth muscle myosin light chain kinase. (h) 1dp5: Interaction of proteinase A with the helical proteinase inhibitor IA3. (i) 1l4x: Homo-octomeric de novo designed helix. (j) 1pq1: Interaction of Bcl-2 family member Bcl-x_L with helical Bim-BH3 peptide. (k) 1syq: Interaction of viniculin with helical talin. (l) 1vzj: Four fold helical interaction of a WWW motif with a left handed polyproline helix (m) 2zta: Coiled-coil leucine zipper that is part of a yeast transcriptional activator. (n) 1x79: Human GGA1 GAT domain complexed with the helical GAT-binding domain of Rabaptin5. (o) 2bni: Antiparallel four-helix bundle. (p) 1g12: Coiled coil of endosomal SNARE core complex. (q) 2siv: Helical SIV gp41 core structure. (r) 1kzu: Helical chain of the integral membrane peripheral light harvesting complex in R. Acidophila. (s) 1jb0: Interaction between photosystem I and its helical subunit PSAX. (t) 1tu3: Interaction of the RAB5 complex with the helical Rabaptin5 c-terminal domain. (u) 1ltg: Heat-labile enterotoxin complex mediated through a helical domain. (v) 1m56: Helical subunit of cytochrome C oxidase interacting with total complex.

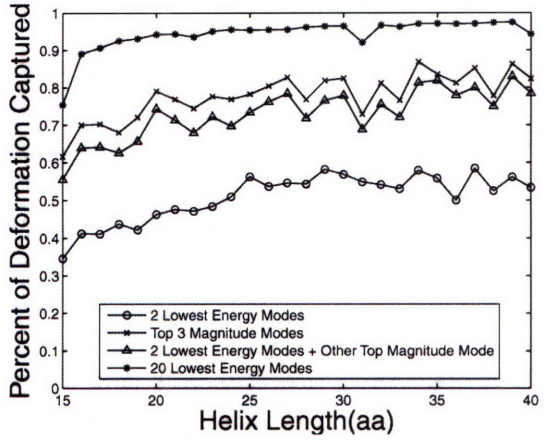
the structural deformations associated with sequence changes for alpha-helical segments, and possibly other structures, in protein design calculations.

Previously, Emberly et al. used the C_α-backbone trace to generate normal modes and fit these to existing protein structures. However, it is quite difficult to predict all backbone atoms from the C_α-trace, as the best methods only fit the main-chain atoms to 0.4 to 0.6 Å rmsd.[25] Here I report on the use of NM analysis to generate deformations associated with the C_α, C and N-backbone atoms of helical peptides. This 3-atom method has an advantage because the C_α, C and N atoms are positioned explicitly, leaving no ambiguity in the construction of the backbone.

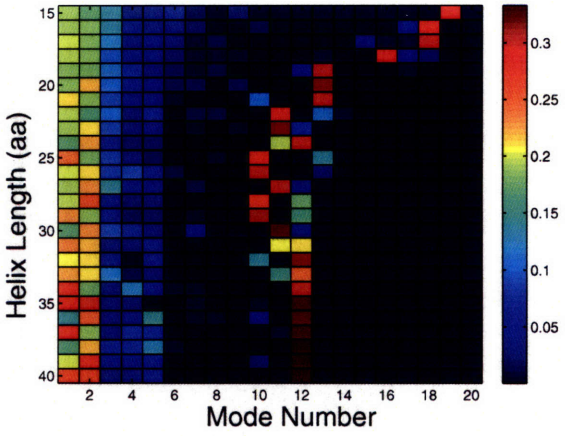
The generation of normal modes is described in detail in the Methods section of this Chapter. Briefly, an ideal helix, generated using CHARMM, serves as the zero point energy structure for these calculations. A simplified distance-dependent function (Eqn. 2-4 in Methods) was used to calculate the harmonic potential. This function has been shown to capture low energy normal modes accurately by Tirion[26] and was later used by Tama et al.[21] The Hessian matrix of this potential was then diagonalized to determine the eigenvectors and eigenvalues. The eigenvectors are the normal modes and the eigenvalues are the frequencies. The deviation of any helix from the ideal helix is then fit to a linear combination of normal modes as described by Equation 2-6 in the Methods.

With this procedure I utilized the test set of helices described above to examine the normal modes that correspond to deformations of helices seen in the PDB. Among these structures, the two normal modes with the lowest frequencies (modes 1 and 2), along with one other mode, captured on average ~70% of the total deformation (Figure 2-3a and 2-3b). In addition, of the three modes with the largest contributions, modes 1 or 2 occurred in the top 3 40-50% of the time. Most importantly, for helices of a given length, modes 1 and 2 had the largest

(a)



(b)



(c)

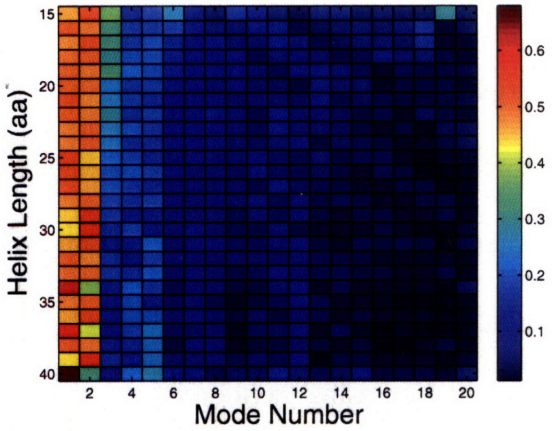


Figure 2-3: Capturing the structural variation of alpha helices using normal modes. A large number of helices with lengths between 15 and 40 residues were extracted from the PDB and analyzed to generate panels a-c (see Methods). (a) Shows the fractional deviation from an ideal-helix geometry that can be described with the indicated normal modes, as a function of the length of the helices being modeled. Circles indicate the two lowest energy modes, crosses the top three magnitude modes, triangles the two lowest energy modes plus the next largest magnitude mode, and asterisks the top 20 modes. (b) Indicates the fractional contribution of modes 1-20 and (c) shows the normalized standard deviation of this contribution. In (b), the color indicates the fraction of occurrences in which a mode is among the top 3 magnitude modes. In (c), the color indicates the standard deviation of normal mode magnitudes across all helices of a given length. Standard deviations are normalized so that the total sum of standard deviations for a given length helix sums to 1.

standard deviations over structures (Figure 2-3c), illustrating that these modes encompassed most of the variability and are good candidates to sample structure space.

For both the C_α and the N-C_α-C backbone, three normal modes made up the largest component of deformation:[13] the single order X-Z and Y-Z bending modes and the helical twisting mode. However, there was a difference in the ordering of relative energy of these modes as compared to the other low-energy normal modes. When describing the backbone using the C_α trace, these three modes were the lowest in energy. For the N-C_α-C backbone, the helical twist mode was only the 10^{th} to 18^{th} lowest energy mode, depending on the length of the helix. This difference could be explained by a more degenerate normal-mode space and the larger total number of normal modes that describe these structural changes.

Sampling structures using NMA

For the N-C_α-C backbone, the two lowest energy normal modes described most of the variability in the normal-mode space. This allowed for backbone flexibility to be included in modeling of helical interfaces with greatly reduced dimensionality. For this type of sampling, I first needed to determine the range of deformations common for the two normal modes. To do this, I fit the deformations of these modes to a normal distribution using the following equation:

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2\text{-}2)$$

where x is the mode value, μ is the mean and σ is the standard deviation. These values and their 95% confidence intervals are shown in Figure 2-4. For neither mode was the value of the mean always zero. However, this value oscillated around zero depending on the length of helices in each set. This was probably due to the small sample size for each set of helices. Additionally, the standard deviation increased as the length of the helix increased. This increase is present because

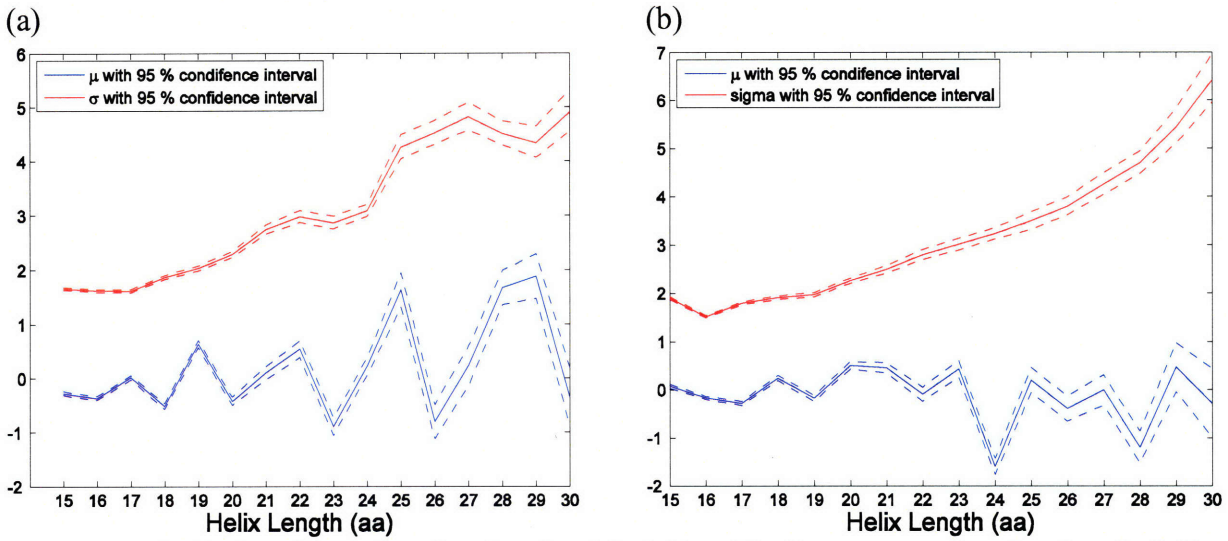


Figure 2-4: Distribution of normal-mode values for alpha helices. The first two normal mode values for helices of length 15 to 30 were fit to normal distributions. The mean (μ) and the standard deviation (σ) are shown with their 95% confidence interval for (a) NM-1 and (b) NM-2. Finally, the first two normal mode values are independently sampled based upon a distribution p (Eqn. 2-3).

the energy required to bend a longer chain the same amount, based on total displacement, as a shorter chain is less, allowing the structure to vary more.

Given these observations, structures can be varied using the following procedure. First, the value of the mean is assumed to be 0 for helices of all lengths. The value of σ is then selected for the appropriate length helix based on a linear fit to the values shown in Figure 2-4. Next, a linear combination of normal modes is fit to the starting structure to determine the native values. This is determined by multiplying a normal distribution associated with the native normal-mode space with a user defined distribution centered on the mode value from the starting structure.

$$p(x \mid \mu, \sigma, m, s) = f(x \mid 0, \sigma) \cdot f(x \mid m, s) \tag{2-3}$$

Here f is defined in Equation 2-2, x is a new normal mode parameter value, μ and σ are the mean and standard deviation from the set of native structures, m is the normal mode value of the starting structure, and s is user defined control parameter. If s is set to a small value, the structures will sample a space that is close to the native mode value of the starting structure

(Figure 2-5a). If s is large, a wide mode space will be sampled, but the space will remain confined to the range of native values (Figure 2-5b). Using the complete set of normal mode parameters, a linear combination of normal mode vectors is added to the ideal helix to generate the new helix. This will completely describe the backbone of the structure.

Principal-component analysis

Principal component (PC) analysis is a mathematical method used to reduce the dimensionality of a data set into its largest contributing components. Like NM analysis, PC analysis has been used to capture the common structural changes in proteins. Examples include capturing dynamical modes during short-time-scale simulation[22,23,27-30] and capturing the components of structural changes between ensembles or sets of homologous proteins.[31-34] Additionally, the deviations of secondary structure elements can be captured using PC analysis.[13,35] The significant difference between NM analysis and PC analysis is that NM analysis describes structural changes by approximating motions using a harmonic potential, while PC analysis describes common deformations by the covariance of different atoms in a set of protein structures. The former uses an external reference that is independent of the test set, while the latter is dependent on the test set and therefore could be biased by it. Here I used the same test set of structures described above to determine if PC analysis could be used to capture the deviation of the N-C_α-C backbone efficiently.

The principal components were determined by calculating the covariance of the protein atoms with respect to the average position of those atoms (Eqn 2-7 in Methods). This average was determined by aligning all helices of a given length helix to the ideal helix and then the average of each coordinate for each atom calculated. The covariance matrix was then

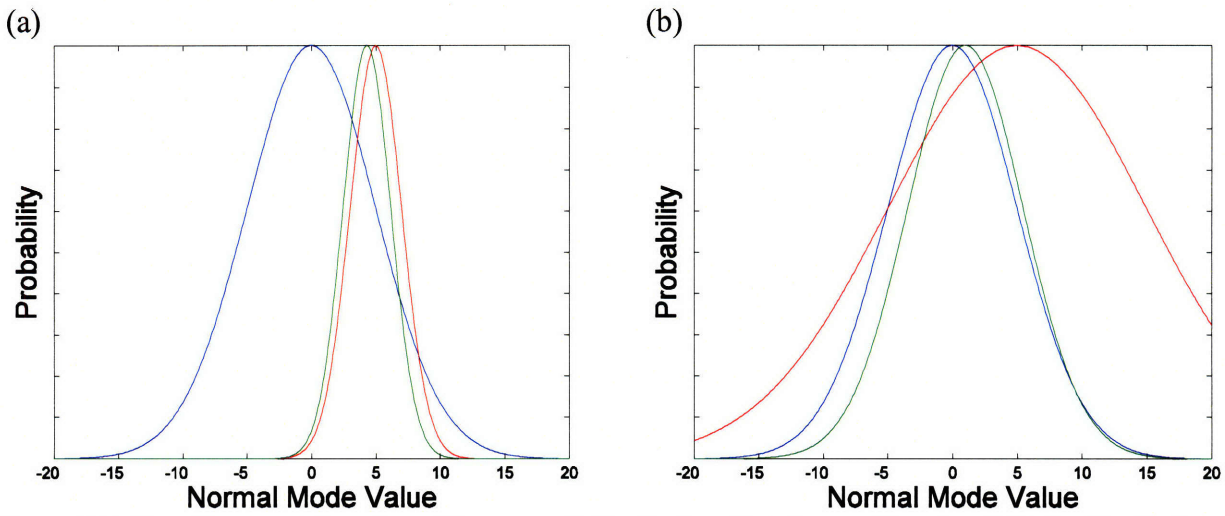


Figure 2-5: Sampling normal modes with a normal distribution. Based upon Equation 2-3, the distribution of normal-mode values sampled (green) is determined by multiplying the normal distribution of the native normal-mode space (shown in blue) by a user defined distribution (shown in red). The width of this distribution is determined by a control variable **s**, which is shown either (a) set to 2 and (b) set to 10.

diagonalized. The resultant eigenvectors were the set of orthogonal principal-components vectors and the eigenvalues indicated how much each component contributed to the overall distributions of deformations. The normalized contribution of all components for helices of each length is shown Figure 2-6. This figure illustrates how many components are required to fit the deformation in the data set. For helices of length 15 to 40 amino acids, the top component captured 40-90% of the deformation, the top two captured 70-95% and the top three captured virtually all the deformation.

Comparison between PC analysis and NM analysis

PC analysis and NM analysis each capture the major types of deformations associated with the changes in helical structure in the PDB. However, PC analysis can capture a significantly larger percentage of the total deformation than NM analysis, using the same number of modes. Given this altered performance between the two methods, there must be some difference between the types of deformation that these large contributing modes capture. This

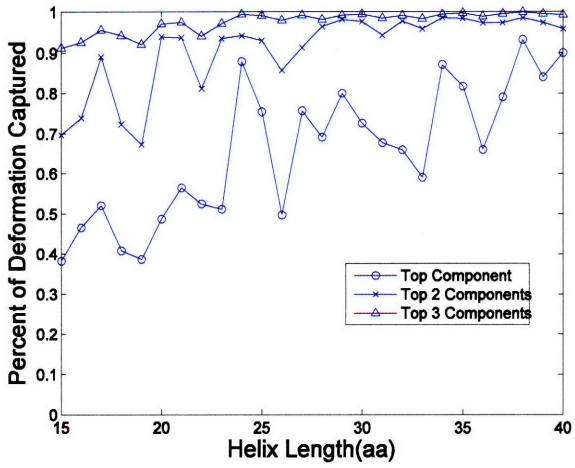


Figure 2-6: Capturing the structural variation of alpha helices using principal-component analysis. Helices of lengths 15 to 40 residues were extracted from the PDB and analyzed to show the fractional deviation from ideal-helix geometry that can be described with the indicated principal components. Circles are the largest component, crosses the largest two, and triangles the largest three.

was examined by comparing the structural difference between the PC analysis and NM analysis eigenvectors. To do this, a linear combination of normal modes was fit to each PC to determine its composition. Since only a few PCs made up most of the deformation, I only looked at the four largest component modes (Figure 2-7). Figure 2-7a, shows that for most helices, the first principal component (PC-1) is made up of the XZ- or YZ-bends from normal modes 1 and 2 (NM-1 and NM-2). However, for helices of length 17, and 24-27 these modes are not the largest contributors. Instead, these modes are the same twisting modes that were seen for the normal mode fit of deformation (Figure 2-3). This difference in normal-mode composition of PC-1 may be more the result of using differing test sets for each length helix than an actual feature of the structural space. PC-2 is almost entirely composed of combinations of the bending modes. The first two principal components make up 70-95% of the total deformations, and are composed primarily of the three types of normal modes that were seen previously. The next two components, PC-3 and PC-4, are mainly composed of other modes and make up the remaining 1

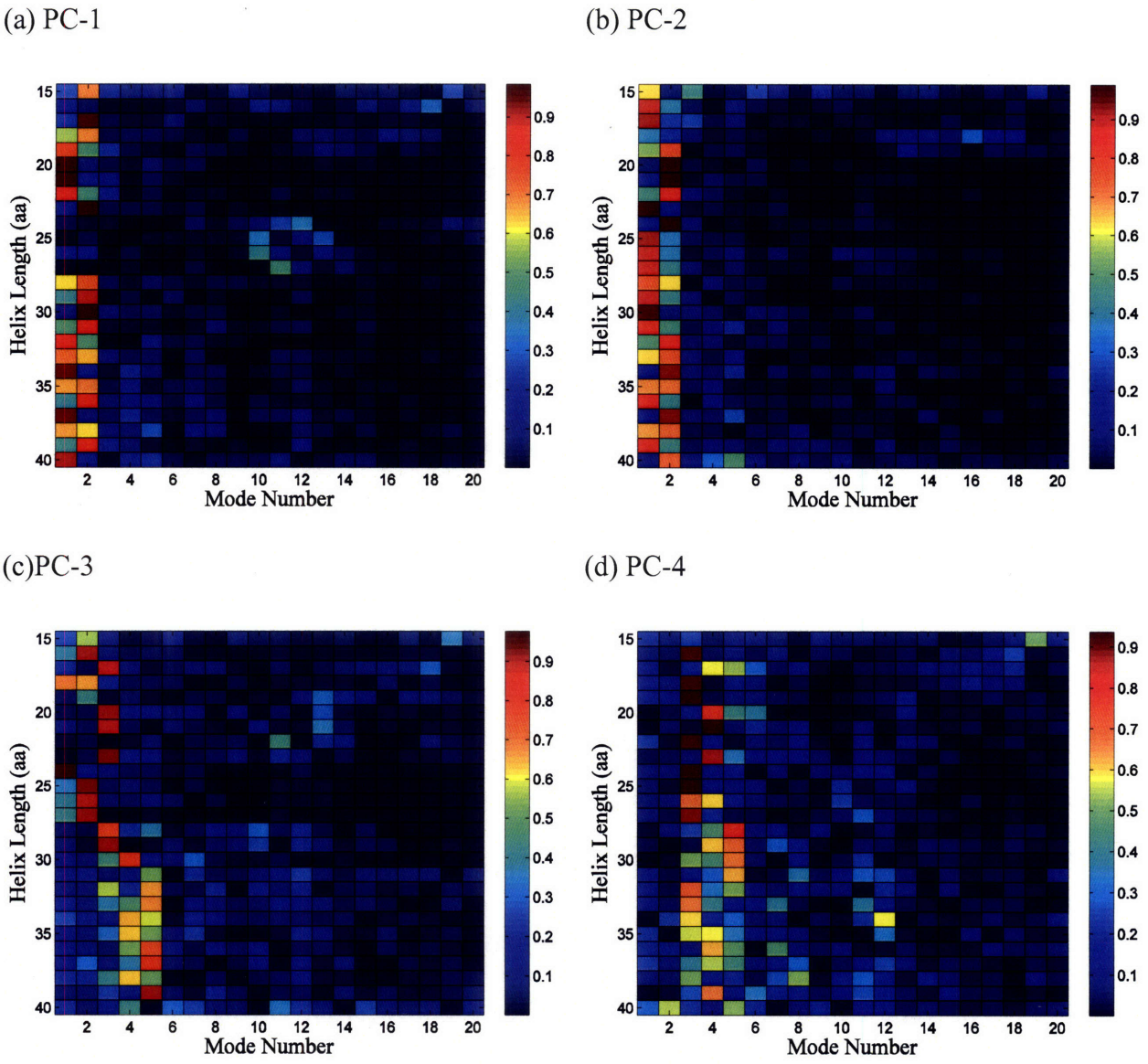


Figure 2-7: Comparison of principal-component analysis and normal-mode analysis for capturing the deformation of helices. For helices of length 15 to 40 a linear combination of normal modes are fit to the first four principal components (PC-1-4), shown as (a-d) respectively. The color indicates the percentage of the PC for that length helix captured by a particular NM.

to 25% of the deformation. These modes sample other higher order normal-mode bends, but are still primarily the lower energy deformations.

Both PC analysis and NM analysis appear to be reasonable methods to sample the structure space of helices. The three modes described for NM analysis seem like the most

common deformation modes for helices and compare well with the PCs. The top PC modes captured a larger fraction of the structural deformation, but there appears to be a strong test set bias. This is seen in the large helix-length dependence on the types of NMs whose linear combinations make up PC vectors. Since the PC vectors must be generated for helices of each length, it will be difficult to determine if the PC deformations are real or just represent a biased non-random set of helices. As a result, NM analysis is likely to be a less biased method to sample the helical structure of a protein. An application of NM analysis for modeling helical proteins that interact with the Bcl-2 family member Bcl-x_L is described in Chapter 3.

Coiled coils

The coiled-coil structure consists of a bundle of supercoiled helices that are encoded by a 7-residue sequence repeat of the form [**abcdefg**]$_n$. With **a** and **d** positions hydrophobic and **e** and **g** positions usually polar or charged, a "sticky" stripe winds its way around an individual helix, dictating the formation of a twisted helical bundle. This structure was first proposed by Crick in 1952.[36] Soon afterwards he described a method whereby the structure of the C_α trace of a parallel coiled coil can be described using as reduced set of parameters: an interhelical radius R_0, a superhelical twist ω_0, and a helical offset parameter ϕ (Figure 2-8).[37] These three parameters described the symmetrical interaction of the two helices about the superhelical axis. In 1995 this method was updated for the use of modern computers and implemented as a user energy function in the molecular mechanics software package CHARMM.[38] The function generates an ideal set of C_α atoms based upon Crick parameters and then imposes an energetic constraint proportional to the distance squared to the target structure. It allows for both the optimization of the Crick parameters to identify those that best fit a known structure, as well as the generation of a parallel

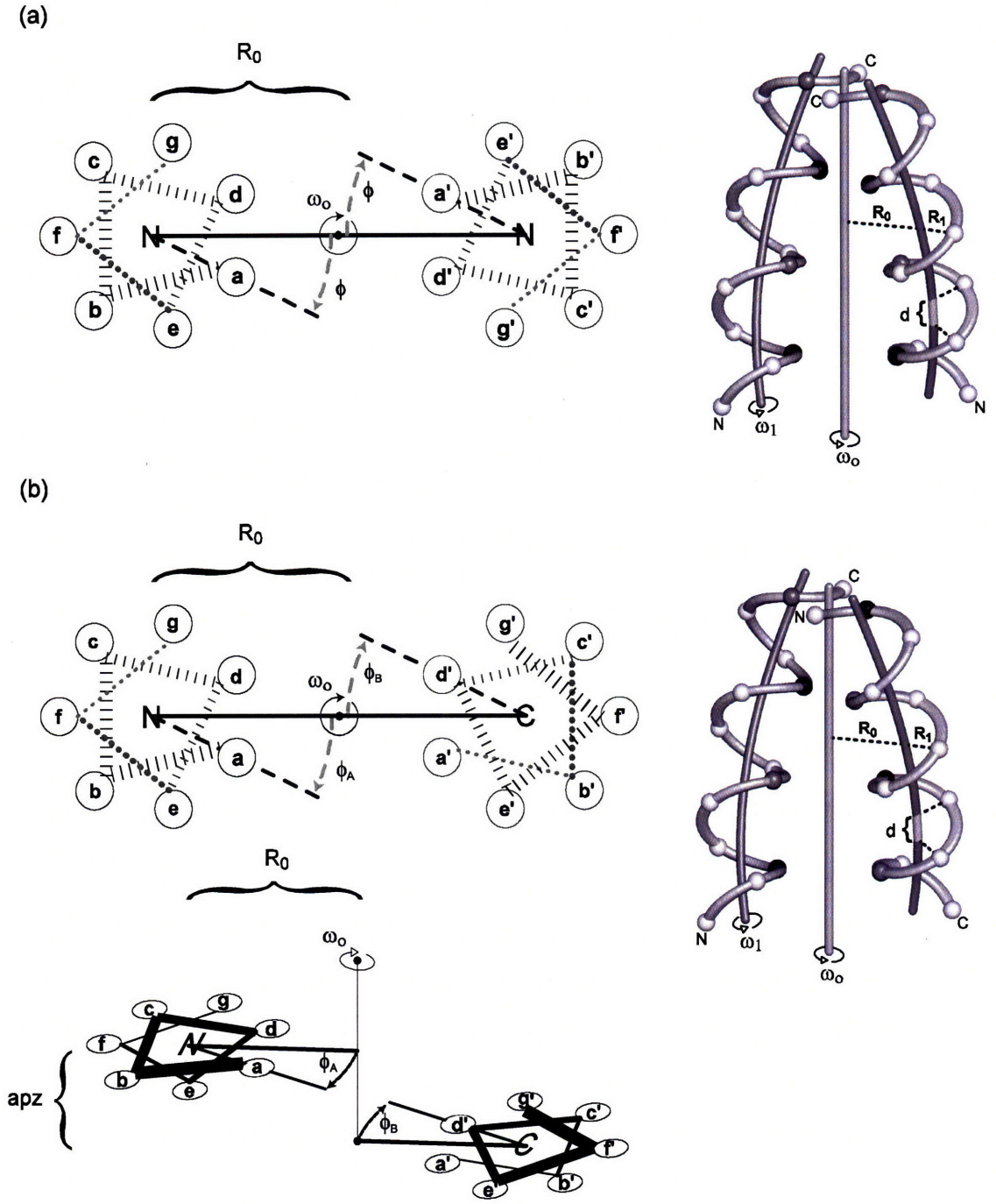


Figure 2-8: Crick parameterization of parallel and antiparallel coiled coils. (a-b) Schematic illustrating parameters used to describe (a) parallel and (b) antiparallel backbone geometries. For each wheel diagram, the heptad positions are indicated in lowercase letters and the direction of the chain is indicated by whether the N or C terminus is out of the page. For the structural diagram, the **a** and **a'** positions are shown in black, the **d** and **d'** positions in gray, and the rest in white

coiled coil with arbitrary parameter values. This method has been used for the design and predication of native-like coiled coils,[38,39] as well as the design of a non-native right handed coiled coil.[40-42]

A test set of parallel dimeric coiled-coil structures was generated to determine the accuracy of this model. These were extracted from the EMBL Protein Quaternary Structure (PQS) using the program SOCKET which detects native-like knobs into holes packing.[43] From this, a set of 54 coiled-coil structures were examined by eye to eliminate any helical bundles, a common false positive of this prediction method.[44] Figure 2-9 shows in blue the ability to fit these structures using the Crick parameterization. The rmsd range is between 0.4 and 2.0 Å, and all but 8 structures have an rmsd of less than 1.0 Å. Figure 2-10 shows a distribution of parameters. Using this set of values, I generated a set of 120 ideal parallel coiled-coil structures that can be used to approximate the range of the native structure space (see methods).

Crick parameterization has not previously been used to generate antiparallel coiled coils. However, it has been speculated that Crick Parameterization could be extended for this purpose.[38,42] This seems reasonable as coiled coils in both orientations are thought to have the same knobs-into-holes packing geometry[45] and the C_α-trace starting from either end of the ideal alpha helix is symmetrical. Because the Crick parameterization relies only on the C_α trace, antiparallel coiled coils have been described using the simple modification of numbering the C_α atoms of one chain in reverse order to generate a pseudo-parallel coiled coil.[42] To see if this modification to Crick parameterization would describe antiparallel coiled coils, a set of antiparallel structures was collected using the same method as for the parallel coiled coils. Here 70 dimeric antiparallel coiled coils were extracted from the PQS using SOCKET after a manual filtering step. These were fit using the Crick parameterization. The results are shown in green in

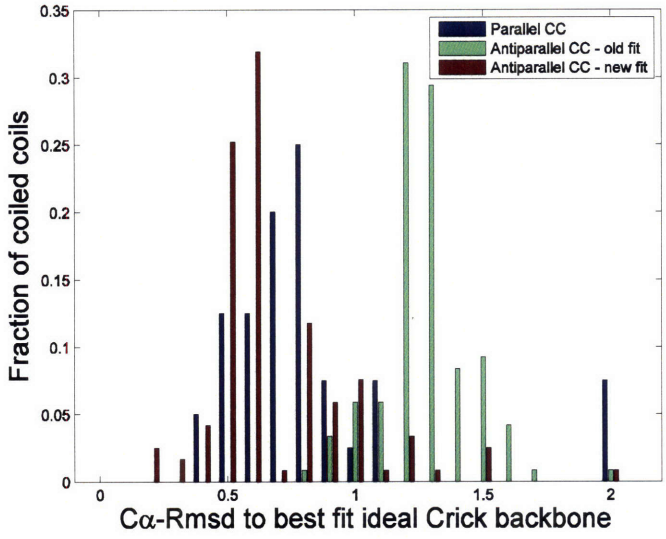


Figure 2-9: Distribution of coiled-coil backbone C_α-rmsd for native crystal structures with respect to the closest ideal structure. Parallel coiled coils fit with the Crick parameterization are shown in blue. Antiparallel coiled coils fit with a modified parallel Crick parameterization are shown in green. Antiparallel coiled coils fit with the new antiparallel-specific Crick parameterization are shown in red.

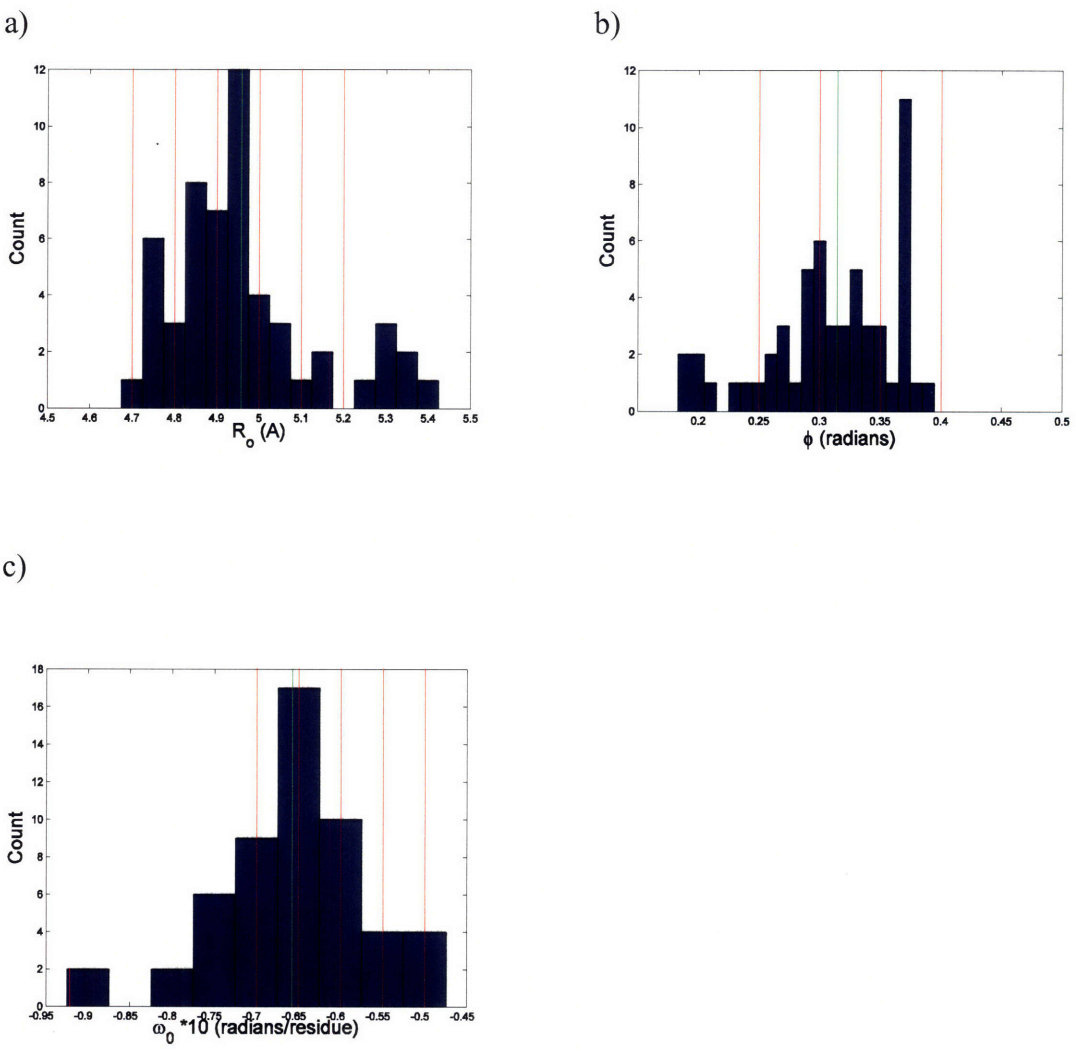


Figure 2-10: Histogram of the parallel Crick parameters generated by fitting parallel test-set structures to the best possible Crick backbone. Parameters a) R_0, b) ϕ and c) ω_0 are shown. The mean value of the parameter is indicated in green, and in red are the values used to generate the parallel-structure templates used for prediction. These parameters were chosen to span the range of values seen in native structures. As a note, the values of ω_0 sampled were skewed towards larger values because no structure with $\omega_0 < -0.07$ radians/residue was chosen as the lowest energy structure for any sequence, parallel or antiparallel.

Figure 2-9. The fit of these structures was not as good as for the parallel coiled coils, with a rmsd range of 0.8 to 2 Å. It was evident from this result that there was something different about the structure of the antiparallel coiled coils.

Previously, Gernet et al. demonstrated that in some antiparallel coiled coils, their core interactions are offset from each other.[46] This is illustrated by comparing two coiled-coil structures (Figure 2-11). To the left, the core residues in opposite chains of the parallel coiled coil lie in the same plane, which is not the case for the antiparallel coiled coil on the right. The current Crick parameterization method assumes that the corresponding C_α atoms lie in the same plane and therefore cannot capture this type of structural difference.

This native offset was accommodated by introducing the new parameter *apz* into the Crick parameterization (Equation 2-8 in Methods). For the first chain *apz* is set to zero, and for the second chain it is altered to account for the degree of offset. Since these two chains are no longer symmetrical about the superhelical axis, the parameter ϕ was allowed to have different values for each chain.

Another key change to this fitting procedure relates to the generation of the superhelical axis. For parallel coiled coils, which are symmetrical about this axis, the superhelical axis was taken as the axis of rotation that gave the greatest amount of superposition of the two chains. For the antiparallel coiled coils, this method would no longer find the superhelical axis because the two helices are not rotationally symmetrical about it. To account for this, the superhelical axis was varied during the fitting process, and the best fit parameters included the best fit superhelical axis. A detailed description of the new Crick parameterization for coiled coils is described in the Methods. These two additional parameters and the new method for identifying the superhelical axis were then used to fit the same test set of 70 antiparallel structures to idealized antiparallel coiled coils. The results are shown in Figure 2-9 in red. These structures are now fit as well as the parallel coiled coils. The distribution of *apz* values for antiparallel coiled coils shows how important the addition of this parameter was (Figure 2-12e). It was also important to allow the

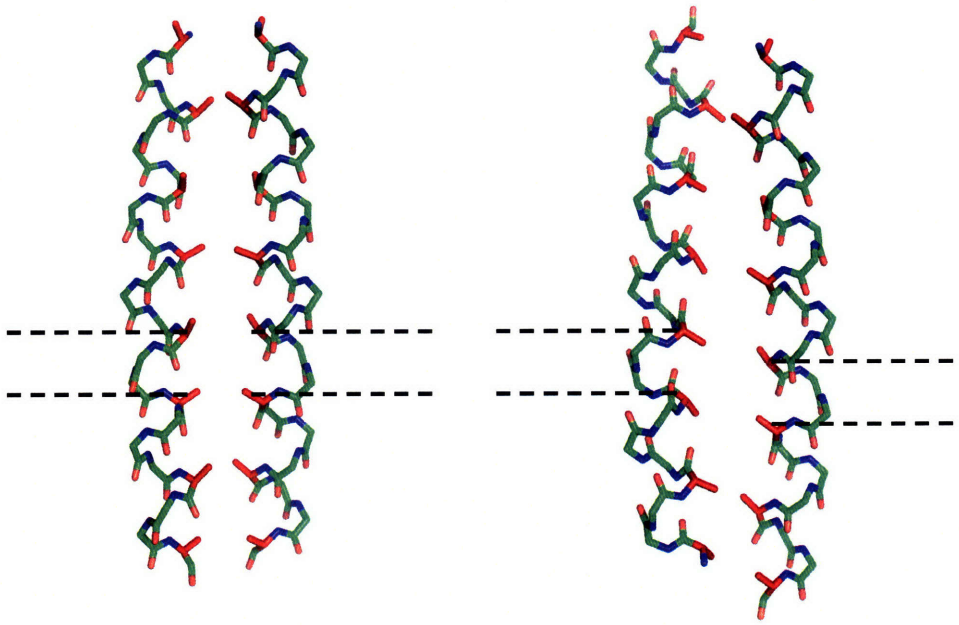


Figure 2-11: Illustration of the helical offset difference between parallel and antiparallel coiled coils. For both structures the C_α-C_β bond of the core **a** and **d** positions are shown in red. At left the parallel coiled coil has opposing **a-a'** pairs and **d-d'** pairs that are in the same plane, perpendicular to the superhelical axis. At right, the antiparallel coiled coil has opposing **a-d'** and **d-a'** pairs that are not in the same plane.

two ϕ parameters to be independent because the values of these two parameters are never identical (Figure 2-13). However, the fact that these two ϕ values were correlated was used in structural sampling.

With this new parameterization method for antiparallel coiled coils, it is now possible to sample the structure space of antiparallel coiled coils in a manner similar to that used for parallel coiled coils. A distribution of these parameters is shown in Figures 2-12 and 2-13. For antiparallel coiled coils, ϕ_A and ϕ_B are well correlated (Figure 2-13), allowing the reduction of the sampling space to 4 parameters. Using this set of parameters, I generated a structure set of 81 ideal antiparallel coiled coils that can be used to model the range of the native structures space

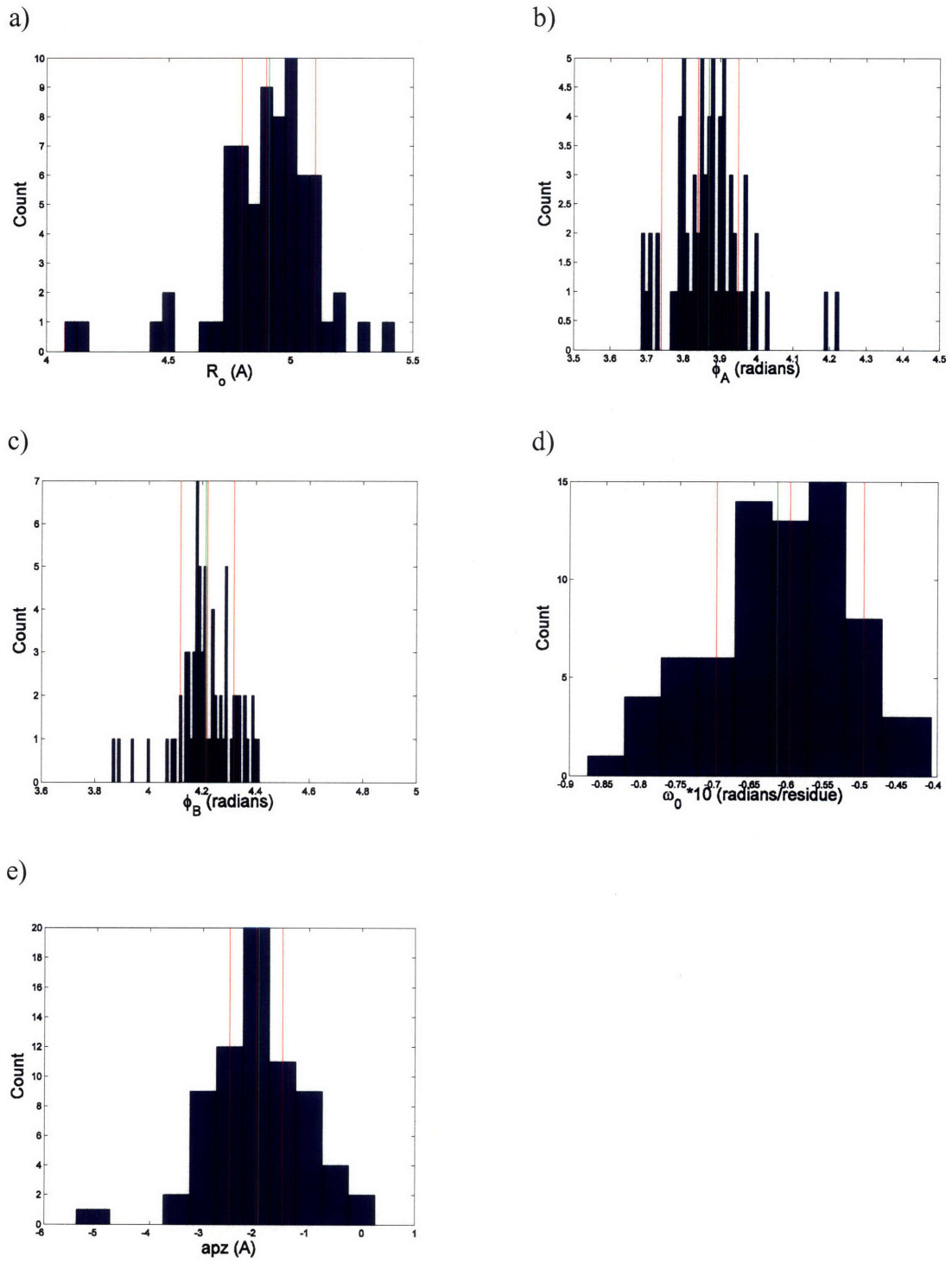


Figure 2-12: Histogram of the antiparallel Crick parameters generated by fitting antiparallel test-set structures to the best possible Crick backbone. Parameters a) R_0 b) ϕ_A, c) ϕ_B d) ω_0 and e) *apz* are shown. The mean value of the parameter is indicated in green, and in red are the sampled values, as in Figure 2-10.

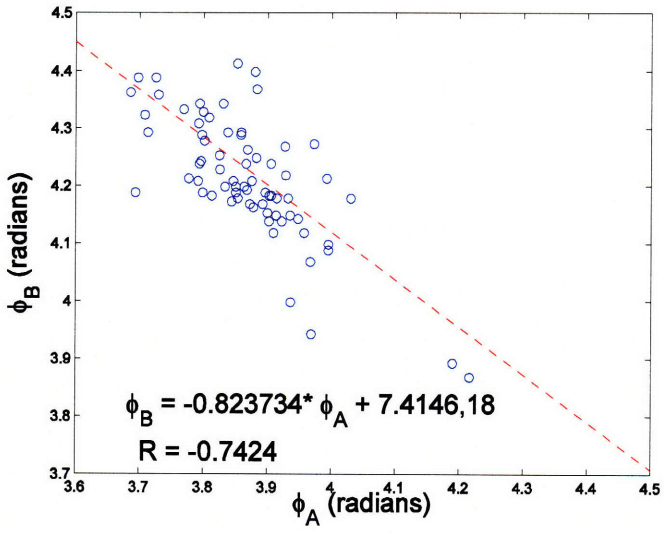


Figure 2-13: Antiparallel ϕ_A and ϕ_B correlation for all structures in the test set. These two parameters were treated independently in antiparallel Crick parameterization, but showed a strong correlation. As a result, ϕ_A was varied in the ideal structure set and ϕ_B was determined using the equation shown in the Figure.

(see methods). This set along with the parallel set were used to compare the binding preference of sequences to be parallel or antiparallel, as described in Chapter 4.

Summary

In this chapter I described two methods for sampling helical structures. First, normal mode analysis was used to determine the important types of helical deformations in existing structures and then uses those vectors to sample the highly populated parts of structure space. This structure-space was also captured through the use of principal-component analysis which showed that the largest deformations were mostly linear combinations of these types of modes. The second method described the structural changes associated with coiled-coil interfaces. Previous work by Crick and by Harbury et al. described the structure of coiled coils using a reduced set of parameters.[37,40] In this chapter, I demonstrated that this method could be extended to describe antiparallel coiled coils accurately with the introduction of two new parameters: an antiparallel offset *apz*, to account for the relative shift between the two chains and an additional

helical offset parameter ϕ for one chain. These parameters allowed for the generation of idealized models that matched closely to structures in the PDB and made it possible to sample new antiparallel coiled-coil structures. Both of these methods allow for an efficient sampling of the structure space and can be integrated into flexible backbone modeling for design and prediction.

Methods

Helix database

The entire PDB was downloaded on August 5, 2005. From these structures, the backbone dihedral angles were calculated using the chi module of the program Jackal.[47] From x-ray crystal structures with resolution of 2.5 Å or better, I extracted over 45,000 protein fragments with at least 15 consecutive residues with ϕ and ψ angles in the range of -50° +/- 30°. These helices were aligned to ideal helices that had CHARMM param19 default bond lengths and bond angles, and dihedral angles defined as $\omega = 180^{\circ}$,$\phi = -57^{\circ}$ and $\psi = -47^{\circ}$.[2] Deformations could then be captured using the difference between the N-C_{α}-C backbones of the helix database and the ideal helix.

All multiple-chain structures containing helices from this database were extracted. From these, the interface residues between each pair of chains were determined, defined as any residue that was within 5 Å of the other chain. If more than 50% of these residues were part of a helix from the database, then this structure was marked as having a helix-mediated interface. This resulted in 170 unique PDB structures.

Backbone variations in normal-mode space

To construct helical backbones, we used a set of NM vectors similar to those described by Emberly et al.,[13,35] with slight modifications. The C, C_α and N backbone atoms, rather than only C_α atoms, were used to compute the NM potential. In Cartesian space, the harmonic potential of a structure was calculated using the following function described by Tirion[26] and later used by Tama et al.[21]:

$$V = \sum_{d_{ij}^o \le R_c} k\left(d_{ij} - d_{ij}^o\right)^2 \tag{2-4}$$

Here k is a force constant that was set to 10 for all pairs of atoms, d_{ij} is the distance between atoms i and j of a structure, and d_{ij}^o is the reference distance between these two atoms in the ideal-helix structure. This potential does not contain values for pairs of atoms with distances larger than the cutoff of R_c. This value was set to 8 Å as suggested by Tama et al.[21] and Bahar et al.[48] From this potential the Hessian (H) can be calculated:

$$H_{i,m,j,n} = \frac{\partial^2 V}{\partial x_{i,m} \partial x_{i,n}} \tag{2-5}$$

Here $x_{i,m}$ is m^{th} Cartesian coordinate of atom i. The eigenvectors $\left\{\overline{X}_i\right\}$ of this matrix are the normal modes, and the eigenvalues are the corresponding frequencies. Modes corresponding to the six rotational and translational degrees of freedom were discarded and the remaining modes were used to sample distortions of a helix about a fixed Cartesian center. A separate set of NM vectors was calculated for each helix of length L (number of residues). To generate a variable helix **F**, a set of $9L$-6 NM amplitudes $\{a_i\}$ was multiplied by the NM vectors and added to the ideal helix **I**:

$$\mathrm{F} = \mathrm{I} + \sum_{i=1}^{9L-6} a_i \bar{X}_i \tag{2-6}$$

To determine the NM values of the native helix, a difference vector between the native helix and the aligned ideal helix was calculated. This vector was fit to a linear combination of NM vectors using linear regression. The fitted linear coefficients gave the $\{a_i\}$ of the native helix.

To generate a new NM structure, all atoms of the helix were removed except for the backbone C, C_α, and N. The backbone was deformed by applying a linear combination of NM vectors to the ideal helix, as described above. We chose random values for the two lowest frequency NM parameters from a Gaussian distribution approximating the mode values seen in helices in the PDB, centered on the starting values for the set. The backbone was reconstructed by regenerating H and O atoms with CHARMM param19 default parameters. The side chains were assigned CHARMM default values for bond lengths and bond angles, but crystal-structure dihedral values. Structures with backbone atoms on different chains within 3 Å were discarded. The remaining NM structures could be used for design.

Principal-component analysis

The backbone deviations found in the database of alpha helices were computed by decomposing the structural variation into its principal components. To do this, I calculated the covariance matrix ($C_{i,j}$) shown in equation 2-7. This matrix contains the relative variance over all *N* structures for each pair of N-, C_α - or C-atom coordinates as compared to their average.

$$C_{i,j} = \frac{1}{N-1}\sum_{m=1}^{N}\left(x_{mi} - \langle x_i \rangle\right)\left(x_{mj} - \langle x_j \rangle\right) \qquad (2\text{-}7)$$

Here x_{mi} and x_{mj} are the i^{th} and j^{th} x-, y- or z-coordinates of structure *m*, and $\langle x_i \rangle$ and $\langle x_j \rangle$ are the average values of those coordinates over all structures. From this covariance matrix I calculated

the eigenvectors and eigenvalues. The eigenvectors are the principal components and the eigenvalues are the amount of structural variation captured by that component.

Coiled-coil database

Parallel and antiparallel coiled-coil dimer structures were obtained by applying SOCKET to the EMBL Protein Quaternary Structure (PQS) database downloaded on April 12, 2007.[43] Structures returned by SOCKET were filtered to exclude those shorter than 18 residues as well as those with a discontinuous heptad assignment. A manual filtering step was used to exclude non-coiled-coil structures, such as certain portions of helix bundles, helix sheets and other extended knobs-into-holes assemblies.[44] The GCN4 coiled-coil family was overrepresented in this set; several sequences containing point mutations were removed. Finally, due to the significant minority of parallel heterodimeric coiled-coil crystal structures, we added seven sequence pairs from the human bZIP family, for which the helix orientation and alignment could be determined by sequence alignment[49,50]: ATF7+MAFK, ATF2+FOS, CREBPA+JUN, CEBPbeta+CEBPalpha, ATF1+CREM, CEBPgamma+ATF4 and the ATF1 homodimer. All complexes contained two chains of the same length and were completely overlapping (i.e. had "blunt" ends) in both parallel and antiparallel orientations. The final set consisted of 61 parallel and 70 antiparallel coiled coils.

Crick parameterization

To describe and generate parallel coiled-coil dimer backbones, I used the parameterization originally proposed by Crick and subsequently implemented by Harbury et al. as a user routine in CHARMM.[37,38] This parameterization has been shown to closely mimic the geometry of several

parallel coiled coils.[38] Additionally, using our parallel coiled-coil test set, we found that this idealized parameterization can be fit to a set of 54 native backbones with C_α RMSD values ranging from 0.25 to 2.5 Å, and with 46 of 54 backbones having an RMSD less than 1.0 Å (Figure 2-14).

I modified the Crick/Harbury approach to describe and generate antiparallel coiled-coil backbones. As in the fitcc program,[42] we used the fact that the C_α trace of the antiparallel coiled coil has approximately the same symmetry properties as the parallel coiled coil. The two relevant exceptions are that a symmetry-breaking axial shift can occur between the two chains, and the ϕ values that describe the angle of side chains relative to the helix-helix interface need not be the same on both chains. The modification of the coiled-coil parameterization accounts for these differences by introducing two new parameters. Parameter apz_i captures the helical shift as described above, and parameter ϕ is replaced with an independent value for each helix: ϕ_A and ϕ_B. The parameterization for antiparallel coiled coils is re-written as:

$$
\begin{aligned}
&CC(\tau) = EC'(\tau) + H(\tau) \\
&E(\omega_0\tau,\alpha,0) = \begin{pmatrix} \cos(\omega_0\tau) & -\sin(\omega_0\tau)\cos(\alpha) & 0 \\ \sin(\omega_0\tau) & \cos(\omega_0\tau)\cos(\alpha) & 0 \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \\
&C' = \begin{pmatrix} R_1\cos(\omega_1\tau+\phi_i) \\ R_1\sin(\omega_1\tau+\phi_i) \\ apz_i \end{pmatrix} \\
&H(\tau) = \begin{pmatrix} R_0\cos(\omega_0\tau) \\ R_0\sin(\omega_0\tau) \\ d\tau\cos(\alpha) \end{pmatrix} \\
&where \quad \sin(\alpha) = \frac{R_0\omega_0}{d}
\end{aligned}
\tag{2-8}
$$

Here R_0 is the superhelical radius, ϕ_i are phase angles that locate the residues on the superhelical backbone trace, and ω_0 is the superhelical frequency, α is the helix-crossing angle, R_1 is the alpha helix radius and ω_1 is the alpha helix frequency. As described above, apz_i is an axial helical

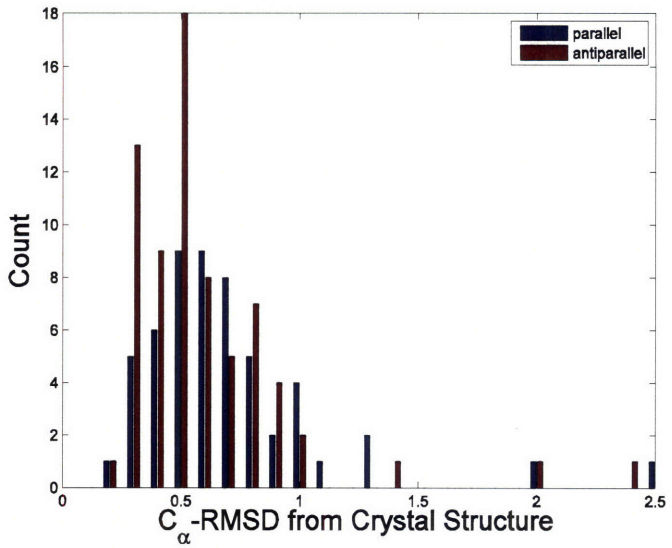


Figure 2-14: Native coiled-coil variation described using the Crick parameterization. Histogram of the backbone RMSD (C_α-only) between each test-set structure and its closest Crick backbone.

offset that is set to 0 for chain A, and is non-zero for other chains. As for the parallel coiled coil, we generate chains by constructing them using this equation and rotating them into position about the superhelical axis. This antiparallel parameterization was coded as a user-defined energy routine in CHARMM, as for the parallel parameterization.

I used the Crick parameterization both to fit idealized backbones to native structures and to generate *de novo* backbones. To fit a native structure, I optimized superhelical parameters, as well as two external parameters that locate the coiled coil in the laboratory frame. It is important that the superhelical axis of the native coiled coil be aligned with the z-axis of the parameterization above. The superhelical axis of a parallel coiled coil can be well-approximated as the rotational axis that maximizes superposition of one helix onto another. However, this is not the case for antiparallel coiled coils. For these, I found the best alignment by adjusting the internal Crick parameters, along with two Euler rotations and three translational degrees of freedom, using a process similar to that of the fitcc program. The center of mass of the helix was translated to the origin, and then the coiled coil was approximately oriented, using two vectors

defined by connecting the first and last C_α atom of each helix. The average of these two vectors was aligned with the z-axis. Starting from this position, the rest of the Crick parameters, along with two Euler angles and translations in three dimensions, were optimized using Matlab's constrained minimization algorithm[51] to minimize the RMSD of the native helix to the closest ideal Crick helix. Given this superhelical alignment, antiparallel Crick parameters were fit in CHARMM by minimizing the energy with respect to these parameters as well as a rotation about the superhelical axis and a translation with respect to this axis. The energy minimized was proportional (with constant 25 kcal/Å^2) to the sum of the distances squared of all C_α atoms from the ideal Crick C_α atom positions.

Generation of Crick backbones

All structures were generated via minimization under a potential that included the user defined Crick energy as well as van der Waals interactions, bond length, bond angle, dihedral and improper dihedral energy terms, and a hydrogen bonding potential, all defined by the param19 force field.[52] Parameters R_1, ω_1 and d, which describe alpha–helix geometry, were set to 2.26 Å, $4\pi/7$ radians per residue and 1.52 Å respectively.[38] Other parameters were sampled as follows: The parallel set contained 120 structures with R_0 values of 4.7, 4.8, 4.9, 5.0, 5.1 and 5.2 Å, ϕ values of 0.25, 0.30, 0.35, and 0.40 radians, and ω_0 values of -0.055, -0.06, -0.065, and -0.70 radians. The antiparallel set contained 81 structures with R_0 values of 4.8, 4.9 and 5.1 Å, ω_0 of -0.050, -0.060 and -0.070 radians, ϕ_A, ϕ_B pairs (in radians) of (0.412, 0.395), (0.422, 0.384), (0.432, 0.374) and apz_i values of 1.5, 2.0 and 2.5 Å. These values span the space of native parallel and antiparallel sequences. As illustrated in Figures 2-10 and 2-12. ϕ_A, ϕ_B values were sampled as pairs due to correlations between these in native structures (Figure 2-13).

References

1. Chothia C, Levitt M, Richardson D. Structure of proteins: packing of alpha-helices and pleated sheets. Proc Natl Acad Sci U S A 1977;74(10):4130-4134.
2. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Advances in protein chemistry 1968;23:283-438.
3. Emberly EG, Wingreen NS, Tang C. Designability of alpha-helical proteins. Proc Natl Acad Sci U S A 2002;99(17):11163-11168.
4. Yue K, Dill KA. Constraint-based assembly of tertiary protein structures from secondary structure elements. Protein Sci 2000;9(10):1935-1946.
5. Ausiello G, Cesareni G, Helmer-Citterich M. ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure. Proteins 1997;28(4):556-567.
6. Vakser IA, Jiang S. Strategies for modeling the interactions of transmembrane helices of G protein-coupled receptors by geometric complementarity using the GRAMM computer algorithm. Methods Enzymol 2002;343:313-328.
7. Ross SA, Sarisky CA, Su A, Mayo SL. Designed protein G core variants fold to native-like structures: sequence selection by ORBIT tolerates variation in backbone specification. Protein Sci 2001;10(2):450-454.
8. Offredi F, Dubail F, Kischel P, Sarinski K, Stern AS, Van de Weerdt C, Hoch JC, Prosperi C, Francois JM, Mayo SL, Martial JA. De novo backbone and sequence design of an idealized alpha/beta-barrel protein: evidence of stable tertiary structure. Journal of molecular biology 2003;325(1):163-174.
9. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A. De novo design and structural characterization of proteins and metalloproteins. Annu Rev Biochem 1999;68:779-819.
10. Schellman JA. The stability of hydrogen-bonded peptide structures in aqueous solution. C R Trav Lab Carlsberg [Chim] 1955;29(14-15):230-259.
11. Zimm BH, Bragg JK. Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. Journal of Chemical Physics 1959;526(3):526-535.
12. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 1951;37(4):205-211.
13. Emberly EG, Mukhopadhyay R, Wingreen NS, Tang C. Flexibility of alpha-helices: Results of a statistical analysis of database protein structures. Journal of molecular biology 2003;327(1):229-237.
14. Makhatadze GI, Privalov PL. Energetics of protein structure. Adv Protein Chem 1995;47:307-425.
15. O'Neil KT, DeGrado WF. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 1990;250:646--651.
16. Pace CN, Scholtz JM. A helix propensity scale based on experimental studies of peptides and proteins. Biophy J 1998;75:422--427.
17. Fersht AR, Matouschek A, Serrano L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. Journal of molecular biology 1992;224(3):771-782.

18. Munoz V, Serrano L. Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. Journal of molecular biology 1995;245(3):275-296.
19. Chakrabartty A, Kortemme T, Baldwin RL. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. Protein Sci 1994;3(5):843-852.
20. Ma J. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. Structure 2005;13(3):373.
21. Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Engineering 2001;14(1):1-6.
22. Kitao A, Hirata F, Go N. The Effects of Solvent on the Conformation and the Collective Motions of Protein - Normal Mode Analysis and Molecular-Dynamics Simulations of Melittin in Water and in Vacuum. Chem Phys 1991;158(2-3):447-472.
23. Horiuchi T, Go N. Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. Proteins 1991;10(2):106-116.
24. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. Biophysical Journal 2005;88(2):1291.
25. DePristo MA, De Bakker PI, Shetty RP, Blundell TL. Discrete restraint-based protein modeling and the Calpha-trace problem. Protein Sci 2003;12(9):2032-2046.
26. Tirion MM. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Physical Review Letters 1996;77(9):1905-1908.
27. Amadei A, Linssen ABM, Berendsen HJC. Essential Dynamics of Proteins. Proteins-Structure Function and Genetics 1993;17(4):412-425.
28. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal component analysis and long time protein dynamics. J Phys Chem-Us 1996;100(7):2567-2572.
29. Garcia AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett 1992;68(17):2696-2699.
30. van Aalten DM, Amadei A, Linssen AB, Eijsink VG, Vriend G, Berendsen HJ. The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. Proteins 1995;22(1):45-54.
31. Theobald DL, Wuttke DS. Accurate Structural Correlations from Maximum Likelihood Superpositions. PLoS Comput Biol 2008;4(2):e43.
32. Velazquez-Muriel JA, Carazo JM. Flexible fitting in 3D-EM with incomplete data on superfamily variability. J Struct Biol 2007;158(2):165-181.
33. Velazquez-Muriel JA, Valle M, Santamaria-Pang A, Kakadiaris IA, Carazo JM. Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies. Structure 2006;14(7):1115-1126.
34. Alber F, Forster F, Korkin D, Topf M, Sali A. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. Annu Rev Biochem 2008.
35. Emberly EG, Mukhopadhyay R, Tang C, Wingreen NS. Flexibility of beta-sheets: Principal component analysis of database protein structures. Proteins-Structure Function and Bioinformatics 2004;55(1):91-98.
36. Crick FH. Is alpha-keratin a coiled coil? Nature 1952;170(4334):882-883.
37. Crick FH. The Fourier Transform of a Coiled-Coil. Acta Cryst 1953;6:685-689.

38. Harbury PB, Tidor B, Kim PS. Repacking Protein Cores with Backbone Freedom - Structure Prediction for Coiled Coils. Proceedings of the National Academy of Sciences of the United States of America 1995;92(18):8408-8412.
39. Keating AE, Malashkevich VN, Tidor B, Kim PS. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. Proc Natl Acad Sci U S A 2001;98(26):14825-14830.
40. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282(5393):1462-1467.
41. Plecs JJ, Harbury PB, Kim PS, Alber T. Structural test of the parameterized-backbone method for protein design. Journal of molecular biology 2004;342(1):289-297.
42. Sales M. FitCC Personal Communication with Tom Alber; http://ucxray.berkeley.edu/~mark/fitcc.html; 2007.
43. Walshaw J, Woolfson DN. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. Journal of molecular biology 2001;307(5):1427-1450.
44. Walshaw J, Woolfson DN. Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. J Struct Biol 2003;144(3):349-361.
45. Crick FH. The packing of alpha helices: simple coiled-coils. Acta Cryst 1953;6:689-697.
46. Gernert KM, Surles MC, Labean TH, Richardson JS, Richardson DC. The Alacoil: a very tight, antiparallel coiled-coil of helices. Protein Sci 1995;4(11):2252-2260.
47. Xiang JZ. JACKAL: A Protein Structure Modeling Package. New York, NY: Columbia University and Howard Hughes Medical Institute; 2002.
48. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and Design 1997;2(3):173-181.
49. Newman JR, Keating AE. Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science 2003;300(5628):2097-2101.
50. Vinson C, Myakishev M, Acharya A, Mir AA, Moll JR, Bonovich M. Classification of human B-ZIP proteins based on dimerization properties. Mol Cell Biol 2002;22(18):6321-6335.
51. Matlab R14: The MathWorks, Inc.; 2005.
52. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comp Chem 1983;4(2):187-217.

Chapter 3

Modeling backbone flexibility to achieve sequence diversity: The design of novel alpha-helical ligands for Bcl-x_L

Portions reprinted with permission of Elsevier B.V. from:

Fu, X., Apgar, J.R., Keating. A.E. (2007) "Modeling backbone flexibility to achieve sequence diversity: The design of novel alpha-helical ligands for Bcl-xL" *J. Mol. Bio.* **31**, 1099-1117

Collaborators Notes:

Xiaoran Fu performed all calculations involving the program SCADS, sequence and profile clustering, and all of the experiments.

Abstract

Computational protein design can be used to select sequences that are compatible with a fixed-backbone template. This strategy has been used in numerous instances to engineer novel proteins. However, the fixed-backbone assumption severely restricts the sequence space that is accessible via design. For challenging problems, such as the design of functional proteins, this may not be acceptable. In this chapter, we present a method for introducing backbone flexibility into protein design calculations and apply it to the design of diverse helical BH3 ligands that bind to the anti-apoptotic protein Bcl-x_L, a member of the Bcl-2 protein family. We demonstrate how normal mode analysis can be used to sample different BH3 backbones, and show that this leads to a larger and more diverse set of low-energy solutions than can be achieved using a native high-resolution Bcl-x_L complex crystal structure as a template. We tested several of the designed solutions experimentally and found that this approach worked well when normal mode calculations were used to deform a native BH3 helix structure, but less well when they were used to deform an idealized helix. A subsequent round of design and testing identified a likely source of the problem as inadequate sampling of the helix pitch. In all, we tested seventeen designed BH3 peptide sequences, including several point mutants. Of these, eight bound well to Bcl-x_L and four others showed weak but detectable binding. The successful designs showed a diversity of sequences that would have been difficult or impossible to achieve using only a fixed backbone. Thus, introducing backbone flexibility via normal mode analysis effectively broadened the set of sequences identified by computational design, and provided insight into positions important for binding Bcl-x_L.

Introduction

Computational protein design holds great promise for guiding the discovery of useful biomolecules. In particular, the design of proteins that form specific interactions could facilitate the development of therapeutic inhibitors or agonists. There have been many experimentally validated examples of protein design, including the design of stable folds, active enzymes and specific receptors.[1-18] Most successful protein design calculations so far have sought to identify a sequence that stabilizes a fixed backbone geometry, as defined by a high-resolution structure. Fixed-backbone design is used to limit the potentially infinite search space and make design problems more tractable. However, the fixed-backbone approximation is an artificial limitation that severely restricts the space of possible design solutions. For example, it has frequently been observed that sequences designed using a fixed backbone are very native-like.[19-21] As the demands placed on protein design problems increase, e.g. as designed proteins are required to be more specific, more highly functional, less aggregation prone or easier to encode in DNA libraries, artificial restrictions such as those imposed by using a fixed backbone become less tolerable. In this chapter, we propose a new method for introducing backbone structural variation using normal mode (NM) analysis and explore it in the context of a protein-protein interaction that is of critical importance for cancer and other diseases – the interaction of pro-apoptotic peptides with anti-apoptotic members of the Bcl-2 family.

The Bcl-2 family comprises both pro- and anti-apoptotic proteins.[22,23] Five mammalian anti-apoptotic family members, Bcl-2, Bcl-x_L, Bcl-w, Mcl-1 and (presumably) A1, have a conserved globular structure, and all known family members, both pro- and anti-apoptotic, share a weakly conserved short BH3 (Bcl-2 homology 3) sequence. Peptides corresponding to the BH3 region have been shown in several instances to adopt an α-helical structure when bound into a

hydrophobic groove on the surface of anti-apoptotic proteins.[24-27] This interaction mode is assumed to be conserved for a larger group of BH3 peptides and anti-apoptotic receptors that have been observed to interact.[28] Recent studies have begun to map the interaction preferences of the Bcl-2 family of proteins and have shown that BH3 peptides have distinct binding profiles, with some binding only a subset of anti-apoptotic receptors and others interacting promiscuously.[29-32] Several models have been proposed to explain how the selectivity of this interaction is important for regulating apoptosis via mitochondrial pathways.[29-31] All of these models support the idea that selective disruption of specific interactions could be a valuable strategy for treating cancers.

Both peptide and small-molecule inhibitors that disrupt Bcl-2 interactions have been identified. In a protein engineering approach, the Schepartz group grafted BH3 sequences onto a mini-protein scaffold derived from an avian pancreatic polypeptide.[33,34] By screening a combinatorial library at selected positions in the BH3 part of the sequence, several peptides were identified that bound to Bcl-2 and Bcl-x_L. Sadowsky et al.[35] designed a novel $\alpha/\beta+\alpha$–amino acid backbone scaffold and identified a sequence that bound to Bcl-x_L with sub-nanomolar affinity. Small-molecule inhibitors that interrupt the interactions between BH3 and Bcl-x_L in the low micromolar range were identified in 2001.[36] More recently, Olterstorf et al.[37] screened hundreds of small molecule fragments using NMR to identify those that bound tightly to Bcl-x_L. A promising compound constructed from these fragments has nanomolar affinity and is now in pre-clinical trials for suppressing certain tumors. Although these inhibitors span a wide range of physical and chemical properties, a common theme in their development was the use of extensive screening and selection to identify compounds with high binding affinity.

BH3 peptides have very diverse sequences and show varying levels of binding to anti-apoptotic Bcl-2 proteins.[30] It would be useful to generate artificial peptides that exhibit diverse binding profiles, distinct from those of native peptides, with respect to Bcl-2 family receptors. Such peptides could serve as reagents to help dissect the biological consequences of different interactions in apoptosis and could lead to the development of more specific inhibitors with better therapeutic properties. Until very recently, however, only one high-resolution crystal structure of a Bcl-2 family receptor/BH3 complex has been solved - a complex of Bcl-x_L with a BH3 peptide derived from Bim.[26,38,39] Ligands designed based on this fixed-backbone structure are likely to sample only a small portion of the sequence space that holds interesting, diverse binding peptides. Introducing backbone flexibility to the design protocol may provide a way to overcome this limitation (see Figure 3-1).

Protein backbones have many degrees of freedom, and sampling these efficiently in protein design is quite challenging, as reviewed by Butterfoss and Kulman.[40] One approach has been to use small sets of parameters to describe variation using a simplified geometry. This technique has been applied to coiled coils and helical bundles,[41,42] and a related approach has been used to vary the orientation of secondary structure elements in the α/β fold of the $\beta 1$ immunoglobulin-binding domain of streptococcal protein G.[15] The Baker group has had tremendous success modeling backbones in structure prediction by sampling from peptide fragments in the PDB. They have also demonstrated that this approach is effective in protein design.[16,43] Kono and Saven used NMR structure ensembles to represent possible backbone conformations,[44] and Larson et al. used a Monte Carlo procedure to sample backbone φ and ϕ angles and generate 'native-like' structure ensembles.[45] In this chapter, we use normal mode (NM) analysis to introduce backbone flexibility. This method has proven useful for modeling

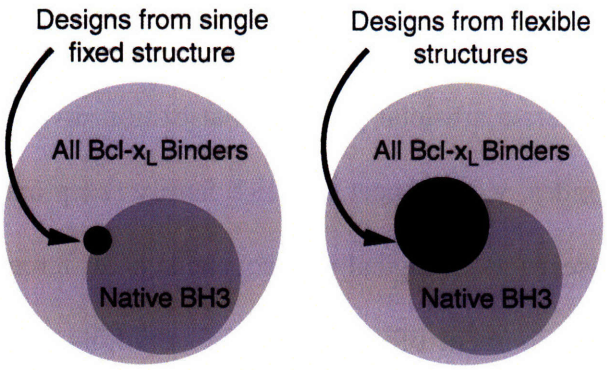


Figure 3-1: Cartoon illustrating the idea of using flexible backbones to expand the accessible BH3 peptide sequence space. Circles encompassing varying areas indicate the sequence diversity of 26mers meeting different criteria.

variations of secondary structure elements.[46,47] It shares the advantages of parameterized sampling but can potentially be applied more broadly.

Any protein motion can be described as a sum of NM distortions, but such a description is most useful if the number of modes making significant contributions to structural variation is small, and if these can be identified. As described in a recent review by Ma,[48] a small number of low-frequency normal modes can be used to model functionally important conformational transitions in several biomolecules that agree with motions observed in molecular dynamics simulations. It has also been noted that a significant amount of the variation seen among different crystal structures of the same, or closely related, proteins can be described by a small set of NM values.[49,50] Specifically for helical regions, Emberly et al. have shown that most of the deformation of the C_α trace can be captured by three low-energy modes.[46] These modes are two perpendicular bends and a helical twist.

We have used NM calculations to generate deformations associated with the C_α, C and N- atom backbone of helical peptides for protein design. We started with the crystal structure of a Bcl-x_L/Bim-BH3* complex[26] and used NM analysis to construct diverse sets of backbones by

* Bim refers to Bim-BH3 throughout this chapter.

fixing the receptor structure and varying the conformation of the binding helix. We then ran computational design calculations on both the crystal structure and on structures in the flexible backbone sets. A larger sequence space could be accessed when flexible backbones were considered. The binding of seventeen designed peptides spanning a range of backbone geometries was tested against three receptor proteins. Eight peptides bound well to Bcl-x_L, as intended, and four more showed weak but detectable binding. Several peptides showed altered binding profiles compared to the wild-type Bim peptide on which the designs were based.

Results

Chapter 2 described how NM analysis can be used to generate structural variation in helical backbones for protein design. The follow sections show how we have used such a strategy to design novel Bcl-x_L ligands.

Flexible backbones generated using normal-mode analysis

As described in Chapter 2, two low-energy normal modes can capture most of the deformation of helices found in the PDB. Given this observations, we used NM analysis to generate two sets of variable templates for protein design. Two hundred I-set (ideal-helix set) and 200 N-set (native-helix set) backbones were generated as described in the Methods. The primary difference between these two sets is in the local deformations. The N-set retains small relaxations associated with the match of the native ligand to the receptor, whereas these have all been removed in the I-set. The purpose of generating two sets of backbones was to reflect different design scenarios that may be encountered. The N-set backbones may be a good choice in cases where a crystal structure complex of the target helix is available. The I-set could be used

in the more general case in which a helix must be constructed *de novo*. Here we use information from the complex structure to position the deformed helices with respect to the receptor, but with docking methods this helix could be placed without this prior knowledge.

Before using the flexible backbone templates for design, we characterized them by repacking the native sequence of Bcl-x_L/Bim on each structure, as described in the Methods. The N-set backbones included solutions that were very close to the native structure in both rmsd and energy, and extended to ~3 Å rmsd (Figure 3-2). Our energy function effectively recognized the native structure, assigning higher energies to structures with higher deviations. Energy minimization of the Bim helix led to minimal structural changes and little change in energy for the best N-set templates, whereas small steric clashes were relieved in the higher-energy structures. The I-set gave structures with larger backbone rmsd from the native structure (up to ~6 Å) and considerably higher energies. Minimization of the I-set Bim helix backbones gave little structural change. However, the energies of the best of these solutions became comparable to those of the minimized N-set, with rmsd values ranging from ~1.5 – 4.3 Å. This analysis suggested that both sets might be reasonable design templates, provided the helix backbone structures were relaxed, with the N-set sampling more native-like structures and the I-set including greater variability.

The sequence landscape over multiple backbones

To evaluate which of the 400 backbones in the N- and I-sets were appropriate for designing helical ligands for Bcl-x_L, we used the SCADS program. SCADS is a statistical protein design method that can rapidly generate sequence profiles that are consistent, in a mean-field sense, with a fixed backbone geometry. We used it to determine which N- and I-set

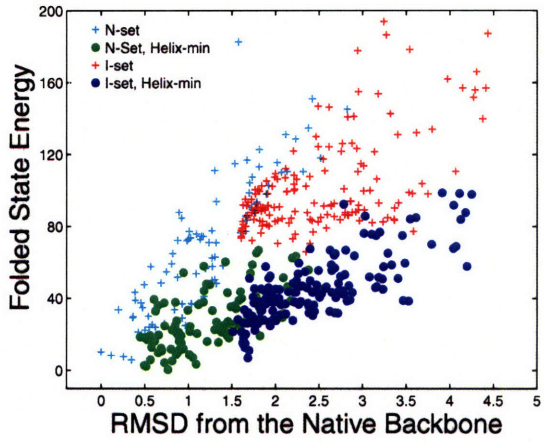


Figure 3-2: Modeling the complex of Bcl-x_L with Bim using different backbone sets. The folded-state energy relative to the crystal-structure energy is plotted against the backbone rmsd from the crystal structure (in Å). Energies were evaluated without (crosses) or with (circles) minimization to relax the structure (see Methods). N-set structures prior to minimization are shown in cyan and those after are shown in green. Energies of I-set structures prior to minimization are shown in red and those after are shown in blue.

backbones were compatible with low-energy sequences by redesigning all 26 residues of Bim on each template. The conformational energies of designed sequence profiles (E_{conf}, see Methods) are plotted as a function of the values of normal mode 1 and normal mode 2 (nm1 and nm2) for each backbone in Figure 3-3c and d. A smooth energy surface with a relatively flat well is observed for both structure sets. As shown in a similar plot of the rmsd from the native backbone (Figure 3-3a and b), we found that the lowest energy region is in the vicinity of the wild-type structure.

To probe the extent to which structural variation can provide diversity in designed sequences, we compared sequence profiles generated from the crystal structure backbone and from both sets of distorted backbones. Backbones were clustered according to sequence profiles derived from them, using a pairwise sequence profile similarity score (ΔSS) and the Xcluster program.[51] Seven clusters were defined in the I-set and eight in the N-set. Structures from the same sequence-profile cluster are indicated with the same symbol in Figure 3-3c and d, showing that the clusters defined in sequence space are also clustered in structure space. The clusters are

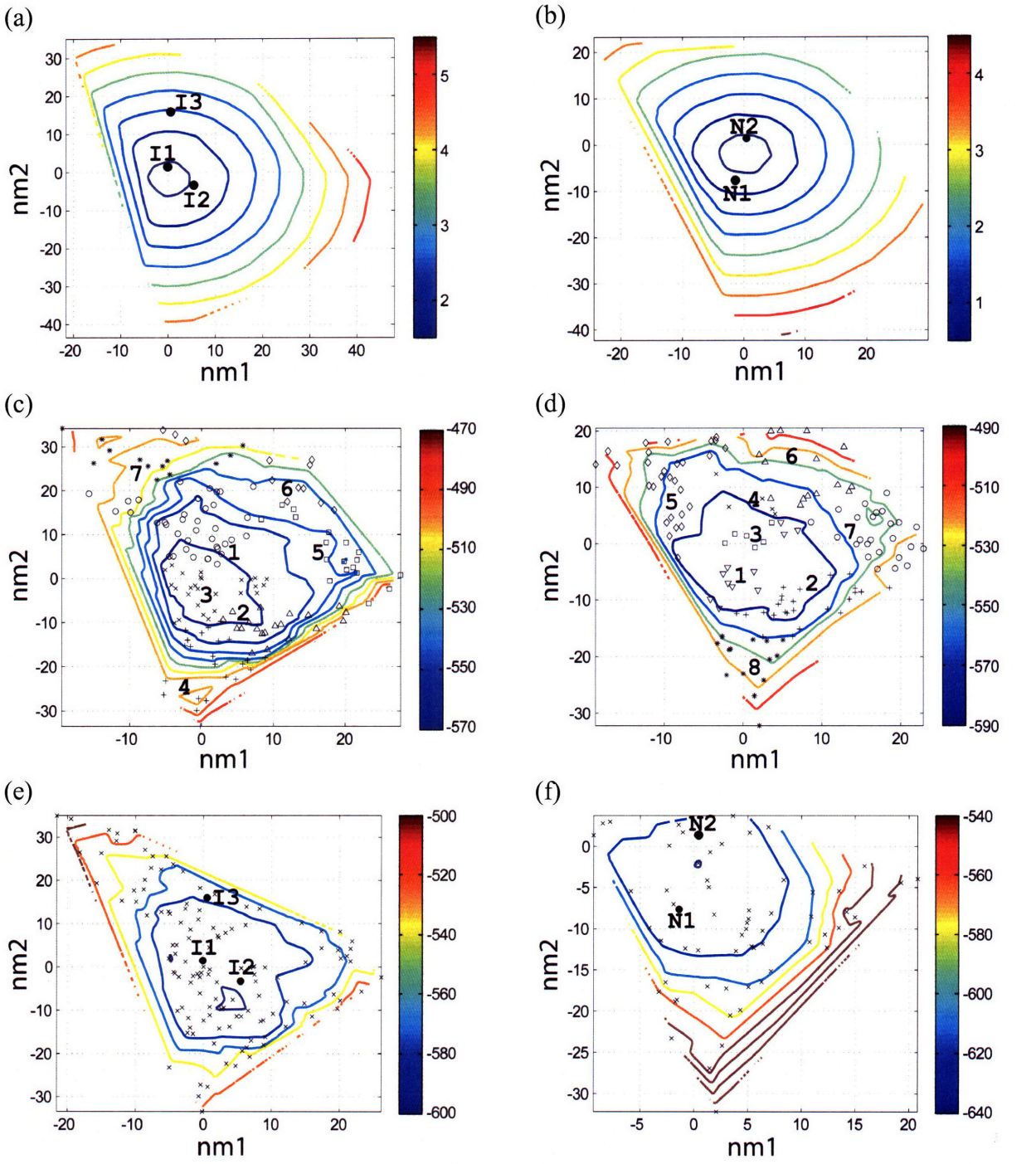


Figure 3-3: Characterization of the I- and N- sets using SCADS protein design. (a) and (b) Contour diagrams of the root-mean square deviation (rmsd) between the native backbone and (A) I-set backbones or (b) N-set backbones. (c) and (d) The E_{conf} energy landscape of the full-sequence design of (c) I-set or (D) N-set backbones. (e) and (f) The E_{conf} energy landscape of the design of the binding interface of (e) I-set or (f) N-set backbones. In figures (c) – (f), each symbol represents a structure for which a sequence profile was computed. In (c) and (d), structures were grouped into clusters based on the similarity of their sequence profiles from SCADS. Clusters are indicated by distinct symbols and are numbered in order of increasing energy. Color in (a) and (b) represents rmsd (Å) from the native backbone, and that in (c) – (f) represents the E_{conf} energy (kcal/mol) interpolated from the points shown in the figures using Matlab.[52] The labeled black dots in (a), (b), (e) and (f) are structures that were selected for experimental testing.

numbered in order of increasing E_{conf} of the lowest energy profile in each cluster. Thus, structures in clusters with low energies, such as clusters 1 to 3 in the I-set and 1 to 4 in the N-set, are potentially good design templates.

Conserved residues may not be conserved for binding

Figure 3-4 shows SCADS design profiles for positions 11 and 16 on the native backbone and on backbones from the I- and N-sets. For the flexible backbones (I- and N- sets), the profiles were averaged within each cluster shown in Figure 3-3c and d. These two residues are highly conserved in native BH3 sequences as Leu and Asp, respectively, and previous alanine-scanning studies by Sattler et al.[24] have shown that they are important for binding. SCADS calculations on the native backbone also indicated that the native residues are strongly preferred at both positions, as shown in the top panels of Figure 3-4a and b. However, when we included backbone flexibility in the re-design of these positions, phenylalanine, a much larger residue than leucine, was preferred in low-energy clusters at position 11 (clusters 1-3 in the I-set and cluster 4 in the N-set). At position 16, the native residue aspartic acid was preferred on the native backbone and for the lowest energy clusters, but lysine was found to be highly probable in cluster 2 in both backbone sets. Alanine is predicted to be unfavorable at both positions on all backbones, consistent with the alanine-scanning experiments.[24]

These results suggest that the conservation of Leu 11 and Asp 16 may not be due to a strict requirement for binding. To test whether residues predicted to be stable using the flexible-helix backbones are indeed competent for binding, two single mutants, Bim-L11F and Bim-D16K were made and their binding to Bcl-x_L was tested using a solution pull-down assay. Wild-type Bim and human Bim with Leu 11 mutated to Asp (hBim-L11D) were used as positive and

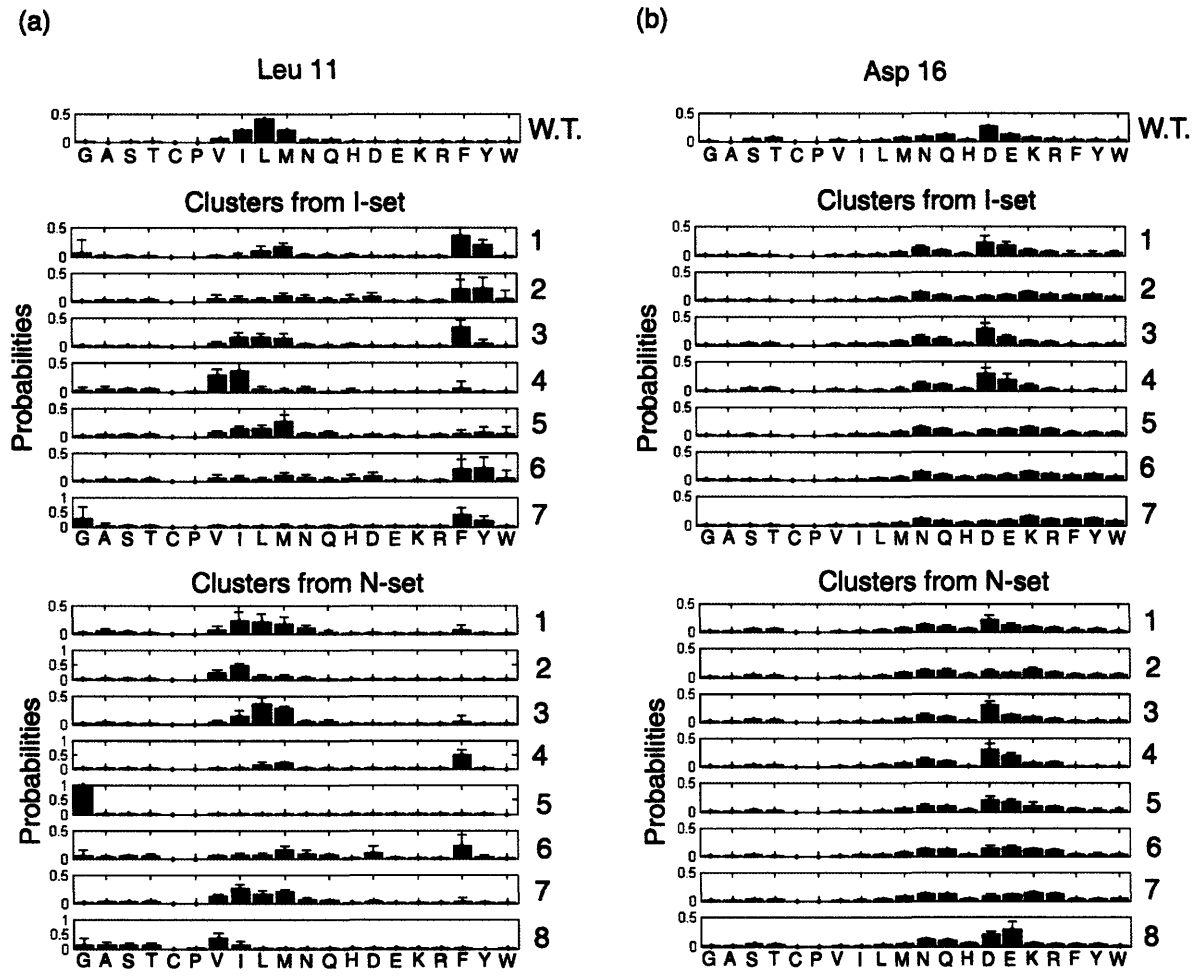


Figure 3-4: Sequence profiles computed with SCADS for positions (a) Leu 11 and (b) Asp 16. "W.T." indicates the native structure whereas numbers indicate clusters of backbones that give rise to similar sequence profiles, as indicated in Figure 3-3. Clusters are numbered by increasing E_{conf}. Profiles for clusters are the average over all structures in that cluster. Error bars show standard deviations.

negative controls, respectively. The results are shown in Figure 3-5. Both single mutants bind to Bcl-x_L approximately as tightly as the native Bim helix.

Design of novel Bcl-x_L binding peptides

As discussed in the Introduction, relieving the fixed-backbone approximation can potentially provide more diverse sequences from protein design calculations than are otherwise available. This is supported by the fact that we could identify point mutations, particularly L11F, that are tolerated at highly conserved positions using flexible backbones, but not the native

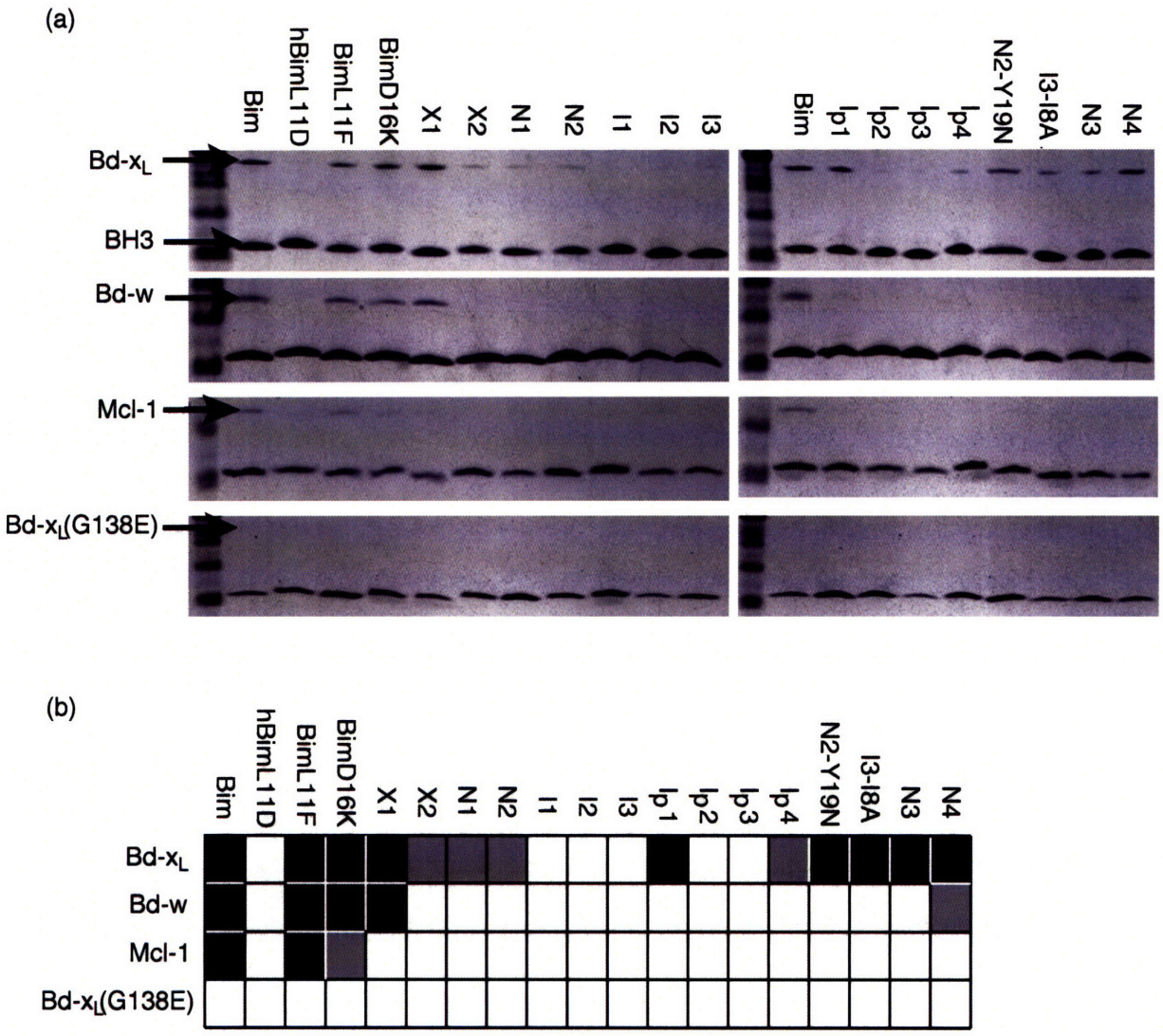


Figure 3-5: (a) Solution pull-down assay for interaction of BH3-like peptides, listed at the top of each lane, with Bcl-x_L, Bcl-w, Mcl-1 and negative control Bcl-x_L G138E, as indicated at left. (b) Summary of (a). Color code: black – strong interaction (similar to the native); grey – weak interaction; white – no discernable interaction.

backbone. To explore this idea further, we redesigned the binding interface of the Bim peptide using the flexible backbone templates. Eleven core and boundary positions were selected for redesign (see Table 3-1). Hydrophobic residues A, F, G, I, L, M, and V were allowed at the core positions, and all amino acids except Cys and Trp were allowed at the boundary positions. Cys was excluded to avoid disulfide bond formation. Trp was excluded to maintain peptide solubility.

Table 3-1. Redesigned positions of Bim. Numbering starts from "1" for residue 2 of Bim chain B in structure 1PQ1.

Subgroup	SASA range	Positions	Allowed residues
Core	<10%	I7, A8, L11, I14, G15, F18	A, F, G, I, L, M, V
Boundary	10%~20%	E4, R12, D16*, N19*, Y22*	A, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, Y
Exposed	>20%	L1, R2, P3, I5, R6, Q9, E10, R13, E17, E20, T21*, T23, R24, R25, V26	Native residues

* These positions were categorized manually based on visual inspection.

Bim residues not in the binding interface were retained with their wild-type identities, but the side-chain conformations were allowed to change.

The I- and N-set backbones were used in this study, along with the crystal structure backbone. Sequences designed using the x-ray structure as a template are referred to as the X-set. We adopted a two-tier design strategy to explore the large sequence-structure space (Figure 3-6). First, SCADS was used to eliminate non-designable backbones and generate profiles of amino acids compatible with each designable backbone. Subsequently, specific sequences were selected using a Monte Carlo (MC) procedure and a different energy function. The two-tier strategy was designed to take advantage of the strengths, and minimize the disadvantages, of these two approaches. SCADS is a method based on the maximization of entropy, and it is ideally suited to identifying the broadest possible set of sequences compatible with a given backbone template at a given design temperature.[53] It is very fast. It can rapidly identify backbone structures that lead to irresolvable clashes or that cannot support good packing interactions. Finally, it has been developed to reproduce patterns of hydrophobic and polar residues that are typical of native structures. Although SCADS has been used alone for many design problems,[2-4] we have found that the results are sensitive to the environmental energy score used (this parameter is constrained to a pre-defined value in these calculations). This can make it difficult to use SCADS

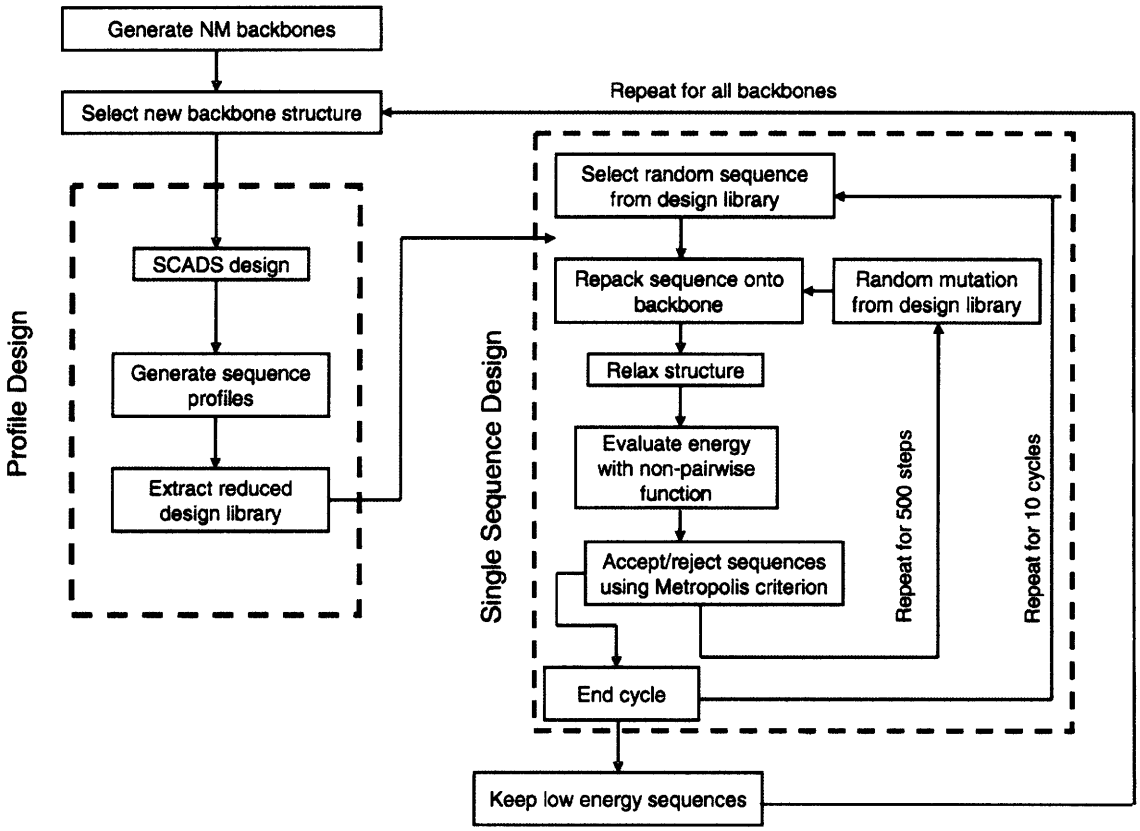


Figure 3-6: Schematic of the two-tiered design method.

to select specific sequences for experimental testing. Thus, we used SCADS to generate restricted amino-acid libraries and analyzed individual sequences selected from these libraries using a Monte Carlo procedure and a different, more physically interpretable, energy function. At each step of the MC search, a structure was generated using side-chain repacking and then relaxed by briefly minimizing all side chains and the helix backbone. This was previously shown to be necessary to give reasonable energies (Figure 3-2). Energies of the relaxed structures were evaluated using the function described in the Methods. The use of two different energy functions with different molecular mechanics parameters for protein design has been suggested to help minimize the error due to biases in either of them individually.[18]

Energies of all sequences visited by the MC search on their respective X-, N- and I-set structures were compared to the energy of the wild-type sequence evaluated in the context of the

crystal structure. Sequences with binding energies lower than the wild-type sequence were considered as possible design candidates and screened further. One hundred and nine sequences were identified using the I-set, and 494 sequences were found from the N-set. Only 35 sequences were found on the crystal-structure backbone. Petros et al. have shown that higher helix propensities for BH3 peptides favor binding.[25] Therefore, we eliminated peptides with helix propensities[54] lower than wild-type Bim from the N-set and I-set. This included 341 sequences from the N-set and 28 sequences from the I-set.

In Figure 3-3e and f, the symbols on the energy landscape indicate I- and N-set backbones on which good design candidates were selected by SCADS. Each symbol represents a backbone. After Monte-Carlo selection, only a few of these backbones, 24 out of 200 in the I-set and 17 out of 200 in the N-set, had one or more sequences that met the two requirements of having lower energy and higher helix propensity than the wild-type structure. Of these, backbones from the N-set had lower SCADS E_{conf} (-640 ~ -620 kcal/mol) than those from the I-set (-600 ~ -560 kcal/mol). The same trend was apparent in energies used for evaluation of single sequences in the Monte Carlo search.

To assess the diversity of sequences generated by this design protocol, all three sets of sequences, N-, I- and X-, were clustered with selected native BH3 sequences using Clustal-X.[55] Only the eleven designed positions were used for clustering. To more clearly visualize the results, we restricted the clustering to the 10 lowest-energy sequences per backbone and up to 50 sequences total for each of the I-, and N- sets (Figure 3-7). Clustering including the entire I- and N-sets gave similar results (data not shown). The 35 sequences in the X-set (blue lines in Figure 3-7) comprise a subfamily of limited diversity. The N-set (green lines) and I-set (red lines) both

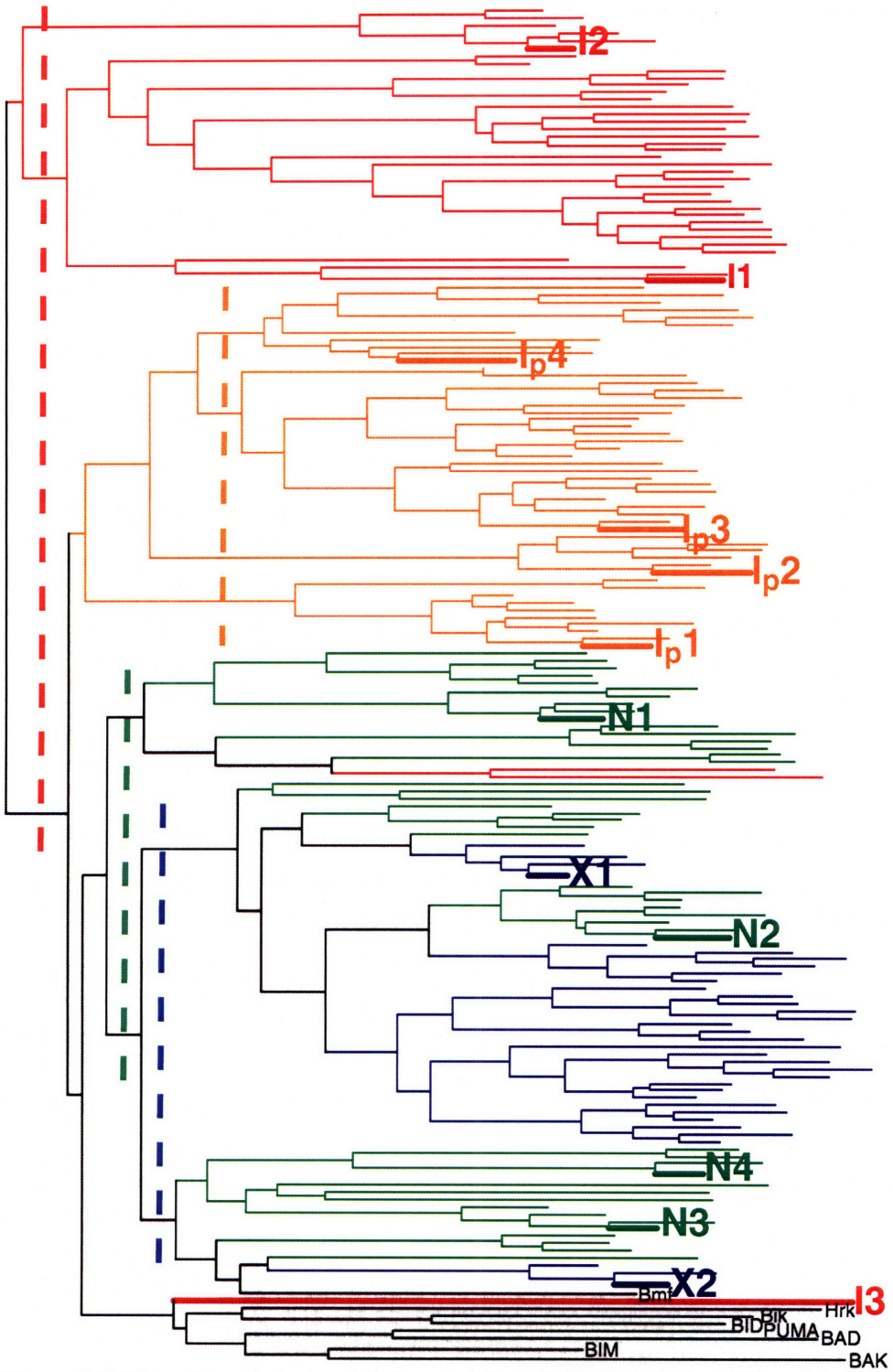


Figure 3-7: Clustering of sequences designed on I-set (red), N-set (green), X-set (blue), and I_p-set (orange) backbones, along with sequences of native BH3 peptides (black) that have been observed to bind Bcl-x_L. The vertical dashed lines show the thresholds used for picking clusters to be tested experimentally.

span a larger space than the X-set, because they contain more backbone structures and provide access to greater sequence diversity.

The results above demonstrate that relieving the rigid backbone approximation can lead to a significantly larger number of sequences that are predicted to have good complementarity with Bcl-x_L and favorable helix propensity. As shown in Figure 3-3a-d, the differences in the backbone can be small but still allow for sequences that would not be designed without the use of an expanded backbone set. There are additional requirements for a sequence to make a good ligand in solution, however. The designed peptide must be soluble, it must not adopt alternative structures not considered in the design procedure, and the energy function used must model not only the bound state but also the unbound state with sufficient accuracy to give high-affinity designs. To test whether our designed peptides met these criteria, the lowest-energy sequences from several clusters in Figure 3-7 were chosen for experimental testing. Thresholds defining clusters for the X-, N- and I-sets, shown as dashed lines in Figure 3-7, were chosen manually to sample the space. The cutoffs give three and two sub-trees for the I- and the N- set, respectively.

Seven sequences were selected for experimental testing: two from the X-set (X1 and X2), three from the I-set (I1, I2 and I3) and two from the N-set (N1 and N2). The sequences selected from the flexible backbones (I- and N- sequences) are shown as the black dots in Figure 3-3e and f. To demonstrate that the I- and N- sequences would not have been identified using the rigid crystal structure, the energies of all sequences evaluated on the crystal-structure backbone and on their respective normal mode design backbones are shown in Table 3-2. When modeled on the crystal structure, the designed sequences are predicted to be at least 8 kcal/mol less stable than the wild-type sequence, with more than 4800 sequences in the combined N-, I- and X-sets

Table 3-2. Sequences designed on flexible backbones that were chosen for experimental characterization.

	Sequences*	energy on NM bb (kcal/mol)	energy on x-ray bb (kcal/mol)
Bim	LRPEIRIAQELRRIGDEFNETYTRRV		0
BimL11F	---E--IA--FR-IGD-FN--Y----	-4.4†	6.6
BimD16K	---E--IA--LR-IGK-FN--Y----	2.2†	6.7
X1	---N--IA--MV-IAR-FH--H----		-2.9
X2	---V--VG--MM-IGR-FF--H----		-2.1
N1	---F--VA--LM-FAH-FD--H----	-14.3	8.3
N2	---V--VA--LM-MGK-FY--M----	-6.7	23.0
I1	---H--IV--FK-FGN-IQ--K----	-10.4	30.8
I2	---F--IA--FM-FAQ-MY--I----	-4.5	26.0
I3	---L--LI--LQ-LGY-FN--A----	-2.6	29.1
I_p1	---A--LA--MR-FAR-FE--A----	-10.7	2.2
I_p2	---R--VA--LE-MAN-LR--N----	-10.3	15.1
I_p3	---I--AA--LQ-FAM-FR--D----	-10.1	13.7
I_p4	---M--IA--LR-FAR-FR--D----	-9.8	7.3
N2-Y19N	---V--VA--LM-MGK-FN--M----	-0.1†	9.5
I3-I8A	---L--LA--LQ-LGY-FN--A----	3.3†	3.3
N3	---V--VG--LM-IAR-FD--T----	-7.8	7.6
N4	---F--VA--LN-IAK-FH--F----	-2.7	4.1

All energies are relative to the energy of the x-ray structure of Bcl-x_L/Bim. Dashes indicate the native Bim residue at that position

* The first Leu was mutated to Glu in experimental tests for all sequences except Bim, BimL11F and BimD16K, see text.

† Sequences for point mutants were evaluated on all backbones. The lowest energy is reported.

predicted to have better binding affinity. Thus, the selected sequences cover a sequence space that cannot be accessed by fixed backbone design.

The designed peptides were tested in a solution pull-down assay. Because previous experiments suggested that designed BH3 peptides can be poorly soluble in aqueous buffers (data not shown), a leucine at the first position of the peptide was mutated to glutamic acid. This site is a surface position and as a result is not expected to influence the binding interaction significantly. Wild-type Bim was used as a positive control and hBim-L11D as a negative control. As a negative control of the receptor protein, we used a Bcl-x_L mutant in which Gly 138, a residue in the hydrophobic binding cleft, was mutated to glutamic acid. The results are shown in Figure 3-5. For the two X-set designs, X1 bound well to Bcl-x_L with X2 binding more weakly. Designed peptides N1 and N2 bound, but more weakly than the positive control. The other three

peptides I1, I2, and I3 did not bind. As expected, none of the peptides, including the native Bim positive control, bound to the Bcl-x_L negative control. We also tested all peptides for binding to anti-apoptotic proteins Mcl-1 and Bcl-w. Pull-down results showed that, except for the X1 design and the two point mutants Bim-L11F and Bim-D16K, none of the designed peptides bound to either protein.

To explore why several peptides from the first round of design did not bind well, we manually designed and tested several point mutants. In most of the native BH3 peptides, position 8 is an alanine or glycine. However two of the I-set designs have a larger side chain at this site. To test whether this could be causing a steric problem, we made an Ile to Ala mutation in design I3. The resulting peptide, I3-I8A, showed improved binding to Bcl-x_L (Figure 3-5). In another case, for design N2, the Tyr residue at position 19 is larger and more hydrophobic than the native asparagine. Gel filtration analysis showed that this peptide eluted somewhat later than native Bim, with a peak that had a long tail, suggesting that it may be sticky and potentially self-associating or aggregating (data not shown). To address this we restored the native Asn at position 19. Again, this peptide bound Bcl-x_L better than the original design (N2-Y19N in Figure 3-5).

All three sequences designed on the I-set backbones performed poorly, suggesting these structures may not be good templates. In our statistical analysis of helices in the PDB we saw that for helices of length 26, the first two normal modes encompass most of the standard deviation (Figure 2-3c) but mode 10 also contributes to the overall difference from the "idealized" helix that we used as a reference (Figure 2-3b). Mode 10 represents a twisting deformation around the helix axis. To test if adjusting the helical pitch would improve the I-set designs, we constructed a new backbone set, the I_p-set, for which the coefficient for mode 10 was

set to the native value of the Bim helix, -6.13. Using this new set, we repeated the design calculations and selected sequences with energy lower than wild-type, giving a total of 249 designed peptides. These sequences were filtered by removing those with helix propensity less favorable than wild type (6 sequences removed), and the 50 lowest-energy sequences remaining were clustered along with the other backbone sets, as shown in Figure 3-7. As with the I-set, the I_p-set designs clustered together, although they were somewhat more similar to the N-set and X-set sequences. Four sequences were chosen for testing by dividing the I_p-set cluster using the dashed yellow line shown in Figure 3-7. Figure 3-5 shows that I_p1 bound Bcl-x_L quite well, I_p4 more weakly, and I_p2 and I_p3 not very much at all. These peptides were also tested against Mcl-1 and Bcl-w; none showed any binding.

Because the N-set designs bound better than the I-set designs, we considered more of these sequences in our next round of experimental tests. We originally chose N1 and N2 from separate clusters, as seen in Figure 3-7, but ignored a third clustering of N-set peptides, because it contained slightly higher energy sequences. We selected two sequences from this cluster, N3 and N4 as shown in Figure 3-7, and found that both bound well to the Bcl-x_L receptor. The binding affinity of these two sequences was also tested against the three other Bcl-2 receptors. N4 showed very weak binding, whereas N3 showed no binding to Bcl-w. Neither of them showed binding to Mcl-1 or Bcl-x_L-G138E. Overall, 12 out of 17 designs considered in this study, which contained from one to eight mutations relative to Bim, showed some level of binding to the Bcl-x_L receptor.

Competition binding experiments

To further characterize the binding of several designed peptides, we tested them in a fluorescence polarization competition assay. Bad-BH3 is a native BH3 peptide that binds in the hydrophobic groove of Bcl-x_L, as determined by previous binding studies and by a solution structure of the complex.[25] In our assay, fluoresceinated Bad-BH3 (FITC-Bad) with a reported K_d of 21.48 nM[56] was competed off of Bcl-x_L by increasing concentrations of Bim, BimL11F, BimD16K, X1, X2, N1, N2, N3, N4, N2-Y19N, I1, Ip1, Ip2, or Ip4. The Bcl-x_L construct used in our assay was slightly different from what was reported, and we measured the K_d of FITC-Bad as 16.7 nM (inset in Figure 3-8). This value was used to fit the competition binding curves shown in Figure 3-8. The K_d values obtained from experiments were: K_d(Bim)=3.2 nM, K_d(X1)=23.4 nM, K_d(N4)=239.7 nM, K_d(I_p1)=73.8 nM, Kd(BimL11F mutant)=0.6 nM, Kd(BimD16K mutant)=33.1 nM, Kd(N2-Y19N)= 73.8 nM, K_d(N2)=490 nM, K_d(I_p4)=790 nM, K_d(N1)=960 nM, K_d(X2)=2400 nM, K_d(N3)=4400 nM. Additionally, I1 and I_p2 showed no binding.

Post-analysis of design results

Given that not all of the designs bound tighter than wild type, as was predicted, we sought to determine what aspects of the calculation was responsible for this. We had noticed that none of the I-set sequences bound to the Bcl-x_L receptor. To determine why, a new method for post-processing of the experimentally tested sequences was introduced to determine whether we could better predict binding affinity. Previously we minimized each helix to allow for local deformation. However, the flexibility introduced by the minimization lowered the energies of I-set backbones, perhaps to an artificially low value (Figure 3-2). This was likely due to the

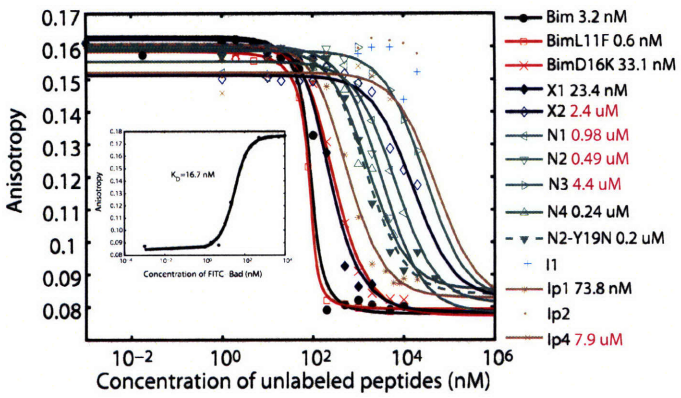


Figure 3-8: Competition binding curves from fluorescence polarization assay. The assay was done at room temperature with murine Bcl-x_L at 100 nM and FITC-Bad at 25 nM. Competing peptides Bim, BimL11F, BimD16K,, X1, X2, N1, N2, N3, N4, N2-Y19N, I1, Ip1, Ip2 and Ip4 are labeled according to the legend on the right . The lines are binding curves fit to the data points. C_L is the initial concentration of the competitor peptide. The inset shows the direct binding of FITC-Bad to murine Bcl-x_L (K_d = 16.7 nM). Here C_R is the initial concentration of Bcl-x_L. The K_d values of the competing peptides (MBM , X1, N4, and Ip1) were derived using our measured K_d (Bad/Bcl-x_L). Competition binding measurements gave the K_d values shown in the legend to the right. I1 and Ip2 did not show binding.

inability of the energy function to accurately describe the relative stability of vastly different backbones. As a way to introduce the flexibility without over-minimization, we used a procedure where each sequence was repacked on all backbones in the set. Additionally, we only allowed the side chains to relax during minimization.

Using this approach we evaluated all 14 sequences with experimental K_ds, along with two non-binders from the pulldown assay, I2 and I3. There were four non-binders in total, which were assigned a K_d value of 10^5 nM, which is the approximate limit of the competition binding assay. Experimental binding data is plotted against the minimum predicted energy of each sequence over all backbones (Figure 3-9). This procedure gives a much better correlation with experimental energies, with an R^2 value of 0.79, and could be used in future rounds of design to improve the prediction performance and the selection of candidate sequences to test experimentally.

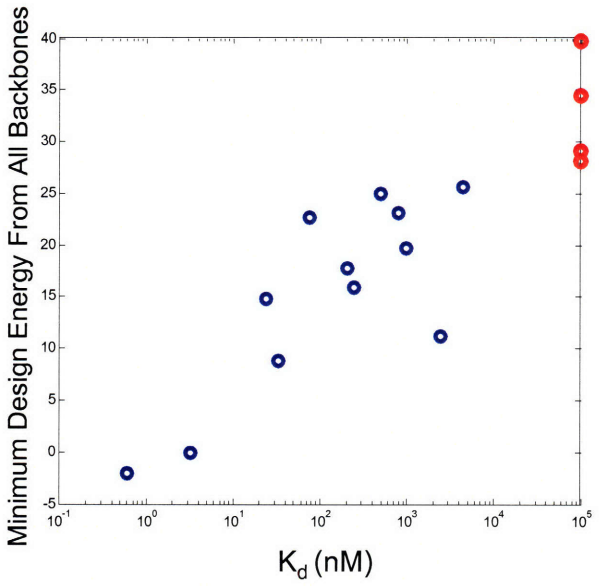


Figure 3-9: Comparison of predicted to experimentally determined binding affinities. The minimum energy for each sequence evaluated on all backbones is compared to K_d values obtained through the competition binding. The calculated affinities are all shown relative that of Bim. In blue are data for Bim, BimL11F, BimD16K,, X1, X2, N1, N2, N3, N4, N2-Y19N, I1, Ip1, Ip2 and Ip4 that showed binding. In red are data for sequences that showed no binding. These non-binders are assigned a binding affinity of 10^5 nM, which is below the detection limit of the competition binding assay.

Designing with side-chain only minimization

Given the success of the post-analysis, we performed a new round of designs using only side-chain minimization and leaving the helix and receptor backbone fixed. In addition to designing for binding to Bcl-x_L, this round also included the design of BH3 peptides to bind to Mcl-1, based on crystal structure 2NL9 (this structure was solved after the initial rounds of design). One issue with this was that in the Mcl-1/Bim structure, Bim contains only 23 residues, 22 of which overlap those found in the Bcl-x_L/Bim structure. Thus 22-mers were designed in this round using the N-set backbones. 200 native-like normal mode backbones that had 22-residue BH3-helices were generated for both the Bcl-x_L/Bim and Mcl-1/Bim structures. With the exception of the minimization step, the same procedure was utilized as in the last round of design. The 200 lowest energy sequences for both Bcl-x_L and Mcl-1 design were re-evaluated using the post-processing step described above. Four low energy sequences were chosen that were

significantly different in sequence identity to those previously validated. Three were designed to bind to Bcl-x_L and one to bind to Mcl-1 (Table 3-3). So far, these sequences have only been tested for binding to Bcl-x_L, and all bound weakly. This may have resulted because the design and the post-processing procedures were very similar and had the same biases. In evaluation of the previous rounds, the sequences were designed and evaluated using different methods. By having different minimization protocols, each method had different biases. As described above, Harbury et al. suggested the benefits of using orthogonal energy functions in design.[18] Perhaps the same is true for minimization procedures.

Discussion

Previous studies aimed at designing protein-protein interactions have focused primarily on identifying one or a few high-affinity and/or heterospecific complexes, often by re-engineering the sequence of both binding partners.[5-7,17,18] There are only a small number of examples in which a protein or peptide has successfully been designed to bind a native target.[8-10] Here, we report the successful design of several new 26-residue peptides that bind to Bcl-x_L. The designs exploited a new method for sampling backbone flexibility using NM analysis. In three rounds of computation and experimental testing, we gained insights into features of the BH3 sequences that are and are not important for binding. We also uncovered important considerations for sampling helical backbone structures. In this section we discuss these issues, as well as the general importance of including backbone flexibility in protein design and some possible areas for future improvements.

Table 3-3. Sequences designed with side-chain only minimization chosen for experimental characterization.

	Sequences*	Design target	Binding Affinity to Bcl-x_L (K_D)
Bim	RPEIRIAQELRRIGDEFNETYT		>1 nM
PE1	--H--VA--LN-AGR-FD--H-	Bcl-xl	1800 nM
PE3	--N--VA--LK-FGR-FD--R-	mcl-1	4200 nM
MN1	--E--MA--VM-IAE-MS--K-	mcl-1	6400 nM

Backbone Templates

Carefully selected backbone structures are key for structure-based computational design.[16] Although native backbone structures determined by x-ray crystallography have been successfully used in many cases, they have obvious limitations. One is that sequences designed on a fixed native backbone are strongly biased by the exact atomic coordinates of the selected structure, as shown in Figure 3-4 and Figure 3-7. However, fixed-backbone design has been successful partly because starting with an x-ray crystal structure guarantees that the template is designable. When flexible or *de novo* backbones are used, additional criteria are needed to select a designable scaffold. Our goal in this study was to increase the sequence space that could be accessed in protein design by introducing backbone flexibility in a way that sampled realistic structures. NM analysis has been shown to be effective for describing structural deformations of helices,[46] and we found that this was also a convenient way to generate structural variants for design. We used this approach to identify a wide range of candidate BH3 ligands for Bcl-x_L.

From our initial round of design, only two of the five peptides that we tested bound to Bcl-x_L. The two that bound were designed from the native-like N-set, and those that did not bind were from the I-set. Additionally, we were able to design binding peptides using the crystal structure as a template. This suggested that the I-set did not provide good templates. The I-set structures were derived *de novo* from an idealized helix backbone using only the two lowest-frequency normal modes to generate structural variation. However, these two modes capture less

than half of the deviation between our reference helix and α-helices in the PDB. For helices of length 26, ~70% of the deformation from the ideal helix can be captured by modes 1, 2 and 10, with mode 10 corresponding to changing the pitch of the helix. The contribution of mode 10 to helices of length 26 is approximately constant (Figure 2-3b and c) and indicates that the pitch of our "ideal" helix is larger than what is found in the PDB. Consistent with this, we found that when we minimized the I-set helices as part of the design procedure, the value of mode 10 changed to be closer to the average value in the PDB. We postulated that modifying the I-set structures to reflect the native value of mode 10 in the Bcl-x_L/Bim structure could improve the quality of the templates. A new I_p-set was used to design four peptides and resulted in two that did bind Bcl-x_L. This suggests that using an 'ideal' helix to construct a new backbone set can be an effective strategy, as long as the pitch is set appropriately.

The I-set sequences that we chose for experimental characterization were scored as low in energy by our design procedure, yet they failed to bind Bcl-x_L. This occurred despite the fact that for the native sequence we were able to distinguish I-set backbone models as higher in energy than N-set models (Figure 3-2). We were also able to relax the I-set backbones towards more native-like structures in the Monte Carlo design procedure. That our energy function was reasonably effective for prediction but showed deficiencies in design is not necessarily surprising. For example, if van der Waals, electrostatic interactions and ϕ and ψ dihedral strain are not balanced, it is possible that the design procedure could systematically exploit this to introduce unrealistic interactions that compensate for poor backbone geometry (e.g. see below). Choosing a backbone set, such as the N-set, that samples more realistic structures can help to address this. It is interesting to note that the choice of energy function and the method for varying backbone

structure may be linked; shortcomings in one can be partially compensated for by adjustments in the other.

Although we successfully introduced flexibility in the binding BH3 helix, the Bcl-x_L receptor was held fixed. It is clear from available NMR and X-ray structures of Bcl-x_L bound to BH3 peptides, [24-26,38,57] as well as to small molecules,[37,58] that there is some variability in the structure of α-helices 3 and 4, which form part of the binding site. This is another degree of freedom that could be sampled to further increase the design diversity. Although normal-mode analysis may not be an efficient way to sample the irregular structural changes involved in this region, one strategy could be to use existing experimental structures as a guide. Qian et al. have shown that principle-component analysis can be used to sample natural variation efficiently, when this is represented by a set of existing structures.[59] With several Bcl-x_L complex structures available,[26,38,57] and more likely to be solved in the future, this represents a possible route towards designing yet more diverse BH3 peptide ligands.

Analysis of designed BH3 sequences

Native BH3 peptides are quite diverse and have only a weak consensus: ------h{A/G}--L--h{G/A}D-h----------, where "h" represents a hydrophobic residue, {x/y} indicates that residues x and y are commonly found at a given site, and "-" indicates no strong consensus. Leu 11 and Asp 16 are the most strongly conserved residues and are present in all native BH3 peptides that are known to bind Bcl-x_L. Our first round of design calculations (using SCADS only) indicated that despite being strongly conserved, Leu 11 and Asp 16 are not strongly favored at their respective positions once backbone flexibility is considered. Slight backbone motions can accommodate the larger Phe residue at position 11 (see Figure 3-10), and several backbones favor Lys over Asp at

position 16. Experiments confirmed that the dramatic sequence changes of Leu to Phe at position 11 and Asp to Lys at position 16, do not disrupt binding of Bim to Bcl-x_L. Thus, these residues are probably conserved for some reason other than maintaining binding affinity to this target.

Two other sequence changes suggested by the designs also contradicted the consensus sequence. These were the designs of a Val or Ile residue at position 8, a site normally occupied by Ala or Gly. Peptides I1 and I3 with these substitutions were designed using the I-set backbones and, when tested experimentally, failed to bind Bcl-x_L. A point mutation of Ile 8 to Ala in design I3 restored binding. Thus, it seems that a small residue at position 8 is probably a requirement for binding Bcl-x_L. Our energy function indicated that Ile or Val at this site could form favorable interactions with the receptor, but only in the context of the I-set backbones. Such β-branched residues could not be accommodated at this site using the N- or X-sets, and only four of the top 50 I_p-sets sequences have valine. Our energy function did not correctly balance the reward of a favorable van der Waals interaction with a suitable penalty for the I-set backbones having an inappropriate pitch. We addressed this by introducing the I_p set, restricting our backbone search to more realistic structures.

In total, our 12 BH3 designs spanned a significant sequence space. All designs (other than the point mutants) had 6-8 sequence changes from native Bim, out of 11 interface positions. All of the designed sequences maintained the four conserved hydrophobic residues that pack into Bcl-x_L, but the identities of these varied according to the backbone structures on which the sequences were designed. Boundary residues varied more significantly, with charged residues such as Glu 4 and Asp 16 in Bim sometimes being replaced by hydrophobic or oppositely

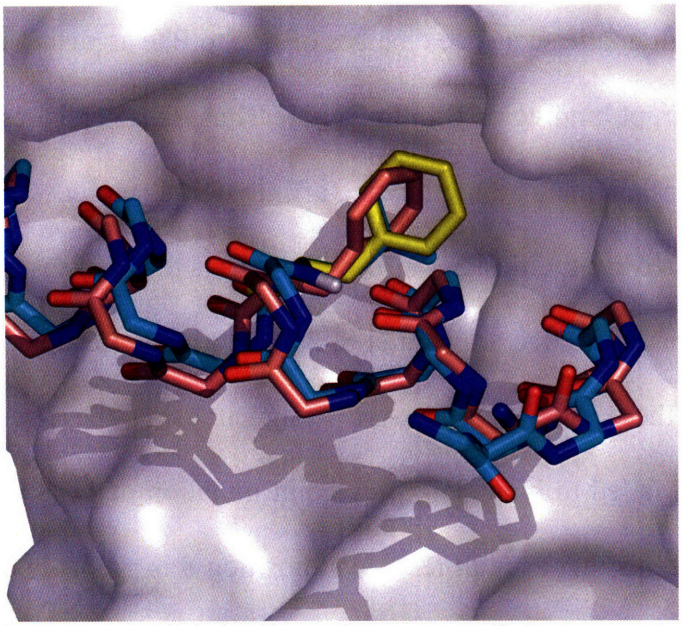


Figure 3-10: Slight backbone changes are sufficient to accommodate the Leu to Phe mutation at position 11 of Bim. Bcl-x_L is shown as a surface representation colored in gray. BH3 backbones and the side chains at position 11 are shown using sticks. The native BH3 backbone and Leu 11 side chain are shown in cyan. The L11F mutant side chain modeled on the native backbone is shown in yellow. The L11F backbone and side chain designed on an N-set normal mode-generated backbone are shown in pink. Figure generated using Pymol.

charged residues. Such changes of residue type may be particularly important for designing BH3 ligands with altered binding specificity (see below).

Backbone flexibility for specificity design

In signaling pathways leading to apoptosis, the binding specificity of native BH3 peptides for multi-domain anti-apoptotic Bcl-2 family members is a key factor in triggering cell death.[29-31] In particular, it is important whether BH3 peptides bind to all or to only a subset of the anti-apoptotic proteins. It would be useful to design synthetic peptides with desired binding specificity profiles, e.g. peptides that bind to Bcl-x_L but not Bcl-w or Mcl-1, in order to understand and manipulate the interactions of these proteins. It might be possible to engineer specificity profiles directly, using a multi-state design procedure. Initially though, we only had the x-ray structure of Bcl-x_L/Bim to use as a template (we have found that available NMR

structures do not give good results in repacking calculations), and our ability to design novel specificity profiles was hindered by a strong bias that causes designed sequences to resemble native Bim in core positions, and have low sequence diversity in all design sites. Including multiple backbones can counteract this structural bias and provide access to a larger sequence space, a space that potentially includes sequences with novel specificity profiles, as illustrated in Figure 3-5.

Our results support this idea. Native Bim is promiscuous and binds to all anti-apoptotic Bcl-2 family members, including Bcl-x_L, Mcl-1 and Bcl-w (Figure 3-5). The two designed point mutants, BimL11F and BimD16K, which are similar in sequence to native Bim, both bound Bcl-w. BimL11F also bound Mcl-1, whereas BimD16K bound Mcl-1 very weakly. In contrast, when 11 positions were redesigned on a range of backbones, only one sequence designed from the crystal backbone (X1) bound Bcl-w, and one from a native-like backbone (N4) bound Bcl-w very weakly. None of the designed sequences show detectable binding with Mcl-1.

Since the completion of this design experiment, several other x-ray crystal structure of Bcl-2 family complexes have been published including Bcl-x_L complexes with Bad and Beclin,[38,57] Mcl-1 complexes with Bim and Noxa,[39] and A1 complexes with Bim, Puma, Bmf, Bak and Bid.[60,61] In the future, it may be possible to use these additional crystal structures to select directly for sequences that bind to certain anti-apoptotic Bcl-2 family members but not others. In such an application, the ability to model backbone flexibility will remain very important. First, with increasing demands on the designed sequences, artificial constraints on the space of possible solutions become less acceptable. Furthermore, backbone flexibility is a critical element of negative design against undesired "decoy" targets. A common problem in negative design is that decoy states must be modeled and their energies accurately evaluated. With fixed

backbone design, this is problematic because structures may have high energies based on slight steric clashes that are easy to resolve with backbone relaxation or flexibility. The BimL11F mutant provides a good example of this (see Figure 3-10). If the complex of Bcl-x_L with Bim was a negative design target, then fixed backbone design would predict that Phe at position 11 would disfavor this structure. In contrast, we find that BimL11F binds well to Bcl-x_L.

Possible directions for future improvements

In this work, we used a range of starting structures as templates for design, with the goal of generating a set of peptides with diverse properties that bind to Bcl-x_L. Practical considerations led us to constrain our search to a sequence space identified as favorable by SCADS, and to use a fairly slow non-pairwise energy function for evaluation. Thus, in an attempt to sample broadly, we have sacrificed local optimization. Although we found many good sequences, we may not have identified minima in either structure or sequence space. A possible strategy for the future is to use sequences from experimentally validated clusters as starting points for further rounds of design. Additionally, Baker and colleagues have demonstrated the power of iteratively optimizing sequence and structure.[16,43] A similar approach could help to identify tighter binding sequences in the space of NM-sampled backbones. Finally, energy functions that are suitable for fixed-backbone design may not be optimal for flexible-backbone design. Further work may be needed to determine how best to balance the internal energy of the template with the interaction energy of the designed side chains. Sampling normal modes in dihedral space rather than Cartesian space may generate backbones that better maintain ideal bond lengths and angles, while retaining suitable dihedral values. Nevertheless, the use of NM analysis focuses backbone sampling to a realistic parts of structure space very efficiently, using

only 2-3 parameters. A set of realistic templates reduces the burden placed on the search and evaluation functions in design. The NM strategy can be used to sample variation of any α-helices present in a design template. Further, the use of NM analysis, which has a very general formulation, may extend well to other secondary or super-secondary structural elements.

The initial design results were promising, but there is much room for improvement in the accuracy of the prediction. This prediction gave 12 out of 17 peptides that bound to Bcl-x_L, but only one that actually had a lower K_d value than the native sequence. Post-processing of the sequences by modeling them on all backbones showed better correlation to the experimental data. This could prove to be a useful method for future rounds of design, and hopefully will improve the hit rate of candidate sequences selected for experimentally testing. It has been suggested that for design, over-optimization of a single function can lead to strong biases.[18] However, these can be overcome by screening solutions with an orthogonal function.[18] The fact that the design process and the post-process do not use identical structural models may indeed help to improve the ability to model particular sequences. Another improvement could include the incorporation of experimental data into the design method. This has shown much promise in optimizing energy functions for particular types of targets.[19,62,63] In addition to our experimentally-tested sequences, there are a number of systematic experiments that have measured the binding affinity of point mutants of Bcl-2 family complexes.[64-67] These data could be used to optimize the relative magnitude of different energy components and the unfolded state energies, or to replace particular interactions that are not presently modeled well. Improvements in the energy function could help to increase the confidence in our design selections.

Methods

Construction of flexible backbone structures

Starting templates

The 1.65 Å resolution structure of Bcl-x_L/Bim was used as a template (PDB code: 1PQ1).[26] The backbone of Bcl-x_L, chain A, was held fixed. The backbone of chain B, corresponding to Bim, was varied. The 1.55 Å resolution structure of Mcl-1/Bim was also used as a template (PDB code: 2NL9).[39] The backbone on Mcl-1, chain A, was held fixed. The backbone of chain B, corresponding to Bim, was varied. An ideal α-helical backbone (φ = -57.0°, ψ = -47.0°, and Ω = 180.0°)[68] was generated using default bond lengths and bond angles from CHARMM param19.[69,70] The C, C_α and N backbone atoms of the ideal helix were aligned to chain B in the crystal structure using ProFit 2.2.[71] The side chains were generated using param19 values for bond angles and bond lengths and the crystal-structure dihedral angles. The original x-ray structure and the structure with chain B replaced by an aligned ideal helix were both used as starting templates.

Backbone variations in normal-mode space

The details of how to sample normal mode space and generate new structures are described in detail in chapter 2. Based upon equation 2-5, we varied the set of normal mode parameters $\{a_i\}$ to generate new sets of backbones. Three choices of $\{a_i\}$ values were used to construct backbone sets in this study. Abbreviations for, and descriptions of, these sets are summarized in Table 3-4. For the I-set (ideal-helix set), all values of a_i except for the two lowest frequency ones were fixed as zero, corresponding to the NM values of an ideal helix. Helices in

Table 3-4. Abbreviations and descriptions for backbone sets used in design calculations.

Set Name	Starting Structure	Modes fixed to 0	Modes fixed to Native
X-set	Native helix	None	All
I-set	Ideal helix	All except 1 and 2	none
N-set	Native helix	None	All except 1 and 2
Ip-set	Ideal helix	All except 1, 2 and 10	10

the I_p set (ideal-helix with native pitch) were constructed in the same manner as the I-set, except that the 10^{th} lowest frequency normal mode, a mode corresponding to the change of the helical pitch, was set to the crystal structure value of -6.13. Finally, for the N-set (native-helix set), all a_i's with i greater than 2 were fixed as those of the native helix. To determine the NM values of the native helix, a difference vector between the native helix and the aligned ideal helix was calculated. This vector was fit to a linear combination of NM vectors using linear regression. The fitted linear coefficients gave the $\{a_i\}$ of the native helix.

Design Calculation

Two types of design calculations were preformed. In the first, Statistical Computationally Assisted Design Strategy (SCADS), developed by the Saven group,[2,44,72] was used to rapidly characterize the sequence and structure space of α-helical ligands of Bcl-x_L. In the second, a two-tiered strategy was implemented to select single sequences for experimental testing. The two tier-procedure included a SCADS profile design, used to narrow the library of amino acids, followed by a single-sequence Monte Carlo design (Figure 3-6). In SCADS, the AMBER force field,[73] with a united atom representation, was used to calculate non-bonded interactions. A statistical environmental score (E_{env}) was included as a constraint to enforce the hydrophobic patterning of native proteins.[44] A tri-peptide model was used to approximate the unfolded/unbound state of the BH3 peptide.[2] The Richardson-Richardson rotamer library[74] was used with the χ_1 angles of Phe, Trp and Tyr expanded by ± 5° and ± 10°, increasing the total

number of rotamers to 254. Bcl-x_L residues with at least one atom located within 10 Å of any atoms of the helix were allowed conformational flexibility. All other residues were held fixed with the crystal-structure coordinates. Sequence profiles, in the form of a set of amino-acid probabilities at each site, were obtained for each backbone structure. A conformational energy (E_{conf}) for each profile was evaluated by averaging non-bonded mean-field energies at each position, weighted by the appropriate amino-acid probabilities. E_{conf} consists of side chain – side chain and side chain - backbone terms and was evaluated at β=0.3 $(kcal/mol)^{-1}$, where β is an effective inverse temperature.

The second tier of design utilized a Monte Carlo strategy. Here a subset of amino acids was chosen based upon SCADS sequence profiles. For each site, the number of amino acids included in the design was determined by the site-specific sequence entropy S_i.

$$S_i = \sum_{\alpha_i} p_{i,\alpha_i} \ln p_{i,\alpha_i} \tag{3-1}$$

Here p_{i,α_i} is the probability of a particular amino acid α_i at site i derived from the SCADS calculation. The probabilities were rescaled from the original β=0.3 calculation to β=1.0 to limit the sequence search to high probability amino acids. The top n_i ($n_i = \exp(S_i)$) most probable amino acids were included in the design at each site (n_i was rounded up to the nearest integer). Using this limited amino-acid library, 10 independent runs of 500 steps of Monte Carlo design were performed for each structure. For each Monte Carlo design step in sequence space, we performed a repacking calculation to model the side-chain conformations, followed by an energy evaluation step to guide the Metropolis sampling. Structures were repacked as described by Ali et al.,[17] with a few modifications. The energy function included CHARMM van der Waals energy with the atomic radii scaled to 90%, EEF1[75] for solvation, distance dependent dielectric

electrostatics with ε=4r, and CHARMM torsional energies. The same rotamer library as for the SCADS calculation was used. All helix residues and all receptor residues within 8 Å of the helix were allowed conformational flexibility. All other residues were held fixed with the crystal structure coordinates. Sequence repacking was performed using Dead-End Elimination (DEE) and the A* algorithm.[76-81] Following repacking, we minimized the structure using CHARMM with 1000 steps of steepest-decent minimization and 1000 steps of adapted bases Newton-Raphson. The energy function for minimization included the van der Waals energy with 100% van der Waals radii, bond angle, bond length, dihedral angle and improper dihedral angle molecular mechanics energies, and ε=r distance dependent dielectric electrostatic interaction energy. The receptor backbone atoms were fixed during minimization. Finally, a non-pairwise decomposable energy function was used to evaluate the energy of the minimized structures. This energy was used to guide the Monte Carlo search. It included terms for van der Waals interactions with 100% van der Waals radii, finite-difference Poisson-Boltzmann (FDPB) solvation energy, Coulombic electrostatic interactions with external and internal dielectric of 4, and a solvent-accessible surface area cavitation energy with a proportionality constant of 10 cal/mol•Å^2. The van der Waals and Columbic energy terms were evaluated using CHARMM and the FDPB calculations using DelPhi V.4;[82,83] the surface area was calculated using NACCESS.[84]

In accord with experimental observation,[25] we modeled the unfolding pathway as a transition from the bound complex (R-H) to an isolated receptor (R) and a random coil (RC). The energy of this transition can be described as follows.

$$\Delta E_{total} = E_{R-H} - \left(E_{RC} + E_R\right) \tag{3-2}$$

The energy of the isolated receptor is the same for all design calculations and can be ignored. It is difficult to evaluate the energy of a random coil, but the contribution of an amino acid to the

transition from a random coil to a folded helix can be captured using experimentally determined helix propensities. Helix propensities are reasonably context independent, with good agreement found between measurements made in different environments.[85,86] Consequently we define $E_{H_i^*}$ as the self energy of a single amino-acid side chain *i* in the context of a poly-alanine peptide, and write:

$$\sum_i HP_i \approx \sum_i E_{H_i^*} - E_{RC} \tag{3-3}$$

HP_i are helix propensities as measured by O'Neil and Degrado[54] and $E_{H_i^*}$ was calculated using the same repacking, minimization and energy evaluation described above. Given our original folding equation, we can add and subtract $\sum_i E_{H_i^*}$ giving

$$\Delta E_{Tot} = (E_{R-H} - \sum_i E_{H_i^*}) + (\sum_i E_{H_i^*} - E_{RC}) \tag{3-4}$$

or

$$\Delta E_{Tot} = (E_{R-H} - \sum_i E_{H_i^*}) + \sum_i HP_i \tag{3-5}$$

E_{R-H} was calculated explicitly as the energy of the complex. ΔE_{tot} is the energy used in the MC design procedure.

Selection of the design positions

Although there are 33 residues in the B chain of Bcl-x_L/Bim structure 1PQ1,[26] some residues at the N- and C- termini do not make direct contact with the receptor protein. In the design calculations, we considered residues 2 to 27, and re-numbered these as 1 to 26. In an initial set of SCADS calculations, all twenty-six residues from chain B were designed and allowed to be any amino acid. When designing individual sequences with our two-tier procedure, only residues at the binding interface were re-designed. The binding interface was defined based

on solvent-accessible surface area (SASA) calculated by NACCESS,[84] followed by manual inspection. Design positions for these calculations, and residues allowed at each position, are given in Table 3-1.

Characterization of sequence space

Comparing sequence profiles

A sequence profile can be either a set of site-specific probabilities, such as those obtained from multiple sequence alignment, a SCADS design calculation, or a single sequence, which is equivalent to a profile with all site-specific probabilities either 1 or 0. The sequence similarity score defined by Panchenko et al.[87] (ΔSS) was used to compare pairs of sequence profiles (*p1* and *p2*):

$$\Delta SS = SS - \overline{SS} \tag{3-6}$$

where SS is a raw pair-wise similarity score, and $\overline{SS}$ is a reference sequence score.[87] Only sequences with the same chain length (L) were studied in this work. SS and $\overline{SS}$ were calculated with the following equations:

$$SS = \frac{1}{L}\sum_{i=1}^{L} score(p_{1,i}, p_{2,i}) \tag{3-7}$$

$$\overline{SS} = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L} score(p_{1,i}, p_{2,j}) \tag{3-8}$$

with $score(p_{1,i}, p_{2,j})$ characterizing the sequence similarity between profile 1 (*p1*) at position i and profile 2 (*p2*) at position j. The BLOSUM62 substitution matrix (M) was used to evaluate the similarity of any pair of amino-acid residues ($\alpha_{p1,i}$ and $\alpha_{p2,j}$):

$$score(p_{1,i}, p_{2,j}) = \sum_{\alpha_{p1,i}=1}^{20} \sum_{\alpha_{p2,j}=1}^{20} P(\alpha_{p1,i}) P(\alpha_{p2,j}) M(\alpha_{p1,i}, \alpha_{p2,j}) \quad (3\text{-}9)$$

Sequence clusters

X-cluster[51] was used to cluster sequence profiles by their sequence similarity scores. The k-mean algorithm was used to find the clusters. Up to ten clusters were defined for all pairs of profiles. Clustal-X[55] was used to cluster single sequences. Only the eleven interface residues listed in Table 3-1 were used in the clustering calculations.

Experimental methods

Sample preparation

Twenty-six-residue peptide ligands (BH3 peptides) were constructed using gene synthesis. Oligonucleotides were designed using DNAWorks 3.0[88], with 5' BamHI and 3' NotI restriction sites and ordered from IDT. Standard PCR conditions were used to synthesize genes, using temperatures suggested by DNAWorks. The PCR reaction products were cloned into a modified pDEST17 vector, containing an N-terminal His_6 tag, a tobacco etch virus (TEV) cleavage site and a C-terminal flag tag, giving the sequence: SYY-HHHHHH-LESTSLYKKAGSGS-ENLYFQ-GGS-**BH3**-GGR-DYKDDDDK. Peptides were expressed in *E. coli* RP3098 or BL21 cell. The expressed peptides were purified by Ni-NTA affinity chromatography followed by HPLC to greater than 99% purity. The molecular weights of the purified peptides were confirmed by mass spectrometry and were accurate to within 1% of the expected molecular weight.

Murine Bcl-x_L (provided by G. Nunez, University of Michigan), residues 1-209, which excludes the C-terminal transmembrane domain, was sub-cloned by PCR with 5' BglII and 3'

XhoI sites. The fragment was ligated into a modified pDEST17 vector, containing an N-terminal His_6 tag followed by a TEV protease cleavage site, using BamHI and XhoI sites. The protein was expressed in BL21-pLysS cells. The protein was purified by Ni-affinity chromatography (Ni-NTA agarose) under native conditions, followed by ion-exchange chromatography using Q sepharose.

The human Bcl-x_L negative control construct was generated by PCR amplification of two halves of the Bcl-x_L gene, 1-138 and 138-209, mutating residue 138 from Gly (5'-GGT-3') to Glu (5'-GAG-3'). The two halves were combined by overlapping extension with end primers containing 5' BglII and 3' XhoI sites. The Bcl-x_L G138E mutant DNA was ligated into pSV282 (using 5' BamHI and 3' Xho), a vector containing an N-terminal His-tagged MBP (maltose binding protein) followed by a TEV protease cleavage site. Human Mcl-1 was subcloned, removing the N-terminal PEST domain and C-terminal transmembrane domain. Residues 166-327 were PCR amplified with 5' BamHI and 3' XhoI sites and ligated into pSV282. Human Bcl-w, residues 1-176, was cloned into pSV282 following the same protocol as for Mcl-1. The human clones of Bcl-x_L and Mcl-1 were obtained from J. Kramer, Harvard Institute of Proteomics. The cDNA of human Bcl-w was provided by D. Huang at WEHI in Australia. The pSV282 vector was provided by L. Mizoue at Vanderbilt University, Center for Structural Biology.

The human Bcl-x_L negative control, Mcl-1 and Bcl-w were expressed in BL21 pLysS and purified by Ni-affinity chromatography under native conditions. Ni-purified proteins were cleaved with TEV protease (~1 μM) in a buffer containing 50 mM Tris, 50 mM NaCl, 0.5 mM EDTA at pH 8.0 for 2.5 hours at room temperature. The untagged TEV cleavage product was purified by Ni-affinity chromatography, separating it from His-tagged MBP and TEV. The Bcl-

x_L and Mcl-1 proteins were further purified by gel filtration chromatography with an S75 column. The Bcl-w protein was purified on a Q sepharose column.

Solution pull-down assay

All pull-down experiments were conducted in TBS buffer (150 mM NaCl, 50 mM Tris, pH=7.4) containing 0.1% Triton X-100 using 200 μM of the receptor proteins and 12 μg/ml of the peptides. Mixtures of the receptor proteins and BH3 peptides were incubated at 4 °C on a rocker for one hour before a fixed amount of α-flag beads (Sigma) was added. The protein and bead solutions were incubated at 4 °C on a rocker for another 30 minutes. Washes and elutions were done following the manufacturer's protocol. Elution fractions were analyzed on polyacrylamide gels stained with Coomassie dye.

Fluorescence polarization assay

Fluoreseinated-Bad (FITC-Bad, Abbott) [56] was dissolved in DMSO at 500 nM. Bcl-x_L and the competing peptides are the same as described above. Both Bcl-x_L and the peptides were dissolved in binding buffer (20 mM phosphate buffer, pH 7.5, 50 mM NaCl, 1mM EDTA, and 0.001% triton X100). The concentration of the Bcl-x_L stock was measured at 280 nM in Edelhoch buffer.[89] Peptide concentrations were measured directly in the binding buffer due to limited solubility.

Direct and competition binding assays were conducted at 25°C in the binding buffer as described.[56] In all samples, FITC-Bad was present at 25 nM, with 5% DMSO. In the competition binding assays, the concentration of Bcl-x_L was fixed at 100 nM. For direct binding, the samples were equilibrated for at least 30 minutes. For the competition binding, the samples were

equilibrated for at least three hours. Fluorescence polarization measurements were done using a PTI QM-2000-4SE spectrofluorometer (Lawrenceville, New Jersey) with excitation wavelength of 485 nm, and emission wavelength of 517 nm. A model considering depletion of the labeled peptides was used to fit the direct binding data, and a model considering depletion of both the labeled and unlabeled peptides was used to fit the competition binding data (see online supplemental material for the fitting models). The ability to determine the saturated baselines was limited by the solubility of the peptides. A single additional data point at [competition peptide] = 1 mM was added with an anisotropy value determined by averaging the values of Bim at 1000 and 2000 nM before fitting the competition curves. Experiments were done in duplicate, with one replicate shown in Figure 3-8 and the range of measured K_d's given in the figure.

Acknowledgements

We would like to thank J.G. Saven for SCADS, J.R. Fisher for peptides and assay development, E. Bare for providing the multi-domain Bcl-2 proteins, G. Grigoryan for code used to evaluate energies, the laboratory of R.T. Sauer for the use of their spectrofluorometer and G. Grigoryan, S. Chen, M. Radhakrishnan, B. Joughin, and C. Taylor for thoughtful comments and discussion. This work was funded by the NIH (GM67681 and P50-GM68762) and used equipment purchased under NSF grant number 0216437.

References

1. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82-87.
2. Calhoun JR, Kono H, Lahr S, Wang W, DeGrado WF, Saven JG. Computational design and characterization of a monomeric helical dinuclear metalloprotein. Journal of molecular biology 2003;334(5):1101-1115.
3. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ, DeGrado WF. Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. Journal of the American Chemical Society 2005;127(5):1346-1347.
4. Slovic AM, Kono H, Lear JD, Saven JG, DeGrado WF. Computational design of water-soluble analogues of the potassium channel KcsA. Proceedings of the National Academy of Sciences of the United States of America 2004;101(7):1828-1833.
5. Ogata K, Jaramillo A, Cohen W, Briand JP, Connan F, Choppin J, Muller S, Wodak SJ. Automatic sequence design of major histocompatibility complex class I binding peptides impairing CD8+ T cell recognition. J Biol Chem 2003;278(2):1281-1290.
6. Green DF, Dennis AT, Fam PS, Tidor B, Jasanoff A. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. Biochemistry 2006;45(41):12547-12559.
7. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D. Computational redesign of protein-protein interaction specificity. Nature Structural & Molecular Biology 2004;11(4):371-379.
8. Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, Schwab MS, Serrano L, Gonzalez C. Computer-aided design of a PDZ domain to recognize new target sequences. Nature Structural Biology 2002;9(8):621-627.
9. Shifman JM, Mayo SL. Modulating calmodulin binding specificity through computational protein design. Journal of molecular biology 2002;323(3):417-423.
10. Shifman JM, Mayo SL. Exploring the origins of binding specificity through the computational redesign of calmodulin. Proceedings of the National Academy of Sciences of the United States of America 2003;100(23):13274-13279.
11. Chevalier BS, Kortemme T, Chadsey MS, Baker D, Monnat RJ, Stoddard BL. Design, activity, and structure of a highly specific artificial endonuclease. Molecular Cell 2002;10(4):895-905.
12. Dwyer MA, Looger LL, Hellinga HW. Computational design of a biologically active enzyme. Science 2004;304(5679):1967-1971.
13. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature 2003;423(6936):185-190.
14. Bolon DN, Mayo SL. Enzyme-like proteins by computational design. Proceedings of the National Academy of Sciences of the United States of America 2001;98(25):14274-14279.
15. Su A, Mayo SL. Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Science 1997;6(8):1701-1707.
16. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302(5649):1364-1368.

17. Ali MH, Taylor CM, Grigoryan G, Allen KN, Imperiali B, Keating AE. Design of a heterospecific, tetrameric, 21-residue miniprotein with mixed alpha/beta structure. Structure 2005;13(2):225-234.
18. Havranek JJ, Harbury PB. Automated design of specificity in molecular recognition. Nature Structural Biology 2003;10(1):45-52.
19. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.
20. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. Protein Science 2000;9:1106-1119.
21. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332(2):449-460.
22. Gross A, McDonnell JM, Korsmeyer SJ. BCL-2 family members and the mitochondria in apoptosis. Genes & Development 1999;13(15):1899-1911.
23. Borner C. The Bcl-2 protein family: sensors and checkpoints for life-or-death decisions. Molecular Immunology 2003;39(11):615-647.
24. Sattler M, Liang H, Nettesheim D, Meadows RP, Harlan JE, Eberstadt M, Yoon HS, Shuker SB, Chang BS, Minn AJ, Thompson CB, Fesik SW. Structure of Bcl-x(L)-Bak peptide complex: Recognition between regulators of apoptosis. Science 1997;275(5302):983-986.
25. Petros AM, Nettesheim DG, Wang Y, Olejniczak ET, Meadows RP, Mack J, Swift K, Matayoshi ED, Zhang HC, Thompson CB, Fesik SW. Rationale for Bcl-x(L)/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. Protein Science 2000;9(12):2528-2534.
26. Liu XQ, Dai SD, Zhu YN, Marrack P, Kappler JW. The structure of a Bcl-x(L)/Bim fragment complex: implications for bim function. Immunity 2003;19(3):341-352.
27. Denisov AY, Chen G, Sprules T, Moldoveanu T, Beauparlant P, Gehring K. Structural Model of the BCL-w-BID Peptide Complex and Its Interactions with Phospholipid Micelles. Biochemistry 2006;45(7):2250-2256.
28. Day CL, Chen L, Richardson SJ, Harrison PJ, Huang DCS, Hinds MG. Solution structure of prosurvival Mcl-1 and characterization of its binding by proapoptotic BH3-only ligands. J Biol Chem 2005;280(6):4738-4744.
29. Kuwana T, Bouchier-Hayes L, Chipuk JE, Bonzon C, Sullivan BA, Green DR, Newmeyer DD. BH3 domains of BH3-only proteins differentially regulate bax-mediated mitochondrial membrane permeabilization both directly and indirectly. Molecular Cell 2005;17(4):525-535.
30. Chen L, Willis SN, Wei A, Smith BJ, Fletcher JI, Hinds MG, Colman PM, Day CL, Adams JM, Huang DCS. Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. Mol Cell 2005;17(3):393-403.
31. Certo M, Moore VD, Nishino M, Wei G, Korsmeyer S, Armstrong SA, Letai A. Mitochondria primed by death signals determine cellular addiction to antiapoptotic BCL-2 family members. Cancer Cell 2006;9(5):351-365.
32. Letai A, Bassik MC, Walensky LD, Sorcinelli MD, Weiler S, Korsmeyer SJ. Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. Cancer Cell 2002;2(3):183-192.

33. Chin JW, Schepartz A. Design and evolution of a miniature bcl-2 binding protein. Angewandte Chemie-International Edition 2001;40(20):3806-3809.
34. Gemperli AC, Rutledge SE, Maranda A, Schepartz A. Paralog-selective ligands for Bcl-2 proteins. Journal of the American Chemical Society 2005;127(6):1596-1597.
35. Sadowsky JD, Schmitt MA, Lee HS, Umezawa N, Wang SM, Tomita Y, Gellman SH. Chimeric (alpha/beta plus alpha)-peptide ligands for the BH3-recognition cleft of Bcl-x(L): Critical role of the molecular scaffold in protein surface recognition. Journal of the American Chemical Society 2005;127(34):11966-11968.
36. Degterev A, Lugovskoy A, Cardone M, Mulley B, Wagner G, Mitchison T, Yuan JY. Identification of small-molecule inhibitors of interaction between the BH3 domain and Bcl-x(L). Nature Cell Biology 2001;3(2):173-182.
37. Oltersdorf T, Elmore SW, Shoemaker AR, Armstrong RC, Augeri DJ, Belli BA, Bruncko M, Deckwerth TL, Dinges J, Hajduk PJ, Joseph MK, Kitada S, Korsmeyer SJ, Kunzer AR, Letai A, Li C, Mitten MJ, Nettesheim DG, Ng S, Nimmer PM, O'Connor JM, Oleksijew A, Petros AM, Reed JC, Shen W, Tahir SK, Thompson CB, Tomaselli KJ, Wang BL, Wendt MD, Zhang HC, Fesik SW, Rosenberg SH. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. Nature 2005;435(7042):677-681.
38. Oberstein A, Jeffrey PD, Shi Y. Crystal Structure of the Bcl-XL-Beclin 1 Peptide Complex: BECLIN 1 IS A NOVEL BH3-ONLY PROTEIN. J Biol Chem 2007;282(17):13123-13132.
39. Czabotar PE, Lee EF, van Delft MF, Day CL, Smith BJ, Huang DC, Fairlie WD, Hinds MG, Colman PM. Structural insights into the degradation of Mcl-1 induced by BH3 domains. Proc Natl Acad Sci U S A 2007;104(15):6217-6222.
40. Butterfoss GL, Kuhlman B. Computer-based design of novel protein structures. Annual Review of Biophysics and Biomolecular Structure 2006;35:49-65.
41. Harbury PB, Tidor B, Kim PS. Repacking Protein Cores with Backbone Freedom - Structure Prediction for Coiled Coils. Proceedings of the National Academy of Sciences of the United States of America 1995;92(18):8408-8412.
42. North B, Summa CM, Ghirlanda G, DeGrado WF. D(n)-symmetrical tertiary templates for the design of tubular proteins. Journal of molecular biology 2001;311(5):1081-1090.
43. Saunders CT, Baker D. Recapitulation of protein family divergence using flexible backbone protein design. Journal of molecular biology 2005;346(2):631-644.
44. Kono H, Saven JG. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. Journal of molecular biology 2001;306(3):607-628.
45. Larson SM, England JL, Desjarlais JR, Pande VS. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. Protein Science 2002;11(12):2804-2813.
46. Emberly EG, Mukhopadhyay R, Wingreen NS, Tang C. Flexibility of alpha-helices: Results of a statistical analysis of database protein structures. Journal of molecular biology 2003;327(1):229-237.
47. Emberly EG, Mukhopadhyay R, Tang C, Wingreen NS. Flexibility of beta-sheets: Principal component analysis of database protein structures. Proteins-Structure Function and Bioinformatics 2004;55(1):91-98.
48. Ma J. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. Structure 2005;13(3):373.

49. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. Biophysical Journal 2005;88(2):1291.
50. Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Engineering 2001;14(1):1-6.
51. Shenkin P, McDonald D. Cluster - Analysis of Molecular Conformations. J Comput Chem 1994;15:899-916.
52. The MathWorks I. MatLab. 7.1: The MathWorks, Inc.; 1984-2005.
53. Fu X, Kono H, North B, Nanda V, Lee O-S, DeGrado WF, Saven JG. Designing proteins by exploring the energy landscape. in preparation.
54. O'Neil KT, DeGrado WF. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 1990;250:646--651.
55. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. Trends in Biochemical Sciences 1998;23(10):403-405.
56. Zhang HC, Nimmer P, Rosenberg SH, Ng SC, Joseph M. Development of a high-throughput fluorescence polarization assay for Bcl-x(L). Analytical Biochemistry 2002;307(1):70-75.
57. Lee KH, Han WD, Kim KJ, Oh BH. The Crystal Structure of Bcl-xL in Complex with Full-length Bad. To be published.
58. Bruncko M, Oost TK, Belli BA, Ding H, Joseph MK, Kunzer A, Martineau D, McClellan WJ, Mitten M, Ng SC, Nimmer PM, Oltersdorf T, Park CM, Petros AM, Shoemaker AR, Song X, Wang X, Wendt MD, Zhang H, Fesik SW, Rosenberg SH, Elmore SW. Studies Leading to Potent, Dual Inhibitors of Bcl-2 and Bcl-xL. J Med Chem 2007;50:641-662.
59. Qian B, Ortiz AR, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proceedings of the National Academy of Sciences of the United States of America 2004;101(43):15346-15351.
60. Herman MD, Lehtio L, Arrowsmith CH, Berglund H, Busam RD, Collins R, Dahlgren LG, Edwards AM, Flodin S, Flores A, Graslund S, Hammarstrom M, Johansson I, Kallas A, Karlberg T, Kotenyova T, Moche M, Nilsson ME, Nyman T, Persson C, Sagemark J, Svensson L, Thorsell AG, Tresaugues L, Van Den Berg S, Weigelt J, Welin M, Nordlund P, Consortium SG. Human Bcl-2A1 in Complex with Bim To be Published 2008.
61. Smits C, Czabotar PE, Hinds MG, Day CL. Structural Plasticity Underpins Promiscuous Binding of the Pro-Survival Protein A1. To be Published 2008.
62. Grigoryan G, Keating AE. Structure-based prediction of bZIP partnering specificity. Journal of molecular biology 2006;355(5):1125-1142.
63. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of molecular biology 2002;320(2):369-387.
64. Lee EF, Czabotar PE, van Delft MF, Michalak EM, Boyle MJ, Willis SN, Puthalakath H, Bouillet P, Colman PM, Huang DC, Fairlie WD. A novel BH3 ligand that selectively targets Mcl-1 reveals that apoptosis can proceed without Mcl-1 degradation. J Cell Biol 2008;180(2):341-355.
65. Lee EF, Czabotar PE, Smith BJ, Deshayes K, Zobel K, Colman PM, Fairlie WD. Crystal structure of ABT-737 complexed with Bcl-xL: implications for selectivity of antagonists of the Bcl-2 family. Cell Death Differ 2007;14(9):1711-1713.

66. Sattler M, Liang H, Nettesheim D, Meadows RP, Harlan JE, Eberstadt M, Yoon HS, Shuker SB, Chang BS, Minn AJ, Thompson CB, Fesik SW. Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis. Science 1997;275(5302):983-986.
67. Petros AM, Nettesheim DG, Wang Y, Olejniczak ET, Meadows RP, Mack J, Swift K, Matayoshi ED, Zhang H, Thompson CB, Fesik SW. Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. Protein Sci 2000;9(12):2528-2534.
68. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Advances in protein chemistry 1968;23:283-438.
69. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comp Chem 1983;4(2):187-217.
70. MacKerell ADJ, Brooks, B., Brooks, C.L. III, Nilsson, L., Roux, B., Won, Y., Karplus, M., . CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. Chichester: John Wiley & Sons:; 1998. 271-277 p.
71. Martin ACR. Profit. London; 2001.
72. Zou JM, Saven JG. Statistical theory of combinatorial libraries of folding proteins: Energetic discrimination of a target structure. Journal of molecular biology 2000;296(1):281-294.
73. Weiner SJ, Kollman PA, Nguyen DT, Case DA. An all atom force field for simulations of proteins and nucleic acids. J Comput Chem 1986;7:230-252.
74. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins: Struct Funct Genet 2000;40:389-408.
75. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35(2):133-152.
76. Desmet J, Demaeyer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 1992;356:539-542.
77. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophysical Journal 1994;66(5):1335.
78. Pierce NA, Spriet JA, Desmet J, Mayo SL. Conformational splitting: A more powerful criterion for dead-end elimination. J Comput Chem 2000;21:999-1009.
79. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins 1998;33(2):227-239.
80. Lasters I, De Maeyer M, Desmet J. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. Protein Engineering 1995;8(8):815-822.
81. Gordon DB, Mayo SL. Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure 1999;7(9):1089-1098.
82. Rocchia W, Alexov E, Honig B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. JOURNAL OF PHYSICAL CHEMISTRY B 2001;105(28):6507-6514.
83. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. J Comput Chem 2002;23(1):128-137.

84. Hubbard S, Thornton JM. NACCESS, Computer Program. 2.1.1; 1996.
85. Munoz V, Serrano L. Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. Journal of molecular biology 1995;245(3):275-296.
86. Pace CN, Scholtz JM. A helix propensity scale based on experimental studies of peptides and proteins. Biophy J 1998;75:422--427.
87. Panchenko AR. Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Research 2003;31(2):683-689.
88. Hoover DM, Lubkowski J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res 2002;30(10):e43.
89. Edelhoch H. Sepctroscopic determination of tryptophan and tyrosine in proteins. Biochemistry 1967;6:1948-1954.

Chapter 4

Predicting helix orientation for coiled-coil dimers

Portions reprinted with permission of John Wiley & Sons, Inc. from:

Apgar, J.R., Gutwin, K.N., Keating. A.E. (2008) "Predicting helix orientation for coiled-coil dimers" *Proteins, In Press*

Collaborators Notes:

Karl Gutwin generated the coiled-coil test sets and implemented the sequence based scoring functions.

Abstract

The alpha-helical coiled coil is a structurally simple protein oligomerization or interaction motif consisting of two or more alpha helices twisted into a supercoiled bundle. Coiled coils can differ in their stoichiometry, helix orientation and axial alignment. Because of the near degeneracy of many of these variants, coiled coils pose a challenge to fold recognition methods for structure prediction. Whereas distinctions between some protein folds can be discriminated on the basis of hydrophobic/polar patterning or secondary structure propensities, the sequence differences that encode important details of coiled-coil structure can be subtle. This is emblematic of a larger problem in the field of protein structure and interaction prediction: that of establishing specificity between closely similar structures. We tested the behavior of different computational models on the problem of recognizing the correct orientation – parallel vs. antiparallel – of pairs of alpha helices that can form a dimeric coiled coil. For each of 131 examples of known structure, we constructed a large number of both parallel and antiparallel structural models and used these to assess the ability of five energy functions to recognize the correct fold. We also developed and tested three sequence-based approaches that make use of varying degrees of implicit structural information. The best structural methods performed similarly to the best sequence methods, correctly categorizing ~81% of dimers. Steric compatibility with the fold was important for some coiled coils we investigated. For many examples, the correct orientation was determined by smaller energy differences between parallel and antiparallel structures distributed over many residues and energy components. Prediction methods that used structure but incorporated varying approximations and assumptions showed quite different behaviors when used to investigate energetic contributions to orientation

preference. Sequence based methods were sensitive to the choice of residue-pair interactions scored.

Introduction

The alpha-helical coiled coil has long served as a model for studying the relationship between protein sequence and structure. The coiled coil consists of a bundle of supercoiled helices encoded by a 7-residue sequence repeat of the form $[\mathbf{abcdefg}]_n$. With **a** and **d** positions hydrophobic and **e** and **g** positions usually polar or charged, a "sticky" stripe winds its way around an individual helix, dictating the formation of a twisted helical bundle (Figure 2-8a and b). Because of this simple relationship, the coiled-coil fold is one of the easiest protein structures to predict. Numerous programs have been developed to detect the presence of coiled-coil forming segments in sequences, and these exhibit respectable sensitivity and specificity.[1-4] However, few methods exist to predict the variety of topologies found in coiled-coil structures.[4-6] Helix content can vary from 2 to 7 helices, and helix orientation can be parallel or antiparallel. Structures can be homo- or hetero-oligomeric, and the helices can align axially in different ways. Thus, the "coiled coil" is really a large family of structures that share many properties but exhibit different topological characteristics.[7]

The difficulty of predicting coiled-coil structure lies in differentiating what can be subtle distinctions in interactions. For example, it has been reported for several designed coiled coils that changing a single **a**- or **d**-position residue can lead to a change or loss of oligomerization specificity.[8-10] Small changes in sequence can also alter helix orientation preferences. In the work of Oakley et al., moving a buried Asn residue by 7 positions in one helix and 3 in its partner helix was sufficient to switch a designed coiled coil from a parallel to an antiparallel

orientation.[11] Lumb and Kim found that a buried Asn can establish both oligomerization and helix orientation specificity.[12] Perhaps surprisingly, this sensitivity to small sequence changes appears to hold for many native sequences as well. Mutation of an Asn residue at an **a** position of the yeast transcription factor GCN4 leads to loss of oligomerization specificity in that coiled coil,[13] and changing 2 residues in the antiparallel coiled-coil dimer of Bcr can give either a mixture of antiparallel higher-order helical assemblies or trimers, depending on the mutations.[14] This plasticity of coiled-coil structure in response to mutation makes the problem of fold recognition challenging. Much of the signal that is typically used to discriminate one structure from another in prediction, including patterns of predicted secondary structure and preferences of residues for different degrees of burial, is of little or no use in classifying coiled coils by type because these properties are largely the same in many of the competing structures. This situation also arises in other structure-prediction problems, where target and decoy structures must be resolved that sometimes include "mirror-image" variants containing the correct secondary structure elements arranged incorrectly with a reversed overall chirality.[15,16]

Despite these challenges, some progress has been made on the problem of predicting coiled-coil interaction preferences from sequence. Several methods have been proposed for discriminating dimers from trimers. Simulations have successfully captured oligomeric preferences, and sequence-based programs have been developed for making predictions on novel coiled coils.[4,5,17] However, these were developed over a decade ago, using extremely small sets of known coiled-coil examples, and frequently fail on additional test cases that are available today. More recently, several methods have been developed to predict interacting partners among the bZIP transcription factors – an important protein family in which dimerization is mediated by a parallel coiled coil.[6,18-21] Relatively little is known about determinants of coiled-coil helix

orientation, however. Various strategies have been used to design coiled coils that specifically adopt a parallel or antiparallel orientation, such as electrostatic charge patterning or the manipulation of **a**- and **d**-position polar residues or shape complementarity.[11,22-27] Alanine in core positions has been proposed to contribute to antiparallel specificity in coiled coils.[28] But in general, it is difficult to recognize sequence patterns that may specify helix orientation in native sequences. Analyzing features that determine orientation specificity via mutagenesis is often confounded by the fact that key residues may encode other types of specificity as well. For example, when probing the possible role of **d**-position Glu in determining the orientation preference of the Bcr coiled-coil domain, mutation to Leu led to the formation of trimers and other higher-order oligomers, as mentioned above.[14]

In this Chapter, we describe the performance of several types of computational models on the problem of predicting coiled-coil orientation. Due to the relatively small number of coiled coils with known orientation preference, learning strategies such as those that have been used in other motif recognition problems are not readily applicable.[1,29-32] Instead, we relied on structural models to evaluate coiled-coil orientation. We developed both explicit structural models and sequence-based models in which our use of structure was implicit. "Out-of-the-box" methods of both types did not perform very well, but small adjustments that took advantage of coiled-coil properties significantly improved the results.

Results

We tested several methods for predicting whether two sequences that can form a coiled coil will assemble as a parallel or an antiparallel dimer. For simplicity, we considered pairs of sequences of equal length that can be fully overlapped in both parallel and antiparallel

orientations, i.e., those sequences that are "blunt ended" when aligned both ways. This test is akin to biochemical assays that can measure the relative stability of these two conformations,[11,33] although it avoids complexities that can be introduced by non-dimer states. An important feature of our calculations is that they do not require an accurate treatment of a dissociated and/or unfolded reference state (because the common unfolded state cancels), and therefore represent a best-case scenario for computational prediction.[18] Significant additional challenges, such as predicting the correct axial alignment of helices, and determining that two sequences will form a dimer rather than some other type of oligomer, must be overcome to develop a general coiled-coil structure prediction method.

Our assessment of different methods used a database of parallel and antiparallel coiled-coil dimers of known structure. To assemble this database, dimers were identified using the program SOCKET,[34] which detects the knobs-into-holes side-chain packing that characterizes coiled-coil interfaces. Additionally, SOCKET was used to determine the coiled-coil heptad assignment (**abcdefg**). Because SOCKET also detects knobs-into-holes packing in non-coiled-coil structures, such as 4-helix bundles and helical sheets,[34,35] these were manually removed. We also included several sequences from the human bZIP family of coiled coils[21,36] in order to increase the number of parallel heterodimers in the database. In total, 61 parallel and 70 antiparallel examples with low sequence similarity and length >= 18 residues were selected and defined as our test set. We made the assumption that the coiled-coil motif itself is sufficient to encode the observed helix orientation for these structures. This may not always be true, and it is less likely to be true for short sequences that are part of a more complex fold. It is also less likely to be true for coiled coils that are highly buried. Nevertheless, local determination of helix orientation has been confirmed experimentally for a small number of cases in the literature, and

it is likely to be true for the majority of our examples.[14,37-40] Due to the limited number of available structures, there are biases in the data set. In particular, the parallel structures include more homodimers and the antiparallel structures more heterodimers. This affected the performance of some methods, as discussed below. A summary of the structures that make up the database is provided in Table 4-1 and a detailed list is available in Table 4-2.

We tested two general categories of methods. The first required explicit models of structure for each orientation. The experimentally determined structure was available for the correct orientation for most of the sequences, but to simulate a real prediction problem we did not use this structure in our evaluations. Instead, models of both parallel and antiparallel complexes were predicted for each dimer. To generate idealized parallel backbones, we used a parameterization first developed by Crick in 1953 and subsequently adapted for use with modern molecular modeling programs by Harbury et al.[41,42] To describe antiparallel coiled-coil backbones, we introduced two new parameters into the Crick parameterization (see Methods). We then generated 120 ideal parallel and 81 ideal antiparallel backbones that spanned the parameter space of the dimeric coiled-coil test set (Figures 2-9 and 2-11 to 2-14). The backbone RMSD between each native structure and its closest idealized backbone was in the range of 0.25-1.8 Å, with all but 12 structures within 1.0 Å (Figure 4-1).

The other class of methods that we tested was based on sequence and did not require structural modeling. These approaches took advantage of characteristics of the coiled coil, such as the heptad repeat and extensive experimental characterization of interfacial residue-residue interactions that are important for dimer stability and specificity. We used this information to select interchain pairs of heptad positions that were scored based upon the residues at those

Table 4-1: Test set of coiled-coil dimers of known orientation

	Sequence Pairs	Avg. Length (range) in residues	Number of intramolecular coiled coils	Avg (range) fraction exposed SASA[a]	Avg (range) RMSD to closest ideal Crick backbone (Å)[b]
Parallel	61				
Homodimer	49	32.9 (18-75)	0	0.71 (0.24 - 0.95)	0.72 (0.29-2.5)
Heterodimer	12	32.9 (18-40)	0	0.80 (0.67 - 0.91)	0.57 (0.35-1.0)
Antiparallel	70				
Homodimer	19	25.0 (18-40)	0	0.59 (0.33 - 0.89)	0.51 (0.21-1.0)
Heterodimer	51	22.5 (18-53)	45	0.58 (0.16 - 0.83)	0.60 (0.28-2.4)

Data for seven bZIP coiled coils without structures not included in averages. [a]Fraction exposed is the ratio of the solvent-accessible surface area (SASA) of the coiled coil as observed in the crystal structure to the SASA of the isolated coiled coil. SASA calculated using NACCESS.[68] [b]RMSD to the closest ideal Crick backbone is the difference between the crystal structure and the best-fitting Crick ideal structure. Data for all structures is shown in Figure 2-12.

Table 4-2: List of PQS structures in the test set.

PQS ID	strand1_loc	strand1_hep	strand2_loc	strand2_hep
1a36	644-668:A	a	684-708:A	a
1a38	48-66:A	d	79-97:A	d
1am9_1	366-391:A	d	366-391:B	d
1ber	117-134:A	a	117-134:B	a
1bjt	1013-1030:A	a	1129-1146:A	a
1c1g_2	740-800:C	d	1024-1084:D	d
1c1g_2	600-674:C	d	884-958:D	d
1cii	229-281:A	a	387-439:A	a
1cnt_2	21-38:2	a	159-176:2	a
1cz7_2	300-345:C	a	300-345:D	a
1dgc	250-274:A	a	250-274:C	a
1e7t	358-404:A	d	358-404:B	d
1ecm	7-38:A	a	7-38:B	a
1ecr	10-27:A	a	110-127:A	a

1egw_1	21-38:A	a	21-38:B	a
1exj	77-116:A	d	77-116:B	d
1few	34-65:A	a	74-105:A	a
1few	101-118:A	a	101-118:B	a
1fos_1	158-190:E	d	282-314:F	d
1fos_2	158-197:G	d	282-321:H	d
1fs0	32-56:G	a	216-240:G	a
1fxk	21-45:A	a	71-95:A	a
1fxk	7-45:C	a	94-132:C	a
1gd2_2	101-132:G	a	101-132:H	a
1gmj_2	60-77:C	a	53-70:D	a
1go4	501-526:E	d	501-526:F	d
1go4	494-526:G	d	494-526:H	d
1h88	299-331:A	d	299-331:B	d
1hlo	57-75:A	d	57-75:B	d
1hw5	117-134:A	a	117-134:B	a
1i1i	165-182:P	a	253-270:P	a
1i1r	13-30:B	a	151-168:B	a
1ii8	166-191:A	d	712-737:B	d
1ik9	123-169:A	d	123-169:B	d
1io1	64-95:A	a	408-439:A	a
1ivs_2	805-823:B	d	840-858:B	d
1jbg	84-101:A	a	84-101:B	a
1jnm	273-305:A	d	273-305:B	d
1k1f_2	42-66:E	a	28-52:F	a
1kd8_3	2-33:E	a	2-33:F	a
1kql	232-270:A	a	232-270:B	a
1kvk	264-282:A	d	290-308:A	d
1lih	58-75:A	a	58-75:B	a
1lj2	213-230:A	a	213-230:B	a
1lj2	274-299:A	d	274-299:B	d
1m5i	134-151:A	a	214-231:A	a
1nkn_1	849-916:A	d	849-916:B	d
1nkn_2	846-912:C	a	846-912:D	a
1nkp_1	953-971:A	d	253-271:B	d
1no4_2	36-67:D	a	36-67:C	a

1nt2	84-102:B	d	131-149:B	d
1nwq	310-335:A	d	310-335:C	d
1nyh	1285-1337:A	a	1285-1337:B	a
1o5l	109-126:A	a	109-126:B	a
1o9c	59-76:A	a	85-102:A	a
1omi	1110-1127:A	a	2110-2127:B	a
1orj	1019-1037:A	d	1103-1121:A	d
1ov9	22-39:A	a	22-39:B	a
1pl5	1274-1341:A	d	1274-1341:S	d
1q05	88-105:A	a	88-105:B	a
1q06	84-109:A	d	84-109:B	d
1q08	86-111:A	d	86-111:B	d
1qp9_1	107-124:A	a	107-124:B	a
1qsd	13-37:B	a	48-72:B	a
1qvr_1	399-417:A	d	435-453:A	d
1qz2	404-422:A	d	404-422:B	d
1r6f	153-171:A	d	284-302:A	d
1r6t	9-26:A	a	36-53:A	a
1r7j	66-90:A	a	66-90:B	a
1rq0_2	435-452:B	a	462-479:B	a
1s1c	992-1009:X	a	992-1009:Y	a
1s4b	164-181:P	a	252-269:P	a
1ses	30-47:B	a	80-97:B	a
1t3j	688-726:A	a	688-726:B	a
1t6f	2-33:A	a	2-33:B	a
1tjl_1	41-65:A	a	82-106:A	a
1tu3_2	807-832:H	d	807-832:I	d
1tu3_3	811-835:J	a	811-835:B	a
1twf	245-262:C	a	88-105:K	a
1uii	99-145:A	d	99-145:B	d
1uuj_2	58-75:C	a	58-75:D	a
1vp7_1	55-73:A	d	55-73:B	d
1wle	75-93:A	d	128-146:A	d
1wlq_1	96-135:A	d	96-135:B	d
1wu9	193-224:A	a	193-224:B	a
1x03	212-244:A	d	212-244:B	d

1x75	366-384:B	d	474-492:B	d
1xd4_2	205-223:B	d	249-267:B	d
1xnp	127-152:A	d	134-159:B	d
1ybz	4-35:A	a	4-35:B	a
1yf2	178-202:A	a	388-412:A	a
1yf2	178-195:B	a	395-412:B	a
1yhn	248-272:B	a	248-272:C	a
1yke_1	175-192:A	a	94-111:B	a
1z0j	736-753:B	a	762-779:B	a
1zik	2-26:A	a	2-26:B	a
1zil	2-26:A	a	2-26:B	a
1zke_1	5-22:A	a	55-72:A	a
1zme	76-93:C	a	76-93:D	a
1zpy	23-41:A	d	53-71:A	d
2b5u	333-372:C	d	393-432:C	d
2bde	340-358:A	d	380-398:A	d
2br9	53-70:A	a	79-96:A	a
2btp	48-66:A	d	79-97:A	d
2d4c_1	216-241:A	d	223-248:B	d
2d4x	70-94:A	a	233-257:A	a
2d8e	5-36:A	a	5-36:B	a
2e7s_4	52-69:G	a	52-69:H	a
2esh	90-115:A	d	90-115:B	d
2etn_3	19-36:C	a	48-65:C	a
2fxo_2	845-905:C	d	845-905:D	d
2fxo_2	912-957:C	a	912-957:D	a
2gau	124-148:A	a	124-148:B	a
2h7v_2	562-580:D	d	590-608:D	d
2hg4_1	16-33:A	a	16-33:B	a
2hl5	200-224:A	a	200-224:B	a
2hld_1	3-34:G	a	227-258:G	a
2iw5	421-445:A	a	335-359:B	a
2jdi	3-20:G	a	236-253:G	a
2ncd	314-345:A	a	314-345:B	a
2nov_1	361-379:A	d	435-453:A	d
2o98	907-924:P	a	907-924:C	a

2ocy	69-87:A	d	69-87:B	d
2pah	430-447:A	a	430-447:C	a
bbz2_C_EBPbeta+44_CEBPalpha*		a		a
bbz3_C_EBPgamma+35_ATF4*		d		d
bbz4_ATF_7+55_MAFK*		a		a
bbz5_ATF_2+10_FOS*		a		a
bbz6_CREBPA+28_JUN*		a		a
bbz7_ATF_1+52_CREM*		a		a
bbz7_ATF_1+7_ATF_1*		a		a

* Sequences taken from previously published bZIP interaction data.[36] Strand1_loc and Strand2_loc indicate the residue numbers and chain for the first and second helix of the coiled coil. Strand1_hep and Strand2_hep indicate the first heptad position of the first and second helices, respectively.

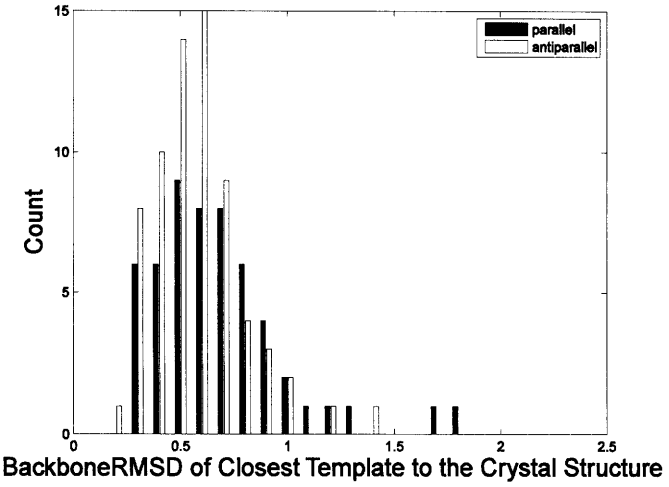


Figure 4-1: Distribution of the backbone RMSD (N, Cα, and C atoms) for the native crystal structures in the test set to the closest ideal structure in the backbone sets. For every example, an idealized model with an RMSD of less than 1.8 Å was available for selection as a template.

positions, thus using structural information implicitly. We refer to the two different types of approaches as ESMs and ISMs, for explicit or implicit structural models, respectively.

These two classes of models have different strengths and weaknesses. The ISMs are much faster to evaluate and can easily incorporate experimental data about relevant heptad pairs and interaction energies. However, they make strong assumptions about the independence of

pair-wise interactions and may obscure potentially significant details of atomic interactions necessary for modeling orientation specificity. ESMs provide advantages for analysis and interpretation of the physical basis of the overall interaction. Finally, ESMs are more generalizable in that they can potentially be applied equally to any structure; ISMs must be created specifically for the structure to be modeled.

Performance of explicit-structure models

Predicting helix orientation using ESMs involved three steps: (1) generating large numbers of parallel and antiparallel dimer backbones, (2) modeling each sequence pair on each backbone, and (3) selecting the lowest-energy model. The first step was carried out using the coiled-coil parameterizations described above. The second step was carried out using Rosetta, or a combination of Rosetta and CHARMM (see below). [43-45] The third step gave rise to differences between models, with each ESM named according to the energy function used at this stage.

In a preliminary set of calculations, we tested two structure-prediction methods for use in step 2. Initially, Rosetta was simply used to place side chains into preferred conformations on each of 81 parallel and 120 antiparallel idealized Crick backbones. When Rosetta was used to select the lowest-energy structure and orientation for each pair of sequences (corresponding to step 3), this procedure predicted the orientation of 42/61 parallel sequences and 48/70 antiparallel sequences correctly. In the second approach, all Rosetta-repacked backbones were relaxed via minimization using the CHARMM param19 force field.[45] Rosetta evaluation of these relaxed structures gave strikingly better results, improving the prediction rate to 50/61 (82%) of parallel sequences and 57/70 (81%) of antiparallel sequences. The performance of these models is shown in Figure 4-2a (left panel). Results are plotted as the fraction of antiparallel sequences

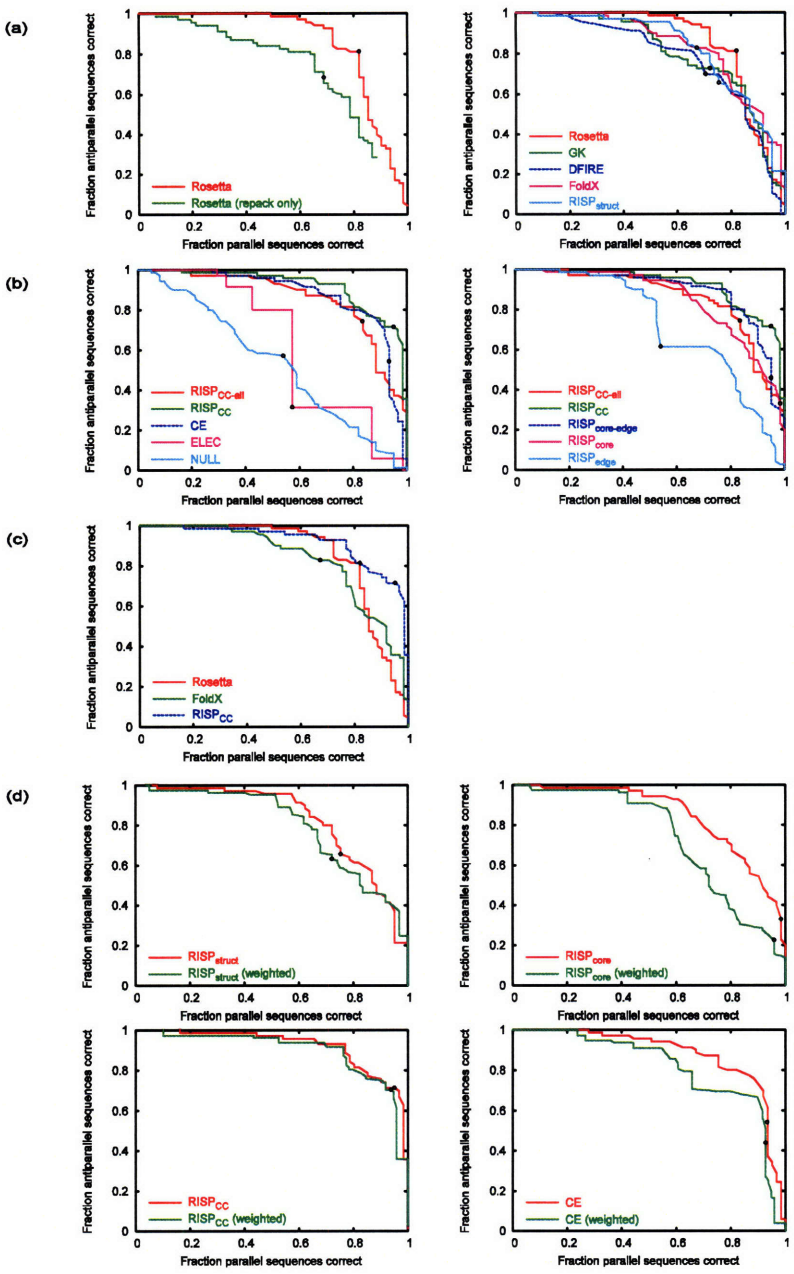


Figure 4-2: Parallel vs. antiparallel discrimination performance of different methods. The fraction of antiparallel structures correctly predicted is plotted versus the fraction of parallel structures correctly predicted. Curves were generated by varying $E_{cut} = E_{AP} - E_P$. A structure was predicted to have an antiparallel orientation if the energy of the antiparallel state was lower than that of the parallel state plus E_{cut}. If this energy was higher, the orientation was predicted as parallel. $E_{cut} = 0$ denoted by black dot. (a) Comparison of ESMs. At left, a comparison of Rosetta evaluated on structures without (repacked only) or with structural relaxation. At right, all candidate ESMs evaluated using relaxed structures. (b) Comparison of ISMs. At left, candidate ISMs including NULL control; at right, several variants of the RISP model. (c) Comparison of best ESM and ISM models. (d) Comparison of the performance on the test set (red) and the performance when hetero- and homodimer results are weighted equally (green). Clockwise from top left, the panels are for $RISP_{stuct}$, $RISP_{core}$, CE and $RISP_{CC}$.

predicted correctly vs. the fraction of parallel sequences predicted correctly. Because including minimization in step 2 significantly improved performance, this protocol was adopted in all remaining calculations, for all ESMs. Using this approach, the predicted structures for the correct orientation provided a good approximation of the real structures, with backbone RMSD values in the range 0.4-2.2 Å (all but 7 within 1.5 Å) and χ-angle recovery rates only slightly lower than can be achieved on the native backbone (Table 4-3).

Models GK, FoldX, DFIRE and RISP used different potentials to select the lowest-energy structures. Model GK, developed by Grigoryan et al.,[18] is based on the CHARMM param19 force field[45] and includes van der Waals interactions and a combination of EEF1 desolvation[46] and generalized Born screening of electrostatic interactions. This model previously showed good performance predicting coiled-coil binding partners.[18] GK describes similar physical terms to those captured by Rosetta, but it is more physical, with no statistical terms or empirical weighting. It performed slightly less well on orientation prediction than Rosetta. FoldX is a scoring function developed by Guerois et al.[47] It consists of physically descriptive terms weighted to predict experimental mutation free energies of primarily large-to-small mutations. Its performance was intermediate between that of Rosetta and GK (Figure 4-2a).

DFIRE and RISP are statistical potentials derived from the frequencies of interactions in the PDB.[48] They were applied to coiled-coil structures by scoring pairs of atoms or residues that met certain criteria. DFIRE is an atom-based potential that has been reported to predict protein-protein complex affinities accurately from experimental structures.[48] On our orientation-prediction test, it performed slightly worse than GK. RISP is a Residue-based Interfacial Statistical Potential consisting of 210 weights for scoring pairs of inter-chain residues that fall

Table 4-3: Chi angle recovery of repacked structures

	Repacked Crystal Structure	Lowest Energy Ideal Structure
Antiparallel Core Residues	0.69 ± 0.14	0.60 ± 0.14
Antiparallel Edge Residues	0.60 ± 0.18	0.55 ± 0.16
Antiparallel other residues	0.55 ± 0.15	0.52 ± 0.15
Parallel Core Residues	0.68 ± 0.12	0.61 ± 0.097
Parallel Edge Residues	0.56 ± 0.14	0.52 ± 0.13
Parallel other residues	0.56 ± 0.11	0.53 ± 0.11

Values shown are for average repacking performance. This is defined as the fraction of χ_1 and χ_2 angles recovered (within $\pm 40^0$), evaluated on the native crystal structure or the lowest-energy repacked structure as evaluated by Rosetta on relaxed structures. Chi angle recovery broken up by core (**a** or **d**), edge (**e**, **g**) and the remaining other residues (**b**, **c** and **f**).

within a distance cutoff; it is very similar to the residue-based potential developed by Lu et al.[49] Applied to the relaxed structure set as $RISP_{struct}$, it performed relatively poorly (Figure 4-2a).

To address test-set bias, we approximated the performance expected if there were equal proportions of homo- and heterodimers in the parallel and antiparallel test sets. This was done by calculating the average performance on homodimeric and heterodimeric examples, weighted equally, for each orientation class, at each E_{cut} value (E_{cut} is defined in Figure 4-2). Figure 4-2d shows that $RISP_{struct}$ was quite sensitive to this adjustment. This potential favored homodimers, and some of its success in predicting parallel structures was a result of this bias. The DFIRE, Rosetta, FoldX and GK potentials, on the other hand, performed similarly in the two tests.

Performance of implicit-structure models

In our ISM models, the energy of a structure is expressed as a sum of contributions from pair-wise residue interactions. The models differ from one another in the choice of pairs and/or the weights assigned to them. Our selection of residue pairs took advantage of the known heptad register of the test-set structures. Heptad assignments for coiled-coil sequences with unknown structures can be made using programs such as Paircoil.[1,3] We considered only interactions

among the **a**, **d**, **e,** and **g** residues that make up the coiled-coil dimer interface. A summary of the notation and residue pairs for all ISM models is shown in Table 4-4. To approximate the $RISP_{struct}$ method using an ISM, we scored seven pairs involving residues that commonly satisfy the $RISP_{struct}$ distance cutoff. These pairs were assigned their RISP weights, giving method $RISP_{CC\text{-}all}$. Like $RISP_{struct}$, $RISP_{CC\text{-}all}$ did not perform very well (Figure 4-2b). Interestingly, however, when we scored only 5 types of interactions for each coiled-coil orientation, giving model $RISP_{CC}$, the performance was much better and rivaled that of the best ESM methods (Figure 4-2c). The pairs in $RISP_{CC}$ include those that have been described many times as being important for coiled-coil associations (i.e. **a-a'**, **d-d'** and **g-e'** for parallel[50-53] and **a-d'**, **g-g'** and **e-e'** for antiparallel[54]) as well as core-to-edge terms (**g-a'** and **d-e'** for parallel and **a-e'**, **d-g'** for antiparallel) that have been investigated in some systems and that were previously predicted to be important.[18,55,56] Further reduction of the number of pairs, i.e. using only core **a-a'**, **d-d'** (parallel) or **a-d'** (antiparallel) pairs, giving model $RISP_{core}$, or only edge **g-e'** or **g-g'** (parallel) or **e-e'** (antiparallel) pairs, giving $RISP_{edge}$, degraded performance (Figure 4-2b).

Given the success of model $RISP_{CC}$, we tested model CE. This model includes the same heptad-position pairs, but draws weights, where possible, from experimentally reported interaction energies. These include weights for **a-a'** and **g-e'** interactions in the parallel orientation, taken from coupling energies measured in the Vinson laboratory. Weights for **a-d'** interactions in the antiparallel orientation were taken from measurements by Hadley et al.[57] This model also did well, despite the limited number of available measurements (Figure 4-2b). The performance of two control models is also shown in Figure 4-2b. Model ELEC scores only the **e-** and **g**-position electrostatic complementarity and did not provide good parallel vs. antiparallel

Table 4-4: Summary of pair terms used in ISM models

Model	Parallel	Antiparallel
ELEC	**g-e'**	**g-g' e-e'**
CE	**a-a' g-e'**	**a-d' g-g' e-e'**
$RISP_{core}$	**a-a' d-d'**	**a-d'**
$RISP_{edge}$	**g-e'**	**g-g' e-e'**
$RISP_{core,edge}$	**a-a' d-d' g-e'**	**a-d' g-g' e-e'**
$RISP_{CC}$	**a-a' d-d' g-e' g-a' d-e'**	**a-d' g-g' e-e' a-e' d-g'**
$RISP_{CC-all}$	**a-a' d-d' g-e' g-a' d-e' a-d' d-a'**	**a-d' g-g' e-e' a-e' d-g' d-d' a-a'**

A prime (') designates a residue on the opposite helix. All interaction pairs listed involve structurally adjacent sites on opposite helices. For edge interactions where there may be some ambiguity as to what pair is indicated, the interactions are as follows: **g-e'** pairs in parallel coiled coils are between a **g** residue and the **e** residue of the next (more C-terminal) heptad of the opposite helix; in antiparallel coiled coils **g-g'** pairs are between a **g** residue and the **g** residue of the previous (more N-terminal) heptad of the opposite helix and **e-e'** pairs are between an **e** residue and the **e** residue of the next (more C-terminal) heptad of the opposite helix.

discrimination. We also illustrate the performance of a null model in which weights were assigned to the restricted set of pairs randomly.

Of the ISM models, $RISP_{core}$ and CE showed significant amounts of homodimer bias, i.e. their performance was worse when we weighted the homo- and heterodimer results equally (Figure 4-2d). For $RISP_{core}$, this effect came from more favorable weights for **a-a'** and **d-d'** homotypic interactions than heterotypic interactions. This bias was somewhat surprising, as the RISP energy function was designed to minimize such effects by excluding cases where a residue interacts with a symmetry-related copy of itself in the training set. Increasing the number of pair terms to make the $RISP_{CC}$ model, e.g. by adding edge and core-edge interactions that occur between positions not related by symmetry, diluted this effect, and the overall bias decreased (Figure 4-2d). The CE model is based on a much smaller number of terms than the RISP models, and so homodimer bias here is likely a result of the unequal numbers of weights available for scoring homo vs. heterodimers.

Analysis

The performance of all methods on all examples indicates that some structures are easier to predict than others. For 23 dimers (18%), all 8 methods predicted the correct orientation, and for 74 dimers (56%), at least 6 out of 8 methods were correct. Seventeen structures (13%) were predicted correctly by three or fewer methods. Some of the examples that are rarely predicted correctly may contradict our assumption that the PDB reflects the structure that coiled-coil fragments would adopt in isolation. For example, 1OV9, VicH H-NS histone-like protein, consists of an antiparallel coiled coil flanked by N-terminal swap domains that pack against it; any influence on helix orientation from these domains was not considered in our models. Another example is 1X75, DNA gyrase subunit A, in which an intramolecular antiparallel coiled coil is packed against a large structured loop. Again, structural elements that we did not model may contribute to the observed orientation.

The various prediction methods work very differently, as is evident when comparing their performance on subsets of the test complexes. Figure 4-3a clusters both methods and examples by the similarity of predicted orientation preferences. Classifying all methods as statistics-based (DFIRE, $RISP_{struct}$ and $RISP_{CC}$), knowledge-based (ELEC, CE) or pseudo-physical (Rosetta, GK, FoldX) shows that the knowledge-based potentials are least similar to the other methods and also not closely related to one another. The simple ELEC model had poor performance overall (Figure 4-2b). Figure 4-3a shows that much of this poor performance resulted from the model's frequent failure to make a prediction (gray boxes), due to equivalent attractive and repulsive charge-charge interactions in both orientations. There are also examples where ELEC made a strong, yet incorrect, prediction. Model CE performed much better than ELEC; in overall

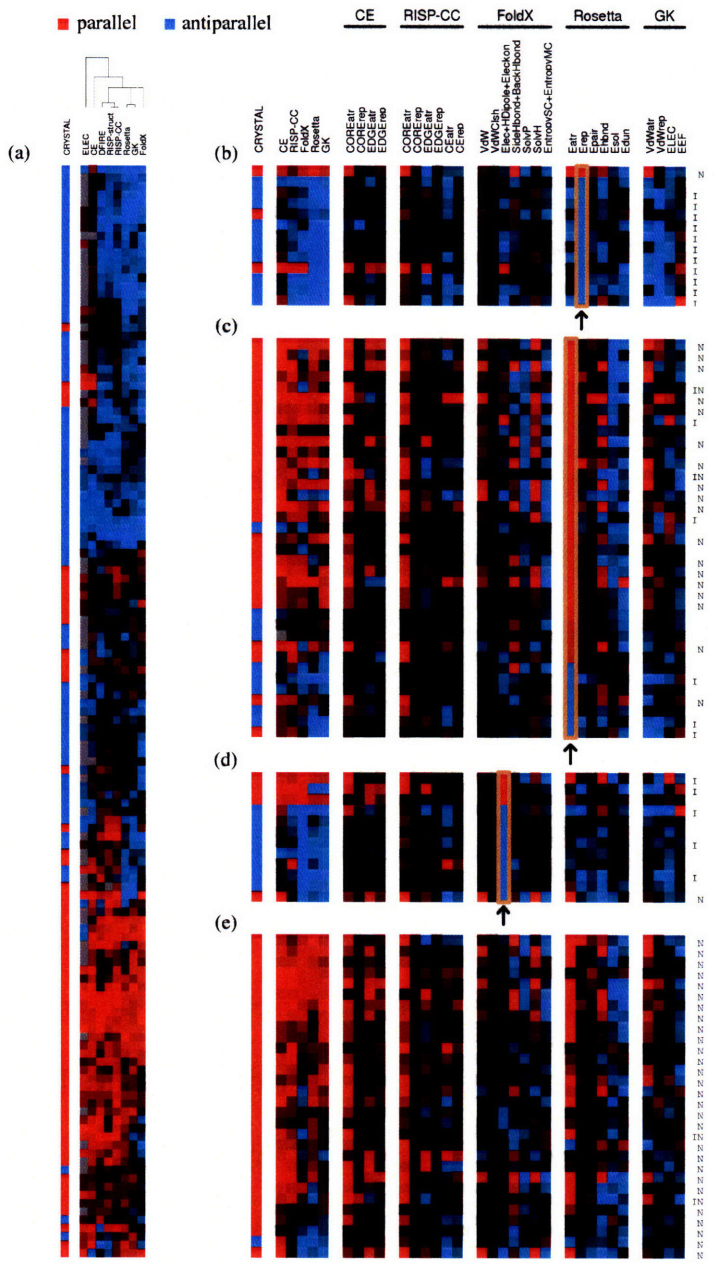


Figure 4-3: Overview of prediction performance and component analysis. All predictions were made using $E_{cut} = 0$. (a) Predictions clustered by method and example. Color (red: parallel, blue: antiparallel) denotes orientation prediction, and intensity (bright to dark) corresponds to the score of that prediction (ΔE), binned into deciles, where darker color indicates low rank (ΔE close to zero). CRYSTAL column denotes orientation in the x-ray structure. (b-e) Prediction results for subsets of sequences, re-clustered. Color scheme as in (a). CRYSTAL column denotes known orientation. Remaining columns are energy terms contributing to overall orientation predictions for the best ESM and ISM methods. Terms favoring parallel orientation are red; those favoring antiparallel are blue. Intensity is in units of sigma (standard deviation of all energy components on all test sequences for a given prediction method), capped at 2.5 σ. In (b-e), energy terms are shown for examples with: (b) the largest absolute magnitude Rosetta Erep, (c) the largest absolute magnitude Rosetta Eatr, (d) the largest FoldX electrostatic components, and (e) paired **a-a'** Asn residues in the parallel orientation. **N** indicates that the sequence pair contains Asn at one or more **a-a'** positions in the parallel orientation; **I** indicates that the sequence pair contains an Ile pair at **d-d'** in the parallel orientation. FoldX, Rosetta, and GK energy components are described further in the Methods and in Table 4-5.

prediction rate it was similar to the very good $RISP_{CC}$ (also an ISM). Yet, the clustering in Figure 4-3a shows that CE is not at all similar to the other ISMs in terms of how orientation is assigned for specific sequences. This is understandable, as CE and RISP are based on completely different methods of deriving pair-wise scoring weights (experiments vs. PDB frequency analysis). Comparisons of ELEC, CE, and $RISP_{CC}$ further illustrate how three types of terms (edge interactions involving **e** and **g** positions, core interactions involving **a** and **d** positions, and core-to-edge interactions) are all important (Figure 4-4a). The inclusion of these heptad-position pairs in $RISP_{CC}$ (absent from ELEC or CE) help to account for its better performance. Finally, it is interesting that the $RISP_{struct}$ and $RISP_{CC}$ methods cluster quite tightly, despite significant differences in their prediction performances, underscoring their basis in the same contact potential.

Differences among the structure-based methods can be dissected using component analysis, which potentially offers insights into physical determinants of helix orientation. For 5 methods (the ISMs CE and $RISP_{CC}$ and the ESMs FoldX, Rosetta and GK), we broke the predicted energy differences into their component terms for all of the examples in the test set. Figures 3b-e show subsets of these (all examples are included in Figure 4-4b). For the ESMs, we also examined the predictive power of individual components, as well as the co-variation of individual energy-term differences with the total parallel vs. antiparallel energy difference. These data are summarized in Figure 4-5 (descriptions of components are included in Table 4-5).

Figure 4-5 panels a-c illustrate the contributions of different energy terms to prediction performance. The prediction accuracy of each important term when used alone is shown, along with the effect of removing terms individually from the total energy. The Rosetta terms Eatr and Erep, which together give the total van der Waals energy, gave reasonable prediction

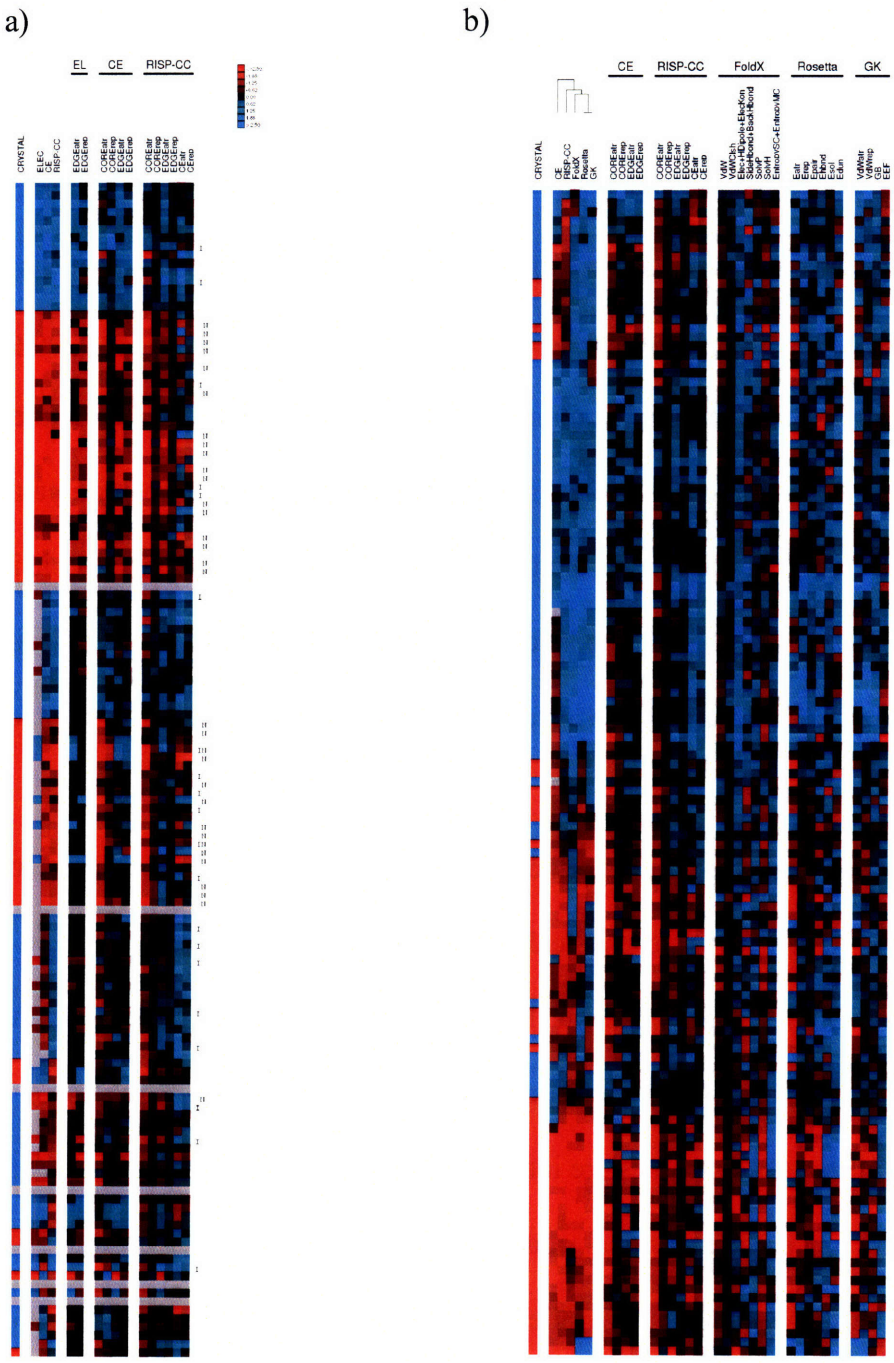


Figure 4-4: Component analysis of ESMs and ISMs. Methods and components are as in Figure 4-3. (a) ISM methods. Sequences are grouped (from top to bottom) as follows: correct with all three methods, incorrect with ELEC only, incorrect with ELEC and CE only, incorrect with all three methods. Remaining groups contain sequences not in the other groups. This figure illustrates the value of including all pairs in $RISP_{CC}$. (b) ESM methods, with color scheme and magnitudes as in Figure 4-3.

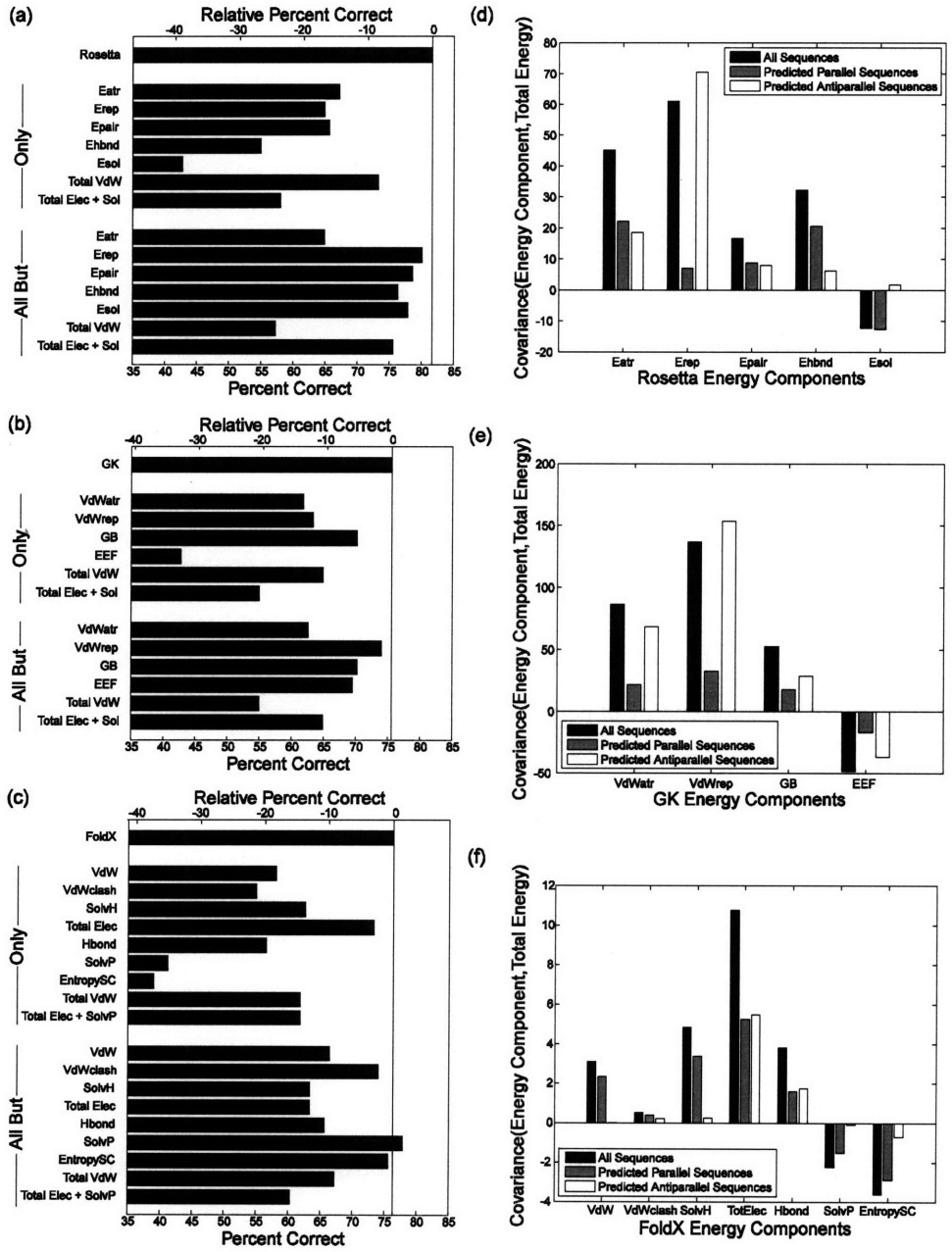


Figure 4-5: Energy component contributions to performance. (a-c) The performance of each component or sum of components was considered alone (Only) or was excluded from the total (All But). The lower axis shows absolute performance and the upper axis shows performance relative to the total energy. (a) Rosetta components as described in the methods with Total VdW including Eatr + Erep, and Total Elec + Sol including Epair + Esol. (b) GK energy components as described in the methods with Total VdW including VdWatr + VdWrep, and Total Elec + Sol including GB + EEF. (c) FoldX energy components as described in the methods with Total Elec including Elec + HDipole + Eleckon, Hbond including SideHbond + BackHBond, Total VdW including VdW + VdWclash and Total Elec + SolvP including Elec + HDipole + Eleckon + SolvP. (d-f) Histograms illustrating how different components of the energy functions co-vary with the overall predicted $E_{parallel} - E_{antiparallel}$ values. Only energy terms with strong covariances are shown. Covariance for all sequences is shown in black, for sequences predicted to be parallel in gray, and for sequences predicted to be antiparallel in white. (d) Rosetta components are the same as in (a). (e) GK energy components are the same as in (b). (f) FoldX energy components are the same as in (c) with TotElec the same as Total Elec.

Table 4-5. List of ESM energy components

Rosetta Energy Components

Eatr	Lennard-Jones attractive term
Erep	Linearized Lennard-Jones repulsive term
Esol	Lazardis-Karplus solvation model EEF1
Edun	Rotamer preference from Dunbrack library
Ehbnd	Hydrogen bonding
Epair	Statistical-based pair term

From Rosetta documentation

FoldX Energy Components

VdW	Surface area based VdW contributions of all atoms with respect to the same interaction in solvent
VdWclash	Steric overlap between atoms
Elec	Electrostatic contribution of charged groups
HDipole	Electrostatic interaction with helix dipole
Eleckon	Electrostatic contribution between chains associated with the k_{on} rate.[69]
SideHbond	Side chain hydrogen bonding
BackHbond	Backbone hydrogen bonding
SolvP	Surface area based solvation energy of polar groups
SolvH	Surface area based solvation energy of hydrophobic groups
EntropySC	Side-chain entropy
EntropyMC	Main-chain entropy

From FoldX documentation and references [47] and [70]

GK Energy Components

Eatr	Lennard-Jones attractive term
Erep	Lennard-Jones repulsive term
GB	Electrostatics with Generalized Born screening
EEF	Lazardis-Karplus solvation model EEF1

From reference [18]

performance when used alone (73%). Although the Rosetta electrostatics terms were poorly predictive in isolation, they significantly enhanced overall performance (removing them reduced performance from 82% to 76%). Interestingly, FoldX relied much more on a single type of term. The electrostatics term alone gave 73% prediction performance (just 3% below that of the FoldX total energy). Removing this term from the total energy reduced performance to 63%. The GK model is more similar to Rosetta than to FoldX, although it describes a more important role for electrostatics than Rosetta does. Interestingly, omitting the repulsive van der Waals energy contribution from the total energy had little effect on the performance of any of the models.

Note, however, that repulsive van der Waals terms were included when selecting the most appropriate backbone structure, and may contribute significantly in this way.

The strong predictive ability of the Rosetta van der Waals energy and the FoldX electrostatics terms suggests that these complementary descriptors could possibly be combined to give a better-performing model. However, we observed that linear combinations of these two terms performed worse than Rosetta on the test set. Extensive fitting of multiple terms to give optimal performance is not appropriate, given that the limited size of the test set restricts our ability to do rigorous cross-validation testing.

Co-variation is another way to assess which energy terms are most important for making predictions. Seeking physical insights, we used this approach to explore whether component terms contribute differently to the total energy depending on whether the final prediction is parallel or antiparallel. For both Rosetta and GK, the van der Waals energy terms co-varied strongly with the total energy (Figures 4d and e). The largest contribution came from the repulsive term, and interestingly, steric clashes were more important for examples predicted to be antiparallel than for those predicted to be parallel. Other Rosetta and GK terms, including those that describe electrostatic and solvation contributions, were smaller and exhibited less dramatic differences between parallel and antiparallel predictions. The FoldX electrostatic terms co-varied to a significant extent with the total energy (Figure 4-5f), consistent with the analysis of Figure 4-5c. However, the FoldX energy terms that differed most between parallel and antiparallel predictions were the van der Waals energy (VdW), solvation terms (SolvP and SolvH) and side-chain entropy contribution (entropySC); these each showed stronger co-variation with the total energy for parallel predictions than for antiparallel. The observations for all three energy functions described above are consistent with parallel structures being packed

more tightly than antiparallel, such that van der Waals interactions are more attractive, side-chain motions are more restricted, desolvation is greater, and clashes are more likely in the parallel orientation.

Figure 4-3 panels b-e further emphasize differences between the methods and also support the characterization of parallel and antiparallel structures suggested by the co-variation analysis. Figure 4-3b illustrates cases where differences in steric repulsion between parallel and antiparallel structures were important, as reflected by a large magnitude for the Rosetta Erep term. The GK model also recognized an effect from repulsive van der Waals interactions for these examples. All but one of the cases with large Erep terms were predicted to be antiparallel by Rosetta and GK, most of them correctly so. Further analysis revealed that 11 out of 13 such examples, including 2 incorrect predictions, had Ile residues paired at **d-d'** positions in the parallel structures; this is an interaction that is known to lead to unfavorable sterics for some well-studied parallel coiled-coil dimers.[51,58] The examples in Figure 4-3b were treated differently by FoldX, $RISP_{CC}$, and CE than by Rosetta and GK, as is expected because the former energy functions do not include a strongly repulsive steric term. Despite this, $RISP_{CC}$ and FoldX performed well on these structures. These methods capture the influence of poor packing due to steric clashes using other terms, in an overall balance that gives correct results.

Because steric clashes involving Ile residues are a candidate motif for determining orientation, we examined all such examples in the test set. There are 18 complexes in which two Ile residues were paired at **d-d'** when modeled in the parallel orientation. Rosetta correctly predicted 10 out of 10 of the antiparallel coiled coils, and only 3 of 8 of the parallel. Notably, all 8 of these parallel-orientation paired Ile residues are in terminal heptads. From the crystal structures, it is clear that the helices often fray slightly towards the ends of the supercoil to

accommodate these β-branched residues (Figure 4-6). Such fraying is not included in our idealized backbone models. To compensate for this, we tested models in which each coiled-coil heptad, or each residue, contributed its minimum energy when evaluated over all backbones. This provided a way for the radius of the supercoiled bundle to effectively vary, potentially accounting more accurately for the local context of key interactions. However, this did not improve overall performance. FoldX, which does not contain a strong repulsive term, did slightly better at predicting these structures, with 5 out of 8 parallel structures predicted correctly but only 9 out of 10 antiparallel structures correct.

Figure 4-3c highlights examples where there was a substantial difference in the Rosetta attractive van der Waals component between the parallel and antiparallel states. In these examples, this component favored the parallel orientation most of the time and indeed, complexes with large values of this term were mostly parallel. Similar patterns are seen in the CE and $RISP_{CC}$ COREatr terms, in the FoldX VdW and SolvH terms and, to a lesser extent, in the GK Eatr term. Favorable packing was offset in most models by solvation penalties, presumably because polar residues were more buried in better-packed structures. Thus, clear preferences for the antiparallel structure showed up in the FoldX SolvP and Rosetta Esol terms for examples in this panel, and, to a lesser extent, in the GK EEF term. These trends support a model where closer packing and more burial (both favorable hydrophobic burial and unfavorable polar burial) can be achieved in the parallel orientation relative to the antiparallel orientation.

Differences in electrostatics between orientations were predicted to be important by some models. For FoldX, electrostatics terms co-varied most strongly with the total energy (Figure 4-5f). Figure 4-3d shows examples that had large contributions from FoldX electrostatics (Elec, HDipole and Eleckon); these terms more often favored antiparallel structures. The GK potential

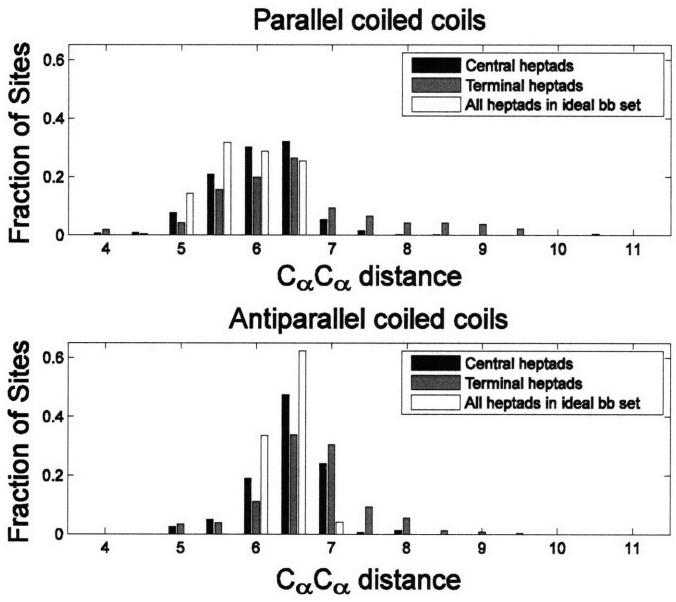


Figure 4-6: Distribution of C_α-C_α distances for core residues in parallel and antiparallel coiled coils. All C_α-C_α distances between core residues (**a-a'**, **d-d'** in parallel and **a-d'** in antiparallel) were binned by distance. For the test-set structures, residues were divided into two sets: Central heptads (black) include positions that are not among the first or last seven residues of a coiled-coil helix, and terminal heptads (gray) include residues that are among the first or last seven in a coiled-coil helix. All core positions of the ideal backbone set are binned together and shown in white.

also showed some of the FoldX trends for these examples, but the overall importance of electrostatics relative to other terms was reduced. Finally, electrostatics contributed very little to the Rosetta potential, which uses a combination of a statistically derived term (Epair) and an orientation-dependent hydrogen bond term (Ehbnd) to account for electrostatic effects.

Figure 4-5d shows a preference for parallel coiled coils in the Rosetta hydrogen-bonding term, which we suspected could include a contribution from Asn residues. A preference for paired, hydrogen-bonding Asn residues at **a-a'** positions in parallel coiled coils has been well documented and described as a determinant of coiled-coil orientation and alignment.[11,12,21] We explored whether this effect was evident in our data. Among all 131 sequence pairs tested, there were 28 examples where two Asn residues could be paired at **a-a'** sites in a parallel model. Of these, 27 were from parallel structures and only one was from an antiparallel structure (Figure 4-

3e). At least in our test set, therefore, the potential to pair Asn residues at **a-a'** is a strong indicator of a parallel orientation. This is recognized by models CE and $RISP_{CC}$. CE includes a strong preference for Asn-Asn pairing, as determined experimentally,[53] and its influence was clear in the CE COREatr term. $RISP_{CC}$ also assigns a favorable weight to this term, reflected in its COREatr term. However, the structure-based prediction methods did not show a strong energy component pattern typifying paired Asn groups. No single term dominated the predictions for these structures, although many seemed to be determined by more favorable packing in the parallel than in the antiparallel orientation. Further analysis at the residue level using Rosetta revealed that Asn hydrogen bonding favored the parallel state for only 16 out of 27 parallel examples, and the total energy of Asn residues at paired **a-a'** positions favored the parallel state in only 14 out of 27 cases. Nevertheless, 23 of 27 parallel dimers containing a pair of Asn residues were predicted correctly by Rosetta, similar to the performance on all sequences. Thus, although Asn pairs at **a-a'** positions correlate strongly with a parallel orientation in the test set, the Rosetta method did not rely heavily on this interaction to make correct predictions. This is consistent with previous observations by Grigoryan et al.[18] that the experimental preference for Asn-Asn over Asn-Val **a-a'** pairs in coiled-coil dimers is difficult to capture using these types of methods.

Confidence

To explore whether the predicted energy differences between parallel and antiparallel models can be used as a measure of confidence, we modified our scheme such that a structure was assigned as parallel (or antiparallel) only if the absolute energy difference $|E_{antiparallel} - E_{parallel}|$ was greater than some cutoff. Increasingly stringent cutoffs left larger numbers of test set

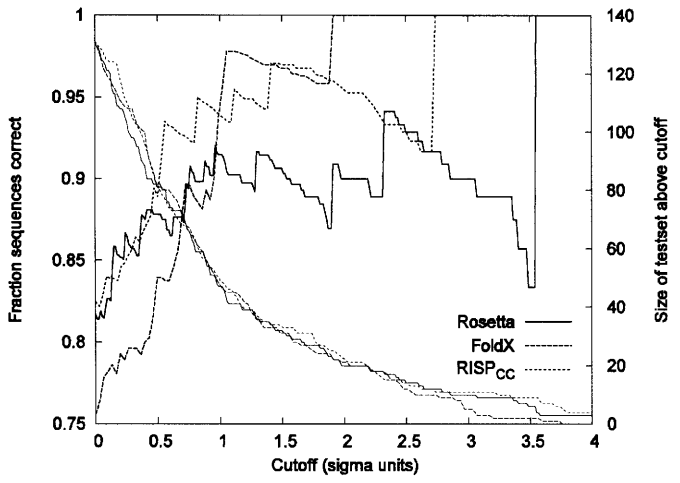


Figure 4-7: Performance as a function of increasing the gap requirement. Performance was evaluated only for those examples with |Eparallel –Eantiparallel| > x*σ and is plotted (thick lines, left axis) as a function of x. The size of the test set at each value of x is plotted using thin lines and the right axis.

examples unclassified. Figure 4-7 illustrates the tradeoff between performance and the number of classifiable structures. For the three best-performing methods, the number of predicted structures falls off quickly as performance improves. A gain of 10% prediction accuracy requires predicting between 40-60% of the test set as "unknown". Thus, although it is possible to improve the confidence of the predictions by imposing a larger energy gap, this comes at a very severe penalty.

Discussion

Our results illustrate that coiled-coil helix orientation prediction is not a trivial problem. Standard methods, applied either at the sequence or structure level, do not give good performance. Nevertheless, refinement of these approaches can provide effective predictors. For our ESMs, we found that allowing structural flexibility was important. To increase the probability that an appropriate backbone was available for each complex, each dimer was modeled on 120 different parallel and 81 different antiparallel templates. This was critical;

ultimately 52 parallel and 44 antiparallel backbones were used to construct the minimum-energy structures of both orientations for the 131 complexes modeled. Although we found in post-analysis that a much smaller set of backbones could provide the same total prediction performance, it would have been difficult to determine in advance which scaffolds these should be. Thus, although it may be possible to capture backbone variability more efficiently than we have done here (e.g. by using a better-targeted backbone library or some different approach), we have found that it is important to model flexibility to achieve good results. We also found that small amounts of structural relaxation following rigid-backbone/rotameric side-chain repacking were important. Comparing the performance of Rosetta on ideal vs. minimized backbones (Figure 4-2a) illustrates the significance of energetically costly clashes that can be removed relatively easily with minimization.

Analysis of the complexes for which ESMs gave incorrect predictions suggested that our models do not yet include sufficient structural plasticity. In particular, we found that our parallel dimer models cannot accommodate pairs of Ile residues at **d-d'** positions. This is consistent with earlier observations by Harbury et al. that β-branched residues confer a preference for trimers or tetramers over dimers when located at the **d** position of parallel homo-oligomers.[13] In native parallel structures, relatively rare Ile residues at **d** positions towards the end of the coiled-coil chain are accommodated by fraying of the ends (Figure 4-6). In contrast to this, the backbones on which we modeled these coiled coils were uniform over the length of the sequence. Incorporating greater local structural variation may be important for improving performance in the future, although our attempts to approach this in a systematic way have not succeeded so far. For now, knowledge that the structure-based methods can fail in cases where there are terminal-heptad β-branched clashes can guide appropriate use of these methods.

In the absence of more structural sampling, softening the steric repulsive term is a way to approximate structural variability. However, it is not easy to modify the ESMs to accommodate small clashes, because such clashes can be important for determining the correct helix orientation. For example, softening the repulsive terms in Rosetta or GK to accommodate Ile pairs at terminal **d** positions may prevent the proper identification of clashes elsewhere. Interestingly, FoldX lacks such a rigid repulsive term, yet is still able to correctly predict the orientation of many sequences that contain these paired residues (Figure 4-3b). Overall, our analyses support a model in which packing constraints are more demanding on parallel than on antiparallel backbones. Features of this model are captured differently by different methods. Models that include steric repulsion use this to predict that certain structures are antiparallel. Yet models that lack these terms can nevertheless recognize better packing in other ways. For FoldX, energy decomposition shows a role for the surface-area based van der Waals and hydrophobic solvation terms in favoring parallel structures. However, for sequences with large clashes (as assessed by Rosetta Erep differences), the preference of these terms for the parallel state is reduced or even reversed (Figure 4-3b). This illustrates that despite a lack of explicit steric repulsion, FoldX can still recognize poor packing that arises in structure prediction of the incorrect orientation.

The models used here, although all quite successful for the task of prediction, do not reach a significant consensus about what sequence features and energy terms are most critical for specific cases. $RISP_{CC}$, FoldX, and Rosetta are based on different sets of assumptions, and each model includes many parameters that are not derived rigorously from physical principles. GK is a more physical model, and although it may be more informative in component analysis, it did not perform quite as well. Thus, although structure-based models supposedly work by accurately

capturing physical phenomena, the large extent to which they differ in their particulars here leaves this premise in doubt (Figures 3 and 4). Our results suggest that despite good performance, caution should be observed when attempting to gain physical insight from individual energy terms in structure-based, yet highly parameterized, calculations. This is especially true given that these methods are optimized to recapitulate native structures and mutational energies, rather than to reproduce individual physical components.

Testing of various ISMs also led to interesting results. The performance of these methods was very sensitive to the choice of interfacial pairs that were scored. In particular, scoring all pairs of residues that satisfied a 4.5 Å distance cutoff in explicitly modeled structures was not effective (model $RISP_{struct}$). Scoring all pairs of residues that could *potentially* be within 4.5 Å, based on sequence and known coiled-coil dimer structures, was also not effective (model $RISP_{CC\text{-}all}$). Strikingly, however, when just 5 types of pairs were included for each orientation, performance was very good ($RISP_{CC}$). The key pairs included those that have been highlighted by many biochemical experiments over the past 10-15 years. In particular, Vinson and colleagues have quantified contributions of **a-a'**, **d-d'** and **g-e'** pairs in parallel bZIP coiled coils,[50-53] and there is an approximate structural correspondence between these and the **a-d'**, **g-g'** and **e-e'** pairs of antiparallel coiled coils, which have been less investigated.[54] The core-to-edge terms (**g-a'** and **d-e'** for parallel and **a-e'**, **d-g'** for antiparallel) provide a slight but detectable improvement in performance (Figure 4-4a). Interestingly, including additional core-core terms (**a-d'** in parallel or **a-a'**, **d-d'** in antiparallel structures) significantly degraded performance, despite recent observations by Hadley et al. that these can be significant in some antiparallel structures.[59] These results suggest that fold-recognition techniques applied to protein complexes, e.g. as are implemented in programs such as InterPreTS and Multiprospector,[60-62] could be

improved if strategies for identifying critical specificity-determining residues in different folds were available. A significant disadvantage of some of the ISMs is that they exhibit a parallel bias for homodimeric structures. It is unlikely that this preference has a physical justification, as it is not supported by the best performing ESM models. Therefore, the use of ISMs to predict coiled-coil orientation may be subject to systematic errors that favor structures in which residues interact with adjacent copies of themselves. This effect is also likely to show up in other related ISM applications.

Our results illustrate that several different types of computational approaches are capable of discriminating parallel from antiparallel coiled-coil helix alignments with reasonable accuracy. By far the most efficient of these are the sequence-based methods, which are easily scalable to evaluate candidate interactions at the proteomic scale. Structure-based methods are less prone to biases, however, and these methods could also be scaled up for some types of applications. Our recently developed cluster-expansion methodology, in which a simple expression for energy as a function of sequence can be fit to the results of more expensive calculations, is a promising way of approaching this problem.[63,64] However, significant challenges remain before accurate tertiary/quaternary annotation can be provided for novel coiled-coil sequences. Techniques must be developed that can recognize the correct set of interacting helices and their appropriate stoichiometries. When sequences are of different lengths, the correct axial alignment must also be selected. Our demonstration of helix-orientation prediction in a rigorously chosen subset of examples represents an important and necessary component of this larger-scale genomic annotation problem.

Possible Future Directions

The work described above and in Chapter 2 has shown that one can capture the structural deviation of parallel and antiparallel dimeric coiled coils using the extended Crick parameterization method. The original Crick Parameterization method had been used to generate higher order parallel coiled coils that have C_n symmetry about the superhelical axis.[23] The extended method could easily be expanded to include any arbitrary number of helices in any predefined orientation. However, even though any structure could be generated, identification of the native structure space is currently limited by the few structural examples of higher-order coiled-coils oligomers in the PDB. This could be overcome with further development of the energy functions to better predict the native-like Crick parameters for coiled-coil sequences.

Another improvement would be to allow for local structural sampling throughout the coiled coil when predicting structures and orientations. This could be included in several different ways: The simplest approach would be to relieve some of the restraints on ends of the coiled coils to allow for some amount of fraying. This could be managed by employing the Crick constraint with variable force constants throughout the structure. Another method would be to allow for differing coiled-coil parameters throughout the structure. These could be varied for every heptad or even changed every C_α position. The minimization procedure could then use these C_α positions to constrain the structure, as has been done previously.

A significant problem that still needs to be addressed is one of alignment prediction. In the orientation prediction test described in this Chapter only parallel and antiparallel coiled coils with maximal overlap were included. This left no ambiguity as to the alignment of both the parallel and antiparallel state, i.e which sets of residues would be interacting. The ability to predict the alignment for an arbitrary pair of sequences is still a difficult problem. Some work

has been done on a related problem, partnering prediction. Here the goal is to predict the relative binding affinity for a set of different partners. This problem has been explored for a set bZIP coiled coils.[16,24,25] However, while the current methods are successful, it has not been shown that methods for partner prediction will perform the same for alignment prediction. Moreover, these tests have only been limited to a small class of parallel sequences. As the test-set of coiled coil sequences becomes more diverse the accuracy of the models and the energy functions will need to be much improved.

Methods

For descriptions of the coiled-coil database, Crick parameterization and generation of Crick backbones, see Chapter 2.

Evaluation of structures

Sequences were repacked on 201 parallel + antiparallel rigid backbones using Rosetta with default parameters and expansion of the first and second dihedral angles in the rotamer library.[44] The energy of these repacked structures was recorded to provide the Rosetta energy. Repacked structures were then converted to CHARMM 19 atom types and minimized using CHARMM with param19 EEF1 parameters and topology.[45,46] The energy function used in minimization included van der Waals; EEF1 solvation; distance-dependent-dielectric electrostatics with dielectric constant of 4r; bond length, angle, dihedral angle, and improper dihedral molecular mechanics energy; hydrogen bond energy; and the Crick user energy. Minimization was done with 1000 steps of steepest decent followed by 1000 steps of adopted-

basis Newton-Raphson. These minimized structures were then re-evaluated using five ESM energy functions.

Energy functions – ESMs

All Crick-minimized backbones were evaluated with each ESM. The lowest energy structure in each orientation was used to determine the energy difference. All structures were held fixed during evaluation.

The Rosetta energy was calculated using the same energy function as for repacking. All energy terms were included in the final score; however, the structure-independent reference state canceled in the final analysis. Energy components labeled in the figures for Rosetta are: Eatr – attractive van der Waals; Erep – repulsive van der Waals; Epair – statistical pair electrostatics; Ehbnd – hydrogen bonding; Esol – solvation; and Edun – Dunbrack statistical energy.

Model GK uses the physical energy function described by Grigoryan and Keating.[18] Briefly, the energy function consists of three terms. First, a van der Waals energy term includes atomic radii from CHARMM param19.[45] Second, an electrostatics energy term combines Coulombic interaction energy in a uniform dielectric of 4 with Generalized Born (GB) screening to account for transfer into an external dielectric of 80 and an internal dielectric of 4. Perfect Born radii for use in the GB formulae were calculated using PEP.[65] Finally, a desolvation energy term is included from the EEF1 function in CHARMM.[46] Energy components labeled in the figures for GK are: VdWatr and VdWrep – attractive and repulsive van der Waals; GB – screened Coulombic interaction energy; EEF – EEF1 solvation component.

The DFIRE statistical potential was applied by using binding energies computed using the dcomplex executable, as obtained from the Zhou lab.[48]

The FoldX energy was calculated with FoldX version 2.5.2 obtained from the Serrano laboratory.[47,66] We used the "Stability" command with all options set to their default values. All energy terms contributed to the final score. Energy components labeled in the figures for FoldX are: VdW – van der Waals; VdWclash – van der Waals clash; Elec+HDipole+Eleckon – sum of electrostatic, helix-dipole electrostatic and electrostatic k_{on}; SideHBond+BackHBond – sum of side-chain and backbone hydrogen bonding; SolvP – polar solvation energy; SolvH – hydrophobic solvation energy; and EntropySC+EntropyMC – sum of side-chain and backbone entropy.

RISP (Residue-based Interfacial Statistical Potential) was derived using the framework outlined by Lu et al.[62] It was based on protein complexes from the QS50 database at 3dcomplex.org,[67] which consists of PDB entries filtered to exclude all complexes with greater than 50% sequence identity. We further excluded all structures showing significant sequence homology (BLAST $E < 10^{-10}$) to structures in our coiled-coil test set. An interface between two chains was defined as the set of all residues with any heavy atom within 4.5 Å of the other chain. Interfaces containing 5 or fewer residues were excluded. To reduce the observed bias of the derived potential towards favoring homodimeric interactions, interfaces were excluded if they contained two or more residues making contact with copies of themselves on other chains. The final database consisted of 2,864 interfaces containing 105,287 residues. Pair-wise residue scores were computed according to:

$$P(i,j) = -\log \frac{N_{obs}(i,j)}{N_{\exp}(i,j)} \tag{4-1}$$

where $N_{obs}(i, j)$ is the number of contacts observed between residues i and j in the training database and $N_{exp}(i, j)$ is the product of the mole fractions of residues i and j in the database multiplied by the total number of residues in the database. This reference state performed better

at orientation discrimination compared to a reference state based on the mole fraction of residues occurring in solvent-exposed positions.[62] The RISP potential was applied to modeled coiled-coil structures as a sum of pair-wise residue contact scores. Contacts were determined according to the same criteria used in the development of the potential.

Energy functions - ISMs

A null control model (NULL) was developed by assigning random scores between +1 and -1 to all possible amino acid pairs at **a-a'**, **d-d'**, and **g-e'** (parallel) or **a-d'**, **e-e'**, and **g-g'** (antiparallel) positions.

Model ELEC assigns all occurrences of **g-e'** (parallel) or **g-g'** + **e-e'** (antiparallel) E-R, R-E, K-E or E-K pairs a weight of -1, while E-E, R-R, R-K, K-R, K-K, D-E, E-D and D-D pairs are given a weight of +1.

The CE model is constructed using 48 experimentally determined coupling energies for each orientation. For parallel coiled coils, coupling energies were obtained from references Krylov et al.[50] and Acharya et al.[52] For antiparallel coiled coils, we computed coupling energies for **a-d'** residue pairs from the ΔG values of Hadley et al. as double mutant thermodynamic cycles relative to alanine.[57] Because no published data are available for antiparallel interactions involving **g** and **e** residues, we applied the analogous values from the Krylov study to the antiparallel pairs **g-g'** and **e-e'**.

To apply RISP to sequence data, we predefined pairs of heptad positions to be scored. Different models included different pairs, as follows: $RISP_{core}$ included core interactions: **a-a', d-d'** (parallel) and **a-d'** (antiparallel) pairs. $RISP_{edge}$ included edge interactions: **g-e'** (parallel) and **g-g'**, **e-e'** (antiparallel) pairs. $RISP_{core,edge}$ included the pairs in both $RISP_{core}$ and $RISP_{edge}$.

$RISP_{CC}$ included all pairs from $RISP_{core,edge}$ as well as the core-edge pairs **g-a'**, **d-e'** (parallel) and **a-e'**, **d-g'** (antiparallel). Finally, the $RISP_{all}$ model further included the pairs **d-a'** (parallel) and **a-a'**, **d-d'** (antiparallel). These lists are summarized in Table 4-4. Energy components used in Figure 4-3 for $RISP_{CC}$ are: COREatr/rep – all core-core interactions; EDGEatr/rep – all edge-edge interactions; CEatr/rep – all core-edge interactions. Based on analyses of coiled-coil crystal structures, $RISP_{all}$ corresponds to selecting all pairs with the potential to be in contact according to the 4.5 Å criterion used to develop RISP.

Acknowledgments

We acknowledge funding from National Institutes of Health grant GM67681 and National Science Foundation CAREER award MCB-0347203. Computer equipment to support this work was purchased under NSF award 0216437. We thank G. Grigoryan for thoughtful discussions and useful computer code, and T. C. S. Chen, O. Ashenberg, X. Fu and M. Radhakrishnan for comments on the manuscript. We also thank Tom Alber and Mark Sales for the fitcc source code.

References

1. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS. Predicting coiled coils by use of pairwise residue correlations. Proc Natl Acad Sci U S A 1995;92(18):8259-8263.
2. Delorenzi M, Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 2002;18(4):617-625.
3. McDonnell AV, Jiang T, Keating AE, Berger B. Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 2006;22(3):356-358.
4. Wolf E, Kim PS, Berger B. MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci 1997;6(6):1179-1189.
5. Woolfson DN, Alber T. Predicting oligomerization states of coiled coils. Protein Sci 1995;4(8):1596-1607.
6. Fong JH, Keating AE, Singh M. Predicting specificity in bZIP coiled-coil protein interactions. Genome Biol 2004;5(2):R11.

7. Lupas AN, Gruber M. The structure of alpha-helical coiled coils. Advances in protein chemistry 2005;70:37-78.
8. Tripet B, Wagschal K, Lavigne P, Mant CT, Hodges RS. Effects of side-chain characteristics on stability and oligomerization state of a de novo-designed model coiled-coil: 20 amino acid substitutions in position "d". Journal of molecular biology 2000;300(2):377-402.
9. Wagschal K, Tripet B, Lavigne P, Mant C, Hodges RS. The role of position a in determining the stability and oligomerization state of alpha-helical coiled coils: 20 amino acid stability coefficients in the hydrophobic core of proteins. Protein Sci 1999;8(11):2312-2329.
10. Liu J, Zheng Q, Deng Y, Kallenbach NR, Lu M. Conformational transition between four and five-stranded phenylalanine zippers determined by a local packing interaction. Journal of molecular biology 2006;361(1):168-179.
11. Oakley MG, Kim PS. A buried polar interaction can direct the relative orientation of helices in a coiled coil. Biochemistry 1998;37(36):12603-12610.
12. Lumb KJ, Kim PS. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. Biochemistry 1995;34(27):8642-8648.
13. Harbury PB, Zhang T, Kim PS, Alber T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science 1993;262(5138):1401-1407.
14. Taylor CM, Keating AE. Orientation and oligomerization specificity of the Bcr coiled-coil oligomerization domain. Biochemistry 2005;44(49):16246-16256.
15. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 2003;53(1):76-87.
16. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A 2001;98(18):10125-10130.
17. Vieth M, Kolinski A, Brooks CL, 3rd, Skolnick J. Prediction of quaternary structure of coiled coils. Application to mutants of the GCN4 leucine zipper. Journal of molecular biology 1995;251(3):448-467.
18. Grigoryan G, Keating AE. Structure-based prediction of bZIP partnering specificity. Journal of molecular biology 2006;355(5):1125-1142.
19. Mason JM, Schmitz MA, Muller KM, Arndt KM. Semirational design of Jun-Fos coiled coils with increased affinity: Universal implications for leucine zipper prediction and design. Proc Natl Acad Sci U S A 2006;103(24):8989-8994.
20. Fassler J, Landsman D, Acharya A, Moll JR, Bonovich M, Vinson C. B-ZIP proteins encoded by the Drosophila genome: evaluation of potential dimerization partners. Genome Res 2002;12(8):1190-1200.
21. Vinson C, Myakishev M, Acharya A, Mir AA, Moll JR, Bonovich M. Classification of human B-ZIP proteins based on dimerization properties. Mol Cell Biol 2002;22(18):6321-6335.
22. Gurnon DG, Whitaker JA, Oakley MG. Design and characterization of a homodimeric antiparallel coiled coil. J Am Chem Soc 2003;125(25):7518-7519.

23. McClain DL, Woods HL, Oakley MG. Design and characterization of a heterodimeric coiled coil that forms exclusively with an antiparallel relative helix orientation. J Am Chem Soc 2001;123(13):3151-3152.
24. Monera OD, Kay CM, Hodges RS. Electrostatic interactions control the parallel and antiparallel orientation of alpha-helical chains in two-stranded alpha-helical coiled-coils. Biochemistry 1994;33(13):3862-3871.
25. Monera OD, Zhou NE, Lavigne P, Kay CM, Hodges RS. Formation of parallel and antiparallel coiled-coils controlled by the relative positions of alanine residues in the hydrophobic core. J Biol Chem 1996;271(8):3995-4001.
26. Myszka DG, Chaiken IM. Design and characterization of an intramolecular antiparallel coiled coil peptide. Biochemistry 1994;33(9):2363-2372.
27. Schnarr NA, Kennan AJ. Strand orientation by steric matching: a designed antiparallel coiled-coil trimer. J Am Chem Soc 2004;126(44):14447-14451.
28. Gernert KM, Surles MC, Labean TH, Richardson JS, Richardson DC. The Alacoil: a very tight, antiparallel coiled-coil of helices. Protein Sci 1995;4(11):2252-2260.
29. Berger B, Singh M. An iterative method for improved protein structural motif recognition. J Comput Biol 1997;4(3):261-273.
30. Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, Moelling K, Volkmer-Engert R, Oschkinat H. Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. Journal of molecular biology 2004;343(3):703-718.
31. Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 2003;31(13):3635-3641.
32. Brannetti B, Via A, Cestra G, Cesareni G, Helmer-Citterich M. SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. Journal of molecular biology 2000;298(2):313-328.
33. McClain DL, Binfet JP, Oakley MG. Evaluation of the energetic contribution of interhelical Coulombic interactions for coiled coil helix orientation specificity. Journal of molecular biology 2001;313(2):371-383.
34. Walshaw J, Woolfson DN. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. Journal of molecular biology 2001;307(5):1427-1450.
35. Walshaw J, Woolfson DN. Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. J Struct Biol 2003;144(3):349-361.
36. Newman JR, Keating AE. Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science 2003;300(5628):2097-2101.
37. Supekar VM, Bruckmann C, Ingallinella P, Bianchi E, Pessi A, Carfi A. Structure of a proteolytically resistant core from the severe acute respiratory syndrome coronavirus S2 fusion protein. Proc Natl Acad Sci U S A 2004;101(52):17958-17963.
38. Strelkov SV, Schumacher J, Burkhard P, Aebi U, Herrmann H. Crystal structure of the human lamin A coil 2B dimer: implications for the head-to-tail association of nuclear lamins. Journal of molecular biology 2004;343(4):1067-1080.
39. Oakley MG, Kim PS. Protein dissection of the antiparallel coiled coil from Escherichia coli seryl tRNA synthetase. Biochemistry 1997;36(9):2544-2549.
40. Lumb KJ, Carr CM, Kim PS. Subdomain folding of the coiled coil leucine zipper from the bZIP transcriptional activator GCN4. Biochemistry 1994;33(23):7361-7367.

41. Harbury PB, Tidor B, Kim PS. Repacking Protein Cores with Backbone Freedom - Structure Prediction for Coiled Coils. Proceedings of the National Academy of Sciences of the United States of America 1995;92(18):8408-8412.
42. Crick FH. The Fourier Transform of a Coiled-Coil. Acta Cryst 1953;6:685-689.
43. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.
44. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302(5649):1364-1368.
45. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comp Chem 1983;4(2):187-217.
46. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35(2):133-152.
47. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of molecular biology 2002;320(2):369-387.
48. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11(11):2714-2726.
49. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys J 2003;84(3):1895-1901.
50. Krylov D, Mikhailenko I, Vinson C. A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. Embo J 1994;13(12):2849-2861.
51. Moitra J, Szilak L, Krylov D, Vinson C. Leucine is the most stabilizing aliphatic amino acid in the d position of a dimeric leucine zipper coiled coil. Biochemistry 1997;36(41):12567-12573.
52. Acharya A, Ruvinov SB, Gal J, Moll JR, Vinson C. 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 2002;41(48):14122-14131.
53. Acharya A, Rishi V, Vinson C. Stability of 100 homo and heterotypic coiled-coil a-a' pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R). Biochemistry 2006;45(38):11324-11332.
54. Oakley MG, Hollenbeck JJ. The design of antiparallel coiled coils. Curr Opin Struct Biol 2001;11(4):450-457.
55. McClain DL, Gurnon DG, Oakley MG. Importance of potential interhelical salt-bridges involving interior residues for coiled-coil stability and quaternary structure. Journal of molecular biology 2002;324(2):257-270.
56. Campbell KM, Sholders AJ, Lumb KJ. Contribution of buried lysine residues to the oligomerization specificity and stability of the fos coiled coil. Biochemistry 2002;41(15):4866-4871.
57. Hadley EB, Gellman SH. An antiparallel alpha-helical coiled-coil model system for rapid assessment of side-chain recognition at the hydrophobic interface. J Am Chem Soc 2006;128(51):16444-16445.

58. Harbury PB, Zhang T, Kim PS, Alber T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science 1993;262(5138):1401-1407.
59. Hadley EB, Testa OD, Woolfson DN, Gellman SH. Preferred Side-chain Constellation at Antiparallel Coiled-Coil Interfaces. Proc Natl Acad Sci U S A 2008;105(2):530-535.
60. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci U S A 2002;99(9):5896-5901.
61. Aloy P, Russell RB. InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 2003;19(1):161-162.
62. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins 2002;49(3):350-364.
63. Grigoryan G, Zhou F, Lustig SR, Ceder G, Morgan D, Keating AE. Ultra-fast evaluation of protein energies directly from sequence. PLoS Comput Biol 2006;2(6):e63.
64. Zhou F, Grigoryan G, Lustig SR, Keating AE, Ceder G, Morgan D. Coarse-graining protein energetics in sequence variables. Phys Rev Lett 2005;95(14):148103.
65. Beroza P, Fredkin DR. Calculation of amino acid pK(a)s in a protein from a continuum electrostatic model: Method and sensitivity analysis. J Comput Chem 1996;17(10):1229-1244.
66. Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proc Natl Acad Sci U S A 2005;102(29):10147-10152.
67. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2006;2(11):e155.
68. Hubbard S, Thorton JM. NACCESS, Computer Program 2.1.1 edit.: Department of Biochemistry and Molecular Biology (U. C. L., ed.); 1996.
69. Selzer T, Albeck S, Schreiber G. Rational design of faster associating and tighter binding protein complexes. Nat Struct Biol 2000;7(7):537-541.
70. Pey AL, Stricher F, Serrano L, Martinez A. Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. Am J Hum Genet 2007;81(5):1006-1024.

Chapter 5

Sequence based evaluation of protein energies using multiple backbones

Abstract

The introduction of backbone flexibility into discrete structural models has proved vital in many different protein design applications. However, this increase in accuracy comes at the cost of computational complexity. Previously, Grigoryan and co-workers introduced a method for increasing the efficiency of protein design utilizing cluster expansion.[1,2] This method allows for the conversion of structure-based energies into sequence-based energy functions that permit for a rapid speed-up in the calculations. Here cluster expansion is extended to include backbone flexibility. Using two different design applications, cluster expansion is shown to be a better predictor of structure-based energies for flexible-backbone models than for the corresponding fixed-backbone models.

Introduction

The fixed-backbone approximation greatly simplifies protein design by reducing the conformation search to a set of side-chain rotamers.[3] This approximation has been used

successfully for a number of computational protein design applications.[4-10] However, as discussed in Chapter 1, the fixed-backbone approximation limits the design space to native-like sequences.[11-13] This limitation was apparent in the Bcl-x_L inhibitor design where all the sequences designed on the crystal structure backbone clustered together (Figure 3-7). Also in this Figure, it is shown that the size of the sequence space was enhanced by including backbone flexibility.[14-16] As discussed in Chapter 1, there are many different methods for incorporating backbone flexibility that have been used in design. However, all of these methods suffer from the same drawback, increased computational complexity.

A significant improvement in the search time in protein design was introduced by Grigoryan and co-workers.[1,2] They demonstrated that structure-based energies could effectively be translated into a sequence-based function using a technique called cluster expansion (CE). CE generates a set of basis functions, called cluster functions (CF) that describe the energetic contributions of particular amino acids. These are broken down into single amino-acids terms, terms for pairs of amino acids, terms for triplets and higher order terms. The energy of each sequence is fit as a linear combination of these CFs and the resultant weights are called effective cluster interactions (ECI). When expanded to all terms the total structure-based energy can be fit exactly. However, in the design of a coiled-coil, zinc finger and WW domain, Grigoryan et al. showed that only single-body, two-body and a few three- or four-body terms were necessary to provide a good approximation of the total energy.[1]

One limitation of this work was that all of the designs utilized a fixed backbone and only the side-chain placement was variable. However, if backbone flexibility is included in this approach, each sequence will still have a unique lowest-energy structure. Therefore, the methodology used to fit the structure-based energies should remain valid. The inclusion of

different backbones means that pairs of side chains will have different interactions environments, depending not only on their local sequence, but also on the backbone on which they are evaluated. The expectation was that this increase in complexity could compromise the quality of the fit or require a greater number of higher order terms to achieve the same quality of fit. Despite these expectations, here we show that using flexible backbones for CE gives similar or better results than those using only a single fixed backbone. This result was validated using two examples previously shown to be effective for designing novel proteins: the design of helical inhibitors of Bcl-x_L, as described in Chapter 3, and the re-design of a zinc finger protein.[6]

Results

To test the validity of CE for the use of multiple backbones I examined two different design cases: the design of Bcl-x_L inhibitors using backbones varied through normal-mode analysis and the re-design of a zinc finger using a set of homologous PDB structures to introduce backbone variation.

Bcl-x_L inhibitor design

It was shown in Chapter 3 that flexible backbone models could be used to design novel BH3 peptides that bind to a Bcl-x_L receptor. In that experiment, the BH3 helix was varied using normal-mode analysis. An ensemble of starting structures was generated from a single x-ray structure and design was performed on 11 sites of the peptide. Here, I limited the structural search to the 16 backbones that produced the most low energy design sequences. At each of 11 design sites, all amino acids that were allowed in any design library over all backbones were included. A list of these is in Table 5-1. From this sequence space, 10,000 random sequences

Table 5-1: List of amino acid allowed at each position in the Bcl-x_L inhibitor design.

Position	Amino Acids
B3	Arg, Asn, Asp, Glu, Gly, His, Met, Phe, Ser, Thr, Tyr
B6	Ala, Gly, Ile, Leu, Met, Phe, Val
B7	Ala, Gly
B10	Ala, Gly, Ile, Leu, Met, Phe, Val
B11	Ala, Arg, Asn, Asp, Gln, Glu, Lys, Met, Phe, Ser
B13	Ala, Gly, Ile, Leu, Met, Phe, Val
B14	Ala, Gly
B15	Arg, Asn, Asp, Gln, Glu, His, Lys, Met, Phe, Tyr
B17	Ala, Gly, Ile, Leu, Met, Phe, Val
B18	Asn, Asp, Gln, Glu, His, Phe, Ser, Tyr
B21	Ala, Arg, Asn, Gly, Ile, Lys, Phe, Ser, Thy, Tyr, Val

were generated. Each of these sequences was repacked onto all 16 backbones in the set and the structure-based energy was evaluated as described in the Methods. The flexible-backbone energy for each sequence was defined as the minimum energy over all backbones. This used a wide distribution of backbones, suggesting the importance of allowing flexibility (Figure 5-1). The ECI values were then obtained by CE of the energy of these sequences, including all point and pair CFs. For comparison, ECI values were also obtained for the energy of the same sequences on the native backbone. After the training set ECI values were fit, another set of 10,000 sequences, none with 100% sequence identity to any training set sequence, was generated from the same distribution. These sequences were also repacked and the energy evaluated on all backbones. The flexible-backbone energy of the test set was also taken to be the minimum energy of each sequence over all backbones. The distribution of backbones is also shown in Figure 5-1.

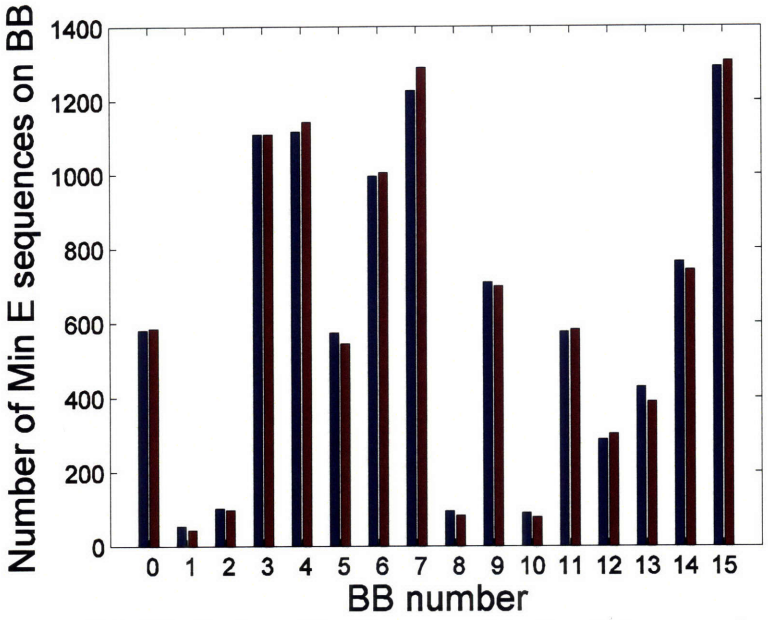


Figure 5-1: Distribution of backbones used in the cluster expansion for the Bcl-x_L inhibitor design. Backbone number 0 corresponds to the fixed backbone reference. In red are the backbones for the training set and in blue those for the test set.

For the fixed- and flexible-backbone designs, the performance of the respective CEs on predicting the energies of the training set sequences and test set sequences is shown in Figures 5-2 a and c, and 5-3 a and c. The leave-one-out cross-validated root mean square deviation (cv_rmsd) for the training set data and the root mean square deviation (rmsd) for the test set data were also calculated. For both the training and test sets, CE was a better predictor of the flexible-backbone structure-based energy than the fixed-backbone structure-based energy. For the fixed-backbone case, the cv_rmsd for the training set was 3.3 kcal/mol and the rmsd for the test set was 3.6 kcal/mol. For the flexible backbone case, the cv_rmsd for the training set was 2.9 kcal/mol and the rmsd for the test set was 2.5 kcal/mol. To try to improve the expansion, I also included triplet CFs. These were selected by identifying point and pairs CFs that were strongly correlated. These point and pair CFs were combined into triplets and added to the original CE one at a time, to see if they significantly improved the cv_rmsd. An additional 144 CFs were added for the single backbone CE and 127 for the flexible backbone CE. The performance of the

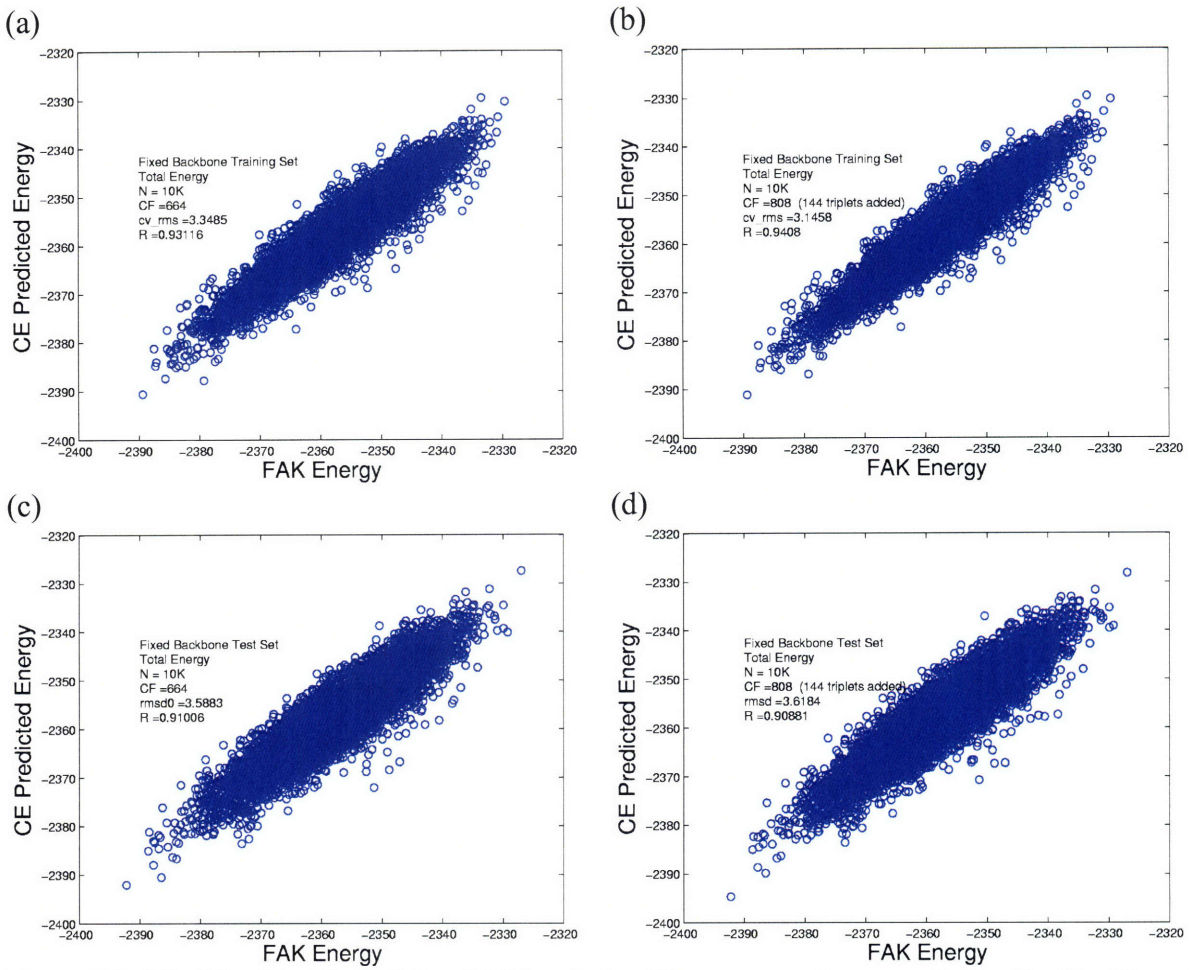


Figure 5-2: CE of fixed-backbone Bcl-x_L inhibitor design. Plots compare the energy from a fixed backbone design (FAK Energy) to the CE energy. (a-b) Training-set data, where (a) is for CE with only point and pair terms, and (b) includes triplet terms. (c-d) Test-set data where (c) is for evaluation using only the point and pair terms, and (d) includes triplet terms.

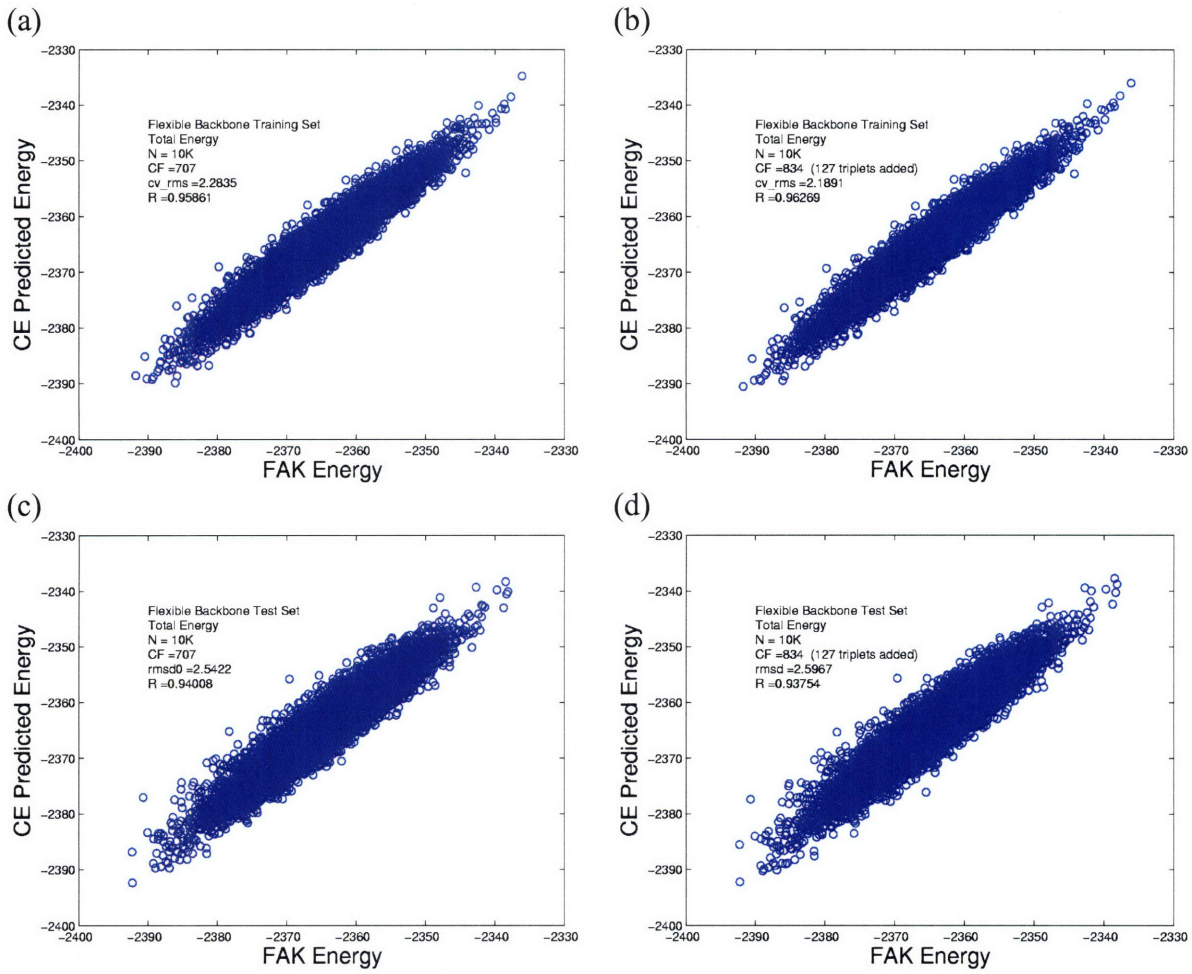


Figure 5-3: CE of flexible-backbone Bcl-x_L inhibitor design. Plots compare the energy from a fixed backbone design (FAK Energy) to the CE energy. (a-b) Training-set data, where (a) is for CE with only point and pair terms, and (b) includes triplet terms. (c-d) Test-set data where (c) is for evaluation using only the point and pair terms, and (d) includes triplet terms.

CEs with added triplets on predicting the energies of the training set sequences and test set sequences are shown in Figures 5-2 b and d and 5-3 b and d. The improvement in cv_rmsd was small for both cases (3.1 kcal/mol for fixed-backbone and 2.2 kcal/mol for flexible-backbone). This incremental improvement was not seen in the test set for the fixed-backbone (rmsd of 3.6 kcal/mol) or the flexible backbone (rmsd of 2.6 kcal/mol). This suggests that the increased numbers of CFs resulted in over-training or the new triplets contained information already described by the pair and point clusters.

Zinc finger design

Dahiyat and Mayo previously performed a *de novo* designs of a zinc finger protein.[6] This was one of the targets selected by Grigoryan et al. for validation of cluster expansion.[1] Given this, we explored the use of CE on multiple zinc-finger backbones. To increase the structure space for the zinc finger, additional backbones were selected using the SCOP database.[17] In the previous design, the second zinc finger domain of chain C from the PDB structure 1zaa was used. I searched the SCOP database to find all x-ray crystal structures of zinc fingers with a continuous 21-residue region homologous to the 1zaa structure. This produced a total of 29 structures. This set was clustered using pairwise C_α-backbone rmsd. The clustergram is shown in Figure 5-4a. From this set of structures, 10 candidate backbones were selected that spanned the range of backbone space and included the original backbone structure (Figure 5-4b).

For all 10 backbones, a set of 50,000 sequences were repacked and evaluated using the same energy function as in Grigoryan et al.[1] The sequences were chosen randomly from a distribution of residues described by Dahiyat and Mayo[6] and Grigoryan et al.[1] Again the flexible-backbone energy for a sequence was taken as the minimum energy over all backbones. The distribution of backbones selected is shown in Figure 5-5. The number of minimum energy sequences for each backbone was not evenly distributed. However, all of the backbones contributed significantly to the total. This uneven distribution most likely was caused by the fact that these backbones do not represent an even sampling of the backbone space, and are of varying quality. As a comparison, the energy for the original structure, 1zaa, was used for fixed-backbone CE. This backbone generated the most minimum energy sequences (Figure 5-5). CE was performed using all point terms and all pair terms for both the fixed- and flexible-backbone

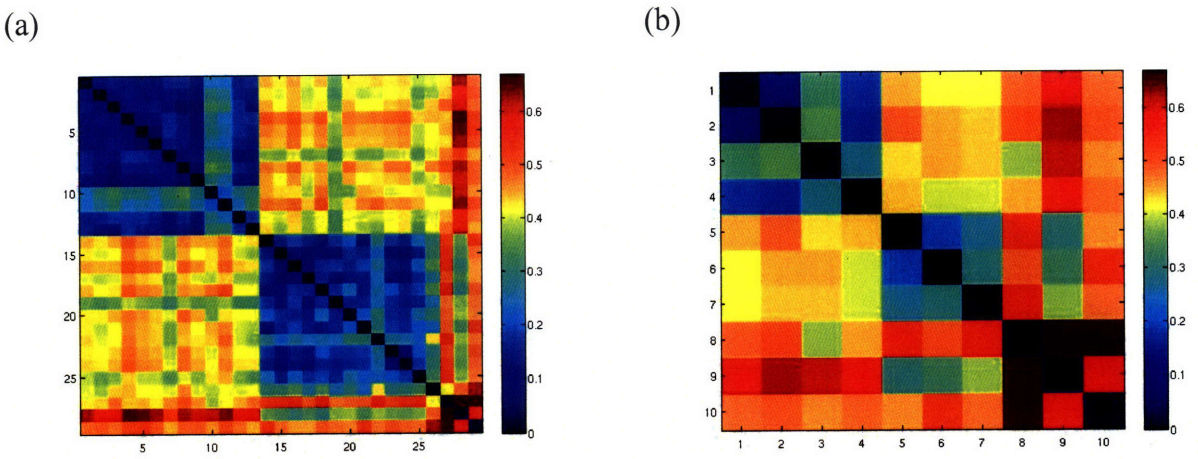


Figure 5-4: Clustering of zinc-finger structures. (a) All 29 zinc-finger structures are clustered by the pairwise C_α-backbone rmsd. The rmsd between structures is shown by the heat map coloring ranging from 0 in dark blue to 0.65 in dark red. (b) Subset of structures chosen for cluster expansion test. Structures are in the same order as in part (a) and the color scheme is the same.

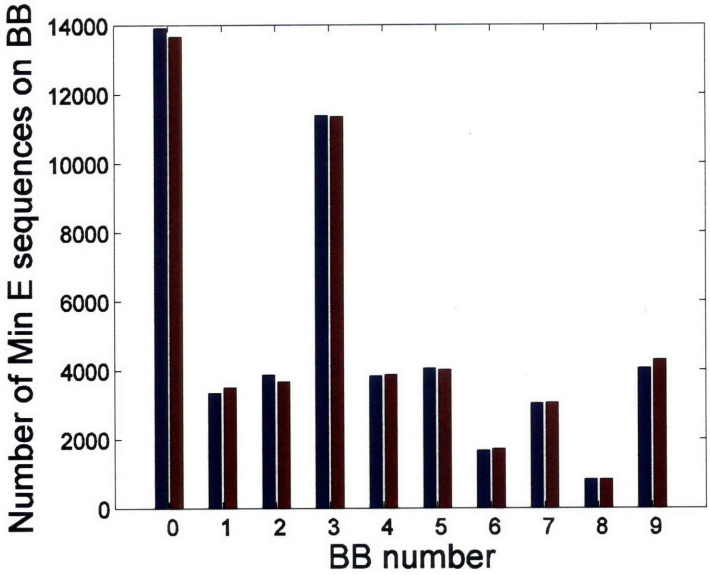


Figure 5-5: Distribution of backbones used in the CE for the zinc-finger design. Backbone 0 corresponds to the fixed backbone reference, 1zaa. In red are the backbones for the training set and in blue those for the test set.

cases. Figures 5-6 a and 5-7 a show the performance of the CE of the training set sequences for the fixed and flexible backbone cases. It is clear from these Figures that inclusion of multiple backbones improved the overall fit of the structure based energies. The cv_rmsd of the training set sequences was 4.2 kcal/mol for the fixed backbone case and 2.8 kcal/mol for flexible-backbone case.

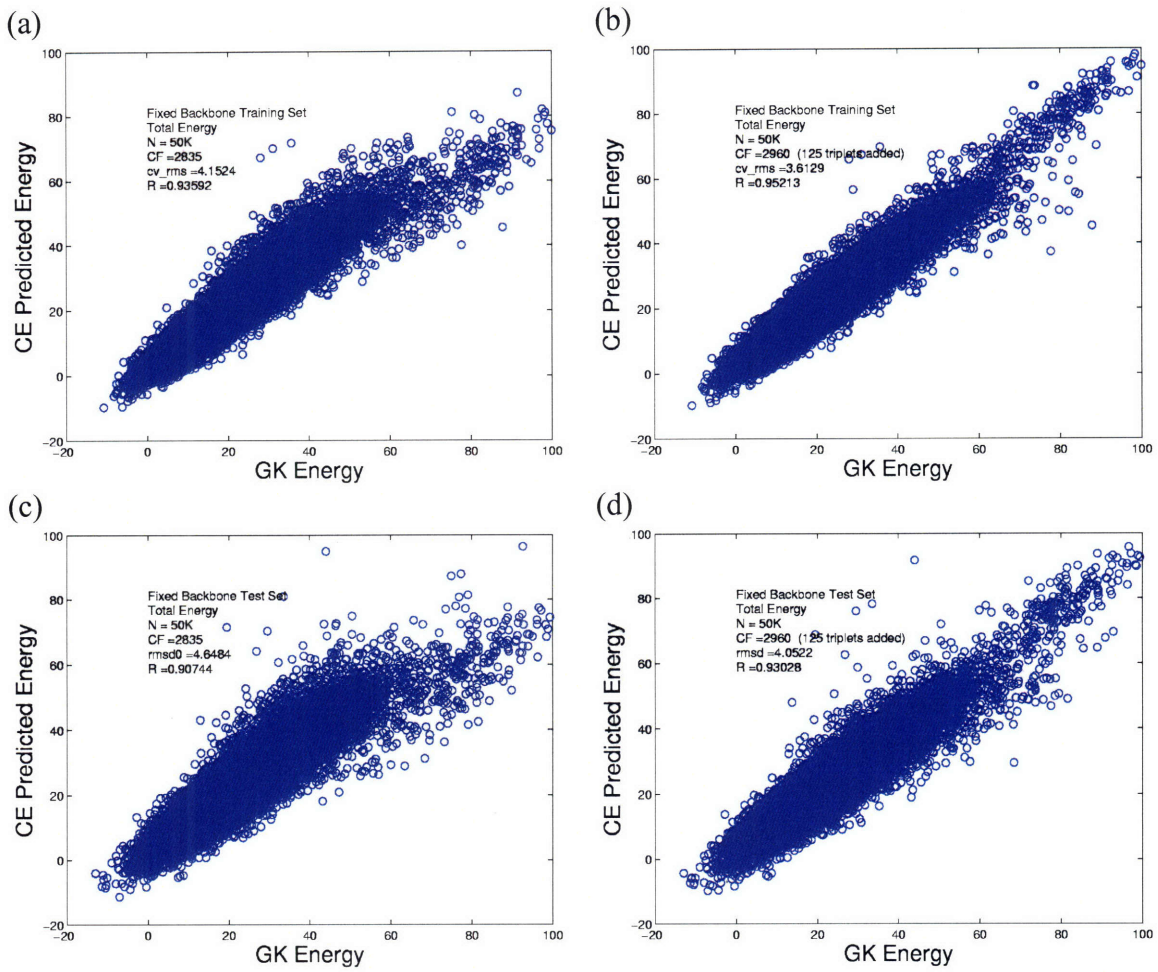


Figure 5-6: CE of fixed-backbone zinc-finger design. Plots compare the energy from a fixed backbone design (GK Energy) to the CE energy. (a-b) Training-set data, where (a) is for CE with only point and pair terms, and (b) includes triplet terms. (c-d) Test-set data where (c) is for evaluation using only the point and pair terms, and (d) includes triplet terms.

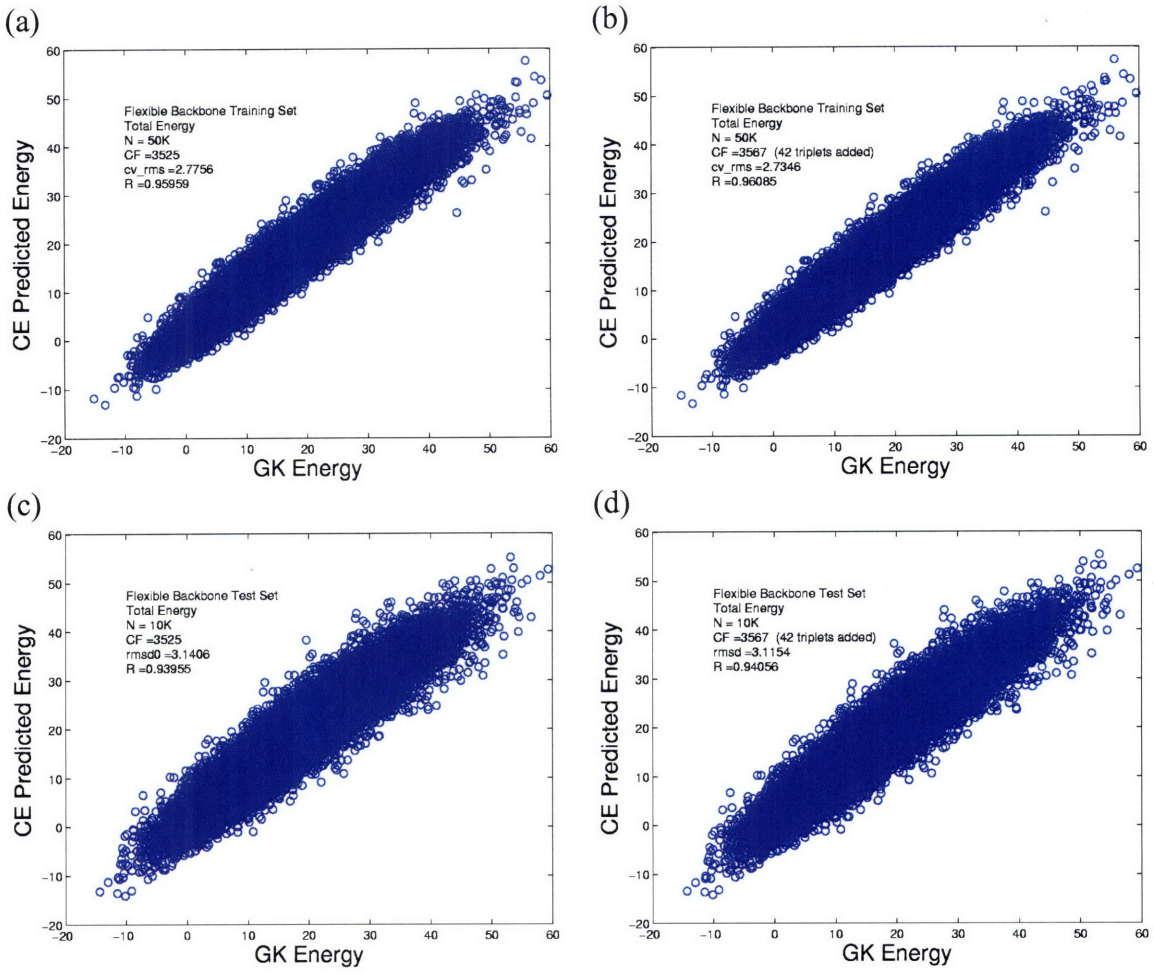


Figure 5-7: CE of flexible-backbone zinc-finger design. Plots compare the energy from a fixed backbone design (GK Energy) to the CE energy. (a-b) Training-set data, where (a) is for CE with only point and pair terms, and (b) includes triplet terms. (c-d) Test-set data where (c) is for evaluation using only the point and pair terms, and (d) includes triplet terms.

Next, 50,000 test-set sequences were generated using the same distribution. The energy of these sequences was evaluated on all backbones using the same structure-based methods. For the fixed-backbone and flexible-backbone cases the performance of the respective CEs on predicting the energies of the test set sequences is shown in Figures 5-6c and 5-7c. The rmsd for the test set was 4.6 kcal/mol for the fixed backbone case and 3.1 kcal/mol for the flexible backbone case. Triplet CFs were sought to improve the performance of the training set using the same covariance method as described above for the Bcl-x_L inhibitor design. This added 125 and

44 triplet CFs to the fixed- and flexible-backbone CEs respectively. The results of the CE prediction of the training and test set sequences are shown in Figures 5-6 b and d and 5-7 b and d. Just as before, there was not a significant improvement for the flexible-backbone case, with a cv_rmsd to the training set of 2.7 kcal/mol and a rmsd for the test set of 3.1 kcal/mol. However, in the fixed backbone case there was a significant improvement for both the training set (cv_rmsd of 3.6 kcal/mol), and the test set (4.1 kcal/mol). Comparison of the training-set and test-set data for the fixed backbone case with and without triplet CFs (Figure 5-6), shows that fitting for high-energy sequences was poor without triplets. Presumably, these triplets helped to describe cases with clashing residues. The same level of improvement was not seen when looking at the flexible backbone case, Figure 5-7. Here, much of the clashing that could occur was compensated by evaluating over multiple backbones.

Discussion

In this Chapter, I have shown that CE is a viable option to convert flexible-backbone structure-based energies into sequence-based functions. The rmsd for the CE prediction of Bcl-x_L inhibitor and zinc finger design energies was much smaller for the flexible-backbone cases than the fixed-backbone cases. As was shown in Chapter 3, the use of multiple backbones in design calculations can improve the quality of results and the diversity of design sequences.[14-16,18,19] Combining the increased performance of this type of flexible-backbone calculation with the relative speed of the CE function allows for a much more complete search of the sequence space, while retaining the benefits of an accurate structural model.

Computational efficiency

The increase in computational complexity is a major limitation of the use of backbone flexibility in design. This is true for all structural sampling, thus the ability to predict energies from sequence allows for the advantage of increased speed and efficiency. Statistical functions are readily available that allow for efficient sampling when examining a large set of different structures.[20-24] However, these types of methods are limited by poor physical interpretability. Statistical functions rely on the probability of amino acids being near or far to each other in known structures, and not on the precise three-dimensional physical interactions between residues. By using CE, we are able to bridge the gap between the speed of the sequence based functions and the interpretability and increased modeling accuracy of structural models with backbone flexibility.

Variations in design methods

In Chapter 1, I reviewed the wide variety of energy components and energy functions used in protein design. CE is not dependent on the type of energy function used, but the energy function could determine the quality of fit. The previous CE work involved the same energy function used for the zinc finger design.[1] This energy function only contained interaction terms that were pairwise decomposable and could be reduced to sums of interactions between sites of residues. Additionally, it was limited to interactions between pairs of side chains or between the side chains and the template, and did not contain backbone-backbone interactions. Here, I tested the design of Bcl-x_L inhibitors using an energy function that included interactions between all atoms and used non-pairwise decomposable energy terms. These energy terms included a finite difference Poisson-Boltzmann continuum solvent model and a solvent-accessible surface-area

based cavitation term. The performance of the CE of this energy function was quite good for both the fixed- and flexible-backbone cases. This result suggests that other non-pairwise energy functions could be included that account for averages over multiple states, as is the case for the implicit solvent model used here.

In addition to differences in the energy function, the Bcl-x_L inhibitor and zinc finger designs applications both used different methods to include backbone flexibility. The former utilized normal-mode analysis to sample a parameterized space. The later utilized multiple x-ray-structure backbones, which is similar to random sampling in fold space. Both of these showed better CE predictions for the flexible-backbone case than the fixed-backbone. This result is mostly likely due to a decrease in steric clashes. It appears that the type of structural flexibility included in the design was less important than that the increased flexibility permit low energy conformations for some sequences and allowed for a reduction of steric repulsions compared to fixed-backbone models. However, each of these methods still contains only a small amount of structural variation that effectively limits the sampling to a very close local-backbone space. Further study is necessary to see how large a local-backbone space could be explored, and if non-local-backbone sampling could be used to give a reasonable fit by CE.

Higher-order terms

For both the Bcl-x_L inhibitor and zinc-finger design, additional triplets were added to try to improve the overall performance of the CE of the structure based energy. For Bcl-x_L inhibitor design additional triplets did not improve either the fixed- or flexible-backbone case. This observation is reasonable, considering the design positions were spread throughout the helix and as such there were not any strong interactions between sets of three or more design sites. This

same result was seen for the flexible-backbone zinc-finger design, but not for the fixed-backbone. In the fixed-backbone zinc-finger design, there was clear improvement by including some sets of triplets.

The ability of CE to predict the energy associated with a set of amino acids is likely related to how much the background sequence affects their interaction environment. When the backbone structure is fixed, the background sequence can only affect the placement of side-chain rotamers. This placement will most likely be determined by close interactions. When the backbone is flexible, the background sequences may also affect the selection of the backbone. This effect could be distributed throughout the structure. This would suggest that collections of amino acids would be present in more varied environments and thus require more higher-order CFs. However, this does not appear to be the case in the work described here. Rather, the role for the triplets in the fixed-backbone structures appears to be to improve the fit of higher energy sequences, which presumably have a larger contribution from clashing residues. Van der Waals clashes are the hardest part of these functions to predict, given their large range of possible contributions to the energy for pairs of residues. This large range was somewhat accounted for by the triplet terms. In the flexible-backbones case, clashing was relieved by allowing for additional sampling. When the van der Waals clashing terms are relaxed, the overall prediction performance was improved with just the pair terms. Triplet terms might still be useful for capturing the environmental differences associated with flexible backbones, but without a large enough number of sequences in the training set, this can not be determined.

Possible future directions

In validation of the method, I demonstrated that cluster expansion can be used to convert

the structure-based energy used in the Bcl-x_L inhibitor design into a sequence based function. This leads to many different possibilities in terms of design objectives. The first would be to perform design using the complete sequence space. This could include using a faster Monte Carlo search,[1] or could be incorporated for use with linear programming.[25] Using the later method, Grigoryan et al. designed coiled coils with multiple constraints including finding sequences with high stability that have an arbitrary energy gap to other undesired targets.[25] This approach could allow for designing specific interactions to Bcl-2 family members that do not bind to any other targets or are designed to bind several targets at once. The key will be to demonstrate that the design method can in fact find good solutions that bind other receptors and that the interaction energies with different targets are on the same scale.

Methods

Cluster expansion

CE was performed as described by Grigoryan et al.[1,2] Briefly, for this method, the optimal energy $f(\vec{\sigma})$ of a sequence $\vec{\sigma} = [\sigma^1, \sigma^2, \cdots, \sigma^N]$ (here σ^x is the amino acid at each position) can be calculated as the perturbation of the energy of a sequence $\vec{\sigma}_o = [\sigma_o^1, \sigma_o^2, \cdots, \sigma_o^N]$, and is described using the following equation:

$$f(\vec{\sigma}) = \sum_I \sum_Z J_Z^I(\vec{\sigma}_o)\psi_Z^I(\vec{\sigma}_o) \tag{5-1}$$

Here I is a cluster of sites, $\psi_Z^I(\vec{\sigma}_o)$ is the Z^{th} CF associated with cluster I and the coefficients $J_Z^I(\vec{\sigma}_o)$ are known as the ECI associated with $\psi_Z^I(\vec{\sigma}_o)$. For a set of CFs with $\vec{\phi} = [\phi^0 \equiv 1, \phi^1_{\sigma^1=V}, \cdots, \phi^1_{\sigma^1=F}, \cdots, \phi^M_{\sigma^M=S}, \cdots, \phi^M_{\sigma^M=L}]$ being the point basis set at a single amino-acid site,

we can simplify this function. This becomes $\phi_a^I(\vec{\sigma},\vec{\sigma}_o) = \delta(\sigma^I - a)\cdot\left(1-\delta(\sigma_o^I - a)\right)$. Here for I > 0, $\phi_a^I(\vec{\sigma})$ is always 0, except if the amino acid at the Ith site of a given sequence is the same as 'a' and does not equal to the amino acid of original sequence at the site ($\sigma^I \neq \sigma_o^I$), then it is 1. Therefore, for any sequence $\vec{\sigma} = [\sigma^1, \sigma^2, ...\sigma^N]$ the only CFs that remain are of the form $\prod_I \phi_{\sigma^I}^I(\vec{\sigma})$ for all 'I' such that $\sigma^I \neq \sigma_o^I$. Therefore the energy can now be expressed as:

$$\begin{aligned} f(\vec{\sigma}) &= J_0^0(\vec{\sigma}_o) + \sum_I J_{\sigma^I}^I(\vec{\sigma}_o)\phi_{\sigma^I}^I(\vec{\sigma},\vec{\sigma}_o) + \sum_I \sum_{J>I} J_{\sigma^I\sigma^J}^{IJ}(\vec{\sigma}_o)\phi_{\sigma^I}^I(\vec{\sigma},\vec{\sigma}_o)\phi_{\sigma^J}^J(\vec{\sigma},\vec{\sigma}_o)\cdots \\ &= J_0^0(\vec{\sigma}_o) + \sum_{\substack{I \\ \sigma^I \neq \sigma_o^I}} J_{\sigma^I}^I(\vec{\sigma}_o) + \sum_I \sum_{\substack{J>I \\ \sigma^I \neq \sigma_o^I, \sigma^J \neq \sigma_o^J}} J_{\sigma^I\sigma^J}^{IJ}(\vec{\sigma}_o)\cdots \end{aligned} \tag{5-2}$$

Here $J_0^0(\vec{\sigma}_o)$ is the reference energy, $J_{\sigma^I}^I(\vec{\sigma}_o)$ are the point CFs, $J_{\sigma^I\sigma^J}^{IJ}(\vec{\sigma}_o)$ are the pair CFs, $J_{\sigma^I\sigma^J\sigma^K}^{IJK}(\vec{\sigma}_o)$ would be the triplets and so on.

For each CE, the CFs chosen for fitting each energy function included all point and pair terms, along with selected triplet terms. Given a set of training sequences, the ECI values were fit to minimize the error. These were incorporated through a leave-one-out cross validation to ensure that they were not over-trained. When triplets CFs were included, these were identified by finding combinations of point and pair terms that show strong correlations, as indicated by the program *look43*.[26] This program compared the ECI value for the selected pair CF, fit using the entire sequence set, to the ECI value of the same pair CF, fit only using sequences that contain the given point CF. Additionally, the standard error for fit of the new ECI value was also determined. If the change in the ECI value for the pair CF was greater than 5 times the standard error then this triplet CF was considered for addition to the CE. These triplets were then added one at a time and included in the CE if the cv_rmsd decreased. Once all the ECI values were fit, they were then used to predict the energy of test set sequences.

Bcl-x_L inhibitor design

The design of the BH3 peptide was performed as described in Chapter 3 with a few exceptions. Briefly, a CHARMM[27] based energy function that included param19 van der Waals, with 90 % radii, dihedral angle energies, distance dependent dielectric electrostatics ($\varepsilon = 4r$), and EEF1 solvation[28] was used to repack the structure using DEE/A*.[29-34] The Richardson and Richardson rotamer library was used.[35] For each repacked structure, 1000 steps of steepest descent followed by 1000 steps of ABNR were used to minimize all the side chains. The energy for the minimization included all CHARMM molecular mechanics terms and distance dependent dielectric electrostatics with dielectric constant of 1r. The energy of a structure was evaluated using CHARMM 19 van der Waals with 100% radii, Coulombic electrostatics with ε=4, finite difference Poisson-Boltzmann solvation using DelPhi[36,37] with internal dielectric of 4 and external dielectrics of 4 and 80, and a solvent-accessible surface-area based cavitation term with coefficient equal to 10 cal/Å^2.[38] A total of 10,000 sequences were chosen randomly for 11 sites from all residues found in design results in Chapter 3. These are summarized in Table 5-1. The energy for each sequence was evaluated on 16 different backbones, including the native. The lowest energy backbone for each sequence was then used for the flexible-backbone CE. CE was also performed on the native structure. Next, a set of 10,000 additional test-set sequences with less than 100 % sequence identity to the training set were generated from the same random distributions. These sequences were evaluated on the same backbone set as the training set. The minimum energy over all backbones and the energy of the native structure were predicted using the ECI values from the training set CE.

Zinc-finger design

Dahiyat and Mayo previously designed the second zinc finger on chain C of 1zaa.[7] To expand the structure space, 28 other x-ray crystal structures of homologous zinc fingers were extracted from the SCOP database that contained a continuous stretch of 21 residues overlapping with chain C of 1zaa (residues 34 to 54). These included the second and third zinc finger from PDB ids: 1a1f, 1a1g, 1a1h, 1a1i, 1a1j, 1a1k, 1a1l, 1aay, 1jk1, 1jk2, 1llm, 1p47 and 1zaa and the only zinc fingers from PDB ids: 1f2i, 1ubd, and 2drp. The pairwise C_α-rmsd values of all sequences were calculated using the program VMD (version 1.8.3)[39] and then clustered using Matlab.[40] From this set, 10 structures were chosen to sample the different clusters. These included the second zinc finger from 1a1h, 1aay, 1jk1, 1p47, and 1zaa, the third zinc finger from 1a1i, and 1llm and the only zinc finger from 1f2i, 1ubd and 2drp.

Random sequences were generated using the distribution of amino acids described by Grigoryan et al[1] and Mayo et al.[6] Positions 1, 3, 5, 7, 9, 10, 12, 13, 14, 15, 16, 18, and 19 allowed Ala, Ser, Thr, Asp, Asn, Glu, Gln, Lys and Arg. Positions 2, 6, 11, 17, 20 and 21 allowed Ala, Ser, Thr, His, Asp, Asn, Glu, Gln, Lys, Arg, Val, Leu, Ile, Phe, Tyr and Trp. Position 8 only allowed Gly. From this sequence space 50,000 random sequences were evaluated on all 10 zinc finger backbones. These sequences were repacked onto the backbones using the same method as for the Bcl-x_L inhibitor design described above. The repacked sequences were evaluated using the same energy function as described in Grigoryan et al.[1] The energy evaluation included calculating the interactions between different side chains and between the side chains and the backbones. The energy function was composed of CHARMM van der Waals with 100 % radii and Coulombic electrostatics with dielectric of 4,[27] and generalized born screening with PEP used to calculate atomic born radii.[41] From these 50,000 sequences, CE was performed using the

minimum energy value over all backbones. Additionally, for a fixed backbone comparison, CE was performed using the second zinc finger from 1zaa. Next, 50,000 test set sequences were generated using the same distribution and the energy was evaluated on all backbones using the same method. Using the ECI values from the CE of the training set sequences, the energy was predicted for the fixed-backbone and flexible-backbones cases.

References

1. Grigoryan G, Zhou F, Lustig SR, Ceder G, Morgan D, Keating AE. Ultra-fast evaluation of protein energies directly from sequence. PLoS Comput Biol 2006;2(6):e63.
2. Zhou F, Grigoryan G, Lustig SR, Keating AE, Ceder G, Morgan D. Coarse-graining protein energetics in sequence variables. Phys Rev Lett 2005;95(14):148103.
3. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of molecular biology 1987;193(4):775-791.
4. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. Protein Sci 1995;4(10):2006-2018.
5. Lazar GA, Desjarlais JR, Handel TM. De novo design of the hydrophobic core of ubiquitin. Protein Sci 1997;6(6):1167-1178.
6. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82-87.
7. Dahiyat BI, Sarisky CA, Mayo SL. De novo protein design: towards fully automated sequence selection. Journal of molecular biology 1997;273(4):789-796.
8. Malakauskas SM, Mayo SL. Design, structure and stability of a hyperthermophilic protein variant. Nat Struct Biol 1998;5(6):470-475.
9. Bryson JW, Desjarlais JR, Handel TM, DeGrado WF. From coiled coils to small globular proteins: design of a native-like three-helix bundle. Protein Sci 1998;7(6):1404-1414.
10. Hellinga HW. Computational protein engineering. Nat Struct Biol 1998;5(7):525-527.
11. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.
12. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. Protein Science 2000;9:1106-1119.
13. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332(2):449-460.
14. Larson SM, England JL, Desjarlais JR, Pande VS. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. Protein Science 2002;11(12):2804-2813.
15. Saunders CT, Baker D. Recapitulation of protein family divergence using flexible backbone protein design. Journal of molecular biology 2005;346(2):631-644.

16. Wollacott AM, Desjarlais JR. Virtual interaction profiles of proteins. Journal of molecular biology 2001;313(2):317-342.
17. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of molecular biology 1995;247(4):536-540.
18. Harbury PB, Tidor B, Kim PS. Repacking Protein Cores with Backbone Freedom - Structure Prediction for Coiled Coils. Proceedings of the National Academy of Sciences of the United States of America 1995;92(18):8408-8412.
19. Keating AE, Malashkevich VN, Tidor B, Kim PS. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. Proc Natl Acad Sci U S A 2001;98(26):14825-14830.
20. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys J 2003;84(3):1895-1901.
21. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001;44(3):223-232.
22. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11(11):2714-2726.
23. Aloy P, Russell RB. InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 2003;19(1):161-162.
24. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18(3):534-552.
25. Grigoryan G, Reinke A, Keating AE. Computational Design of Globally Specific Anti-bZIP Peptides. Unpublished 2008.
26. Grigoryan G. "Look43: A program to identify significant triplet cluster functions"; 2008.
27. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comp Chem 1983;4(2):187-217.
28. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35(2):133-152.
29. Desmet J, Demaeyer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 1992;356:539-542.
30. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophysical Journal 1994;66(5):1335.
31. Gordon DB, Marshall SA, Mayo SL. Energy functions for protein design. Curr Opin Struct Biol 1999;9(4):509-513.
32. Lasters I, De Maeyer M, Desmet J. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. Protein Engineering 1995;8(8):815-822.
33. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins 1998;33(2):227-239.
34. Pierce NA, Spriet JA, Desmet J, Mayo SL. Conformational splitting: A more powerful criterion for dead-end elimination. J Comput Chem 2000;21:999-1009.
35. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins: Struct Funct Genet 2000;40:389-408.

36. Rocchia W, Alexov E, Honig B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. Journal of Physical Chemistry B 2001;105(28):6507-6514.
37. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. J Comput Chem 2002;23(1):128-137.
38. Hubbard S, Thorton JM. NACCESS, Computer Program 2.1.1 edit.: Department of Biochemistry and Molecular Biology (U. C. L., ed.); 1996.
39. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14(1):33-38, 27-38.
40. Matlab R14: The MathWorks, Inc.; 2005.
41. Beroza P, Fredkin DR. Calculation of amino acid pK(a)s in a protein from a continuum electrostatic model: Method and sensitivity analysis. J Comput Chem 1996;17(10):1229-1244.

Chapter 6

Conclusions

The computational complexity associated with the large sequence space used in protein design and prediction requires efficient sampling of the structure space. In Chapter 2, I described two different methods for capturing the structure space of alpha helices in protein-protein interfaces, and showed how to sample this space to generate novel interfaces. For individual helices, the two lowest-energy modes capture most of the deformation and can be used to sample the native structure space. Chapter 3 described the use of normal-mode backbone sampling to design novel BH3 helices that bound to Bcl-x_L. For coiled coils, I extended Crick parameterization to allow for modeling of both parallel and antiparallel dimers. This method was shown to be a useful tool in generating sets of idealized coiled coils that spanned the native structure space. In Chapter 4, these structure sets were used to predict the binding orientation of 131 coiled-coil sequences with 81% accuracy. Finally, in Chapter 5, I demonstrated that flexible backbone methods were compatible with cluster expansion. Design energies from the structure-based energy function used to evaluate the energy of the Bcl-x_L/Bim interaction and the stability of a zinc-finger protein were converted to sequence-based energies to allow for rapid evaluation of the flexible backbone energy. The flexible backbone methods

showed better fitting by cluster expansion than fitting on a single fixed backbone. These three applications have provided substantial information regarding the benefits and problems of flexible-backbone models, in particular, how they relate to the accuracy of predictions and computational efficiency. These findings are discussed below.

Flexible backbones in design and prediction.

An important aspect of the structural search in protein design is the identification of compatible side-chain conformations in the protein core.[1] This identification is typically done quite successfully when repacking the native sequence on the native structure; standard methods are able to find the correct rotamer ($\chi1$ + $\chi2$ correct) with greater than 80% accuracy.[2,3] This results in reasonable modeling of the wild-type structure. However, when this approach is used to model non-native sequences using the wild-type backbone structure, many amino-acid side chains will not fit due to steric clashes. This problem explains in part why the native structure contains a native sequence bias,[4-6] especially in the core. The bias was demonstrated by the design of Bcl-x_L inhibitors in Chapter 3, where SCADS favored the native sequences when evaluated on the native structure (Figure 3-4). When mutations are made, real proteins overcome this steric complementarily problem through subtle relaxations of the side chains and/or the backbone.[7,8] Expanding the backbone-structure space allows this type of relaxation to occur. It greatly reduces steric clashes as seen by the increased ability to design novel sequences (Figure 3-7) and by the observation that repulsive van der Waals no longer plays as large a role in the prediction (Figure 4-4). Reduction is steric clashes is also the primary reason that sequences evaluated on the best backbone out of a set were more easily fit using cluster expansion. Because van der Waals clashes have the largest variability in magnitude and greatest sensitivity

to precise three-dimensional structure, this energy component has the largest deviation among pairs of amino acids in variable background sequence environments. Relieving this strain through the use of backbone flexibility and structure relaxation lowered this energy term and thus allowed for better fitting of the total energy by cluster expansion.

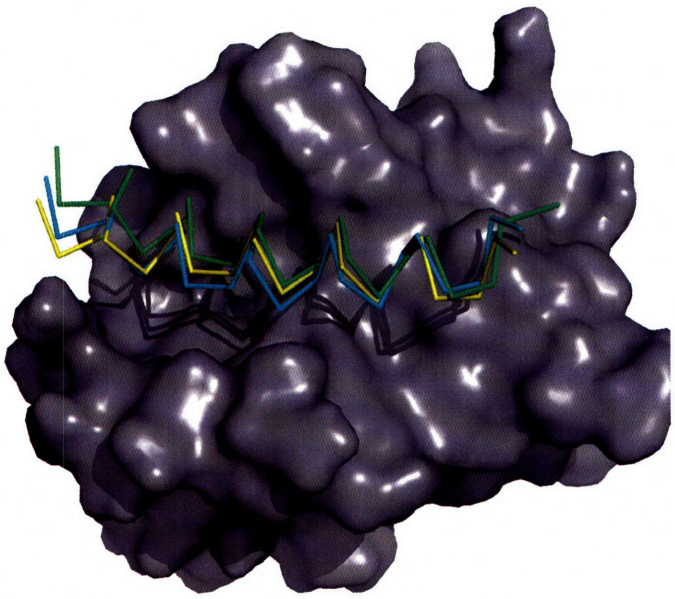
Although there were positive benefits resulting from increased backbone sampling, some problems were also encountered. As described in Chapter 3, some sampling of the backbone, particularly when replacing the native helix with an ideal helix, did not give good predictive results. This was somewhat remedied when the correct native helical pitch was used. Replacing helix minimization with the evaluation of the energy over many backbones gave yet better results (Figure 3-9). This highlights the fact that energy functions, which may be good at differentiating sequences evaluated on similar structures, may not be able to differentiate sequences evaluated on vastly different backbones. It appears that by sampling the normal-mode space and not including backbone minimization we restricted the search to more realistic interfaces, and made it easier to differentiate the relative stabilities of different sequences. Allowing too much unrestrained minimization gave poor results, but some minimization is beneficial, as seen in the prediction of the binding orientation of coiled coils (Figure 4-2a). In that case structures were minimized under constraints on the C_α position derived from the Crick parameterization. This allowed for slight relaxations of the backbone, but not so much as to introduce a large error. Relaxations of this type improved performance over using no minimization of the backbone or free minimization (data not shown). It appears that for both these methods staying in the native-like structure space improves the ability to make accurate predictions.

Another important limitation found with these types of structural sampling methods is that they only sampled global parameters of the helical structure. This was a reasonable first

approximation in order to search over a large sequence space. However, there are examples of native structures where deformation features were not present in any of the backbone sets. This was particularly true for the ends of the coiled-coil structures. Many coiled coils exhibited significant fraying, as seen in several structures with paired Ile residues at **d** positions in parallel coiled coils (Figure 4-5). The C_α-C_α distance of core pairs in the ideal coiled-coil structures ranged from 5 to 7 Å. However, there were native cases where the distance could be as large as 10.5 Å. By not allowing for local deformation, we cannot accurately describe these types of structures.

The structural modeling used for Bcl-x_L inhibitor design also does not sample the entire structure space. Recently Fire et al solved the crystal structure of a single mutant of the Mcl-1/Bim-I6Y complex,[9] which illustrates a structural change that was not present in the set of normal-mode backbones. Figure 6-1a shows native Bim in yellow and Bim-I6Y mutant in green. The rmsd between the BH3 helices in these two structures is 1.70 Å. The normal-mode differences between these two helices show that the largest deviation is in fact one of the two modes that we sampled (Figure 6-1b). Despite this, when these mode vectors are applied to transform the native helix, the resultant helix (blue in Figure 6-1a) still has a 1.44 Å rmsd to the mutant helix. This results from the fact that there are a number of other normal-mode deformations in this helix, along with additional rotations and translations that are not accounted for. Without these other types of sampling, the accuracy of the structural model for this mutant would be limited.

(a)



(b)

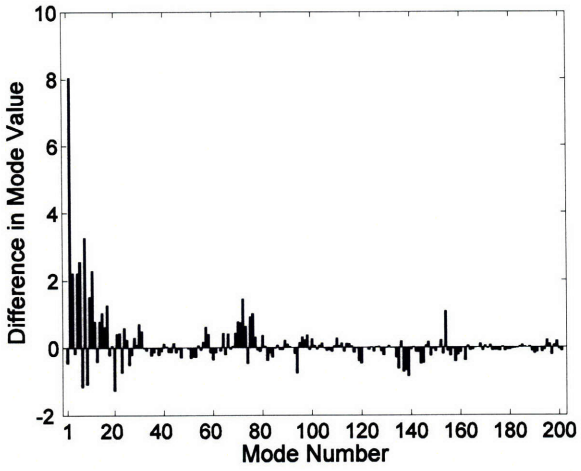


Figure 6-1: Alignment of a native Mcl-1/Bim complex with one involving a mutant Bim. (a) The native Mcl-1 structure from the PDB structure 2nl9 is shown in gray. The Bim helix from PDB structure 2nl9 is shown in yellow. The Bim-I6Y mutant from the structure solved by Fire et al. is shown in green. The native helix fit to the mutant helix by varying the first and second normal mode values is shown in blue. (b) Difference between the normal-mode values for the Bim helix and the mutant BimI6Y helix. The largest difference is seen in mode 2, but other modes also contribute.

Computational efficiency

The computational complexity of protein design methods is an important fact to consider when determining the types of calculations that can be conducted. Many approximations have been made to account for this including using the fixed backbone approximation, limiting the side-chain structure search to a small set of rotamers [10] and employing approximate energy functions that are pairwise decomposable.[11] The applications described in the proceeding chapters all took computational time into account.

Backbone-structural search methods were selected that increased the ability to predict the relative energy of a large number of sequences, while keeping the problems computationally

feasible. Sampling the highly probable parts of structure space using normal-mode analysis and Crick parameterization limited the search space to as few degrees of freedom as possible. Thus these methods were able to search efficiently search a large native-like structure space that improved performance of energy prediction.

As seen in Chapter 1, there are many methods used for backbone-structure searches appropriate for discrete modeling. These methods can be broken down into two categories: global backbone searches and local backbone searches. The global search methods sample the backbone space using parameters that describe the structure of the entire protein. For design this has involved parameterized searches,[12-16] including the normal-mode analysis and Crick parameterization methods described here, as well as ridged-body docking.[17,18] Local backbone searches sample the local changes in backbone-structure space that are independent of changes occurring in other parts of the protein. This type of local search includes fragment library searching,[19,20] random backbone sampling[21,22] and continuous minimization. There are advantages and disadvantages for both types of methods. For the global search, a structurally diverse set of backbones can be sampled, allowing for a much broader range of sequences to be found. However, since the types of backbone deformations are pre-determined, there is no fine-tuning of the backbone space for any particular sequence. Local searches can sample the backbone space much more finely and may find more appropriate backbones for single sequences. However, to sample a diverse set of backbone structures will require much more time than the global search. These two types of searches can be merged together to incorporate the beneficially aspects of both. Global searches could be used to identify parts of structure space particularly suitable for design, and then local searches could be made to fine-tune each of these structures. An example of this is described in Qian et al., where they used principal-component

analysis to define the structure space of a protein family and then used a fragment-based method to search locally in this confined space.[23] This combination of global and local searches was also used in both the Bcl-x_L inhibitor design and coiled-coil orientation prediction applications described in this thesis. In both cases a global parameterization space was searched, and then for each structure local relaxation was allowed through continuous minimization. The global search step was important to increase the diversity of sequences with reasonable structural models, and the minimization step fine-tuned the interaction and allowed for many of the clashes associated with the discrete structural model to be removed.

In addition to the structural search, constraints were imposed on the sequence search. This was addressed in the sequence library used for the design of novel BH3 proteins. For each design experiment, there were 11 variable positions with a total of 5×10^{11} possible combinations of amino acids. To sample quickly over this large sequence space we used the efficient statistical algorithms SCADS[24,25] allowing all sequence combinations to be explored simultaneously. This reduced the sequence library for each backbone to a range of 10^3 to 10^6 for use in the Monte Carlo design phase. Reducing the sequence space enabled the search for individual sequences to use a more accurate energy function.

The final speed-up approach was the use of cluster expansion to convert the structure-based energy function to a sequence-based function. This approach removed the bulk of the computational time for the calculation of each sequence and, as has been demonstrated by Grigoryan et al., can decrease the search time by 7 orders of magnitude.[26] This allows for a much more complete search of the sequence space. Moreover, this additional benefit does not come at the cost of reduced interpretability that is associated with the use of statistical methods.

Summary

I have demonstrated that normal-mode analysis and the Crick parameterization are reasonable methods to sample the global structure space of helices in protein-protein interfaces. Increased structural flexibility has improves the ability to design a widely diverse set of Bcl-x_L inhibitors and to accurately predict the relative stability of parallel and antiparallel orientations of coiled coils. By limiting the search to the highly probable parts of native-structure space, these methods sample structural deformations as efficiently as possible. Finally, I have shown that cluster expansion is a reliable method for fitting a sequence-based energy function to flexible-backbone structural models. This technique will allow for additional sampling of sequence space, despite the increased computational complexity of the structure space. These tools enable the exploration of a wealth of possible helix-mediate interfaces using design and prediction.

References

1. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. Proc Natl Acad Sci U S A 1997;94(19):10172-10177.
2. Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci 2004;13(3):735-751.
3. Grigoryan G, Ochoa A, Keating AE. Computing van der Waals energies in the context of the rotamer approximation. Proteins 2007;68(4):863-878.
4. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. Journal of molecular biology 2003;332(2):449-460.
5. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.
6. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. Protein Science 2000;9:1106-1119.
7. Alber T, Bell JA, Sun DP, Nicholson H, Wozniak JA, Cook S, Matthews BW. Replacements of Pro86 in phage T4 lysozyme extend an alpha-helix but do not alter protein stability. Science 1988;239(4840):631-635.
8. Baldwin EP, Hajiseyedjavadi O, Baase WA, Matthews BW. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. Science 1993;262(5140):1715-1718.

9. Fire E, Keating AE. Crystal structure of mcl-1 bound to Bim ILE to TYR mutant. Unpublished 2007.
10. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of molecular biology 1987;193(4):775-791.
11. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82-87.
12. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282(5393):1462-1467.
13. North B, Summa CM, Ghirlanda G, DeGrado WF. D(n)-symmetrical tertiary templates for the design of tubular proteins. Journal of molecular biology 2001;311(5):1081-1090.
14. Summa CM, Lombardi A, Lewis M, DeGrado WF. Tertiary templates for the design of diiron proteins. Curr Opin Struct Biol 1999;9(4):500-508.
15. Su A, Mayo SL. Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Science 1997;6(8):1701-1707.
16. Offredi F, Dubail F, Kischel P, Sarinski K, Stern AS, Van de Weerdt C, Hoch JC, Prosperi C, Francois JM, Mayo SL, Martial JA. De novo backbone and sequence design of an idealized alpha/beta-barrel protein: evidence of stable tertiary structure. Journal of molecular biology 2003;325(1):163-174.
17. Huang PS, Love JJ, Mayo SL. Adaptation of a fast Fourier transform-based docking algorithm for protein design. J Comput Chem 2005;26(12):1222-1232.
18. Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. Journal of molecular biology 2007;373(2):503-519.
19. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309(5742):1868-1871.
20. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.
21. Wollacott AM, Desjarlais JR. Virtual interaction profiles of proteins. Journal of molecular biology 2001;313(2):317-342.
22. Larson SM, England JL, Desjarlais JR, Pande VS. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. Protein Science 2002;11(12):2804-2813.
23. Qian B, Ortiz AR, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proceedings of the National Academy of Sciences of the United States of America 2004;101(43):15346-15351.
24. Kono H, Saven JG. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. Journal of molecular biology 2001;306(3):607-628.
25. Zou JM, Saven JG. Statistical theory of combinatorial libraries of folding proteins: Energetic discrimination of a target structure. Journal of molecular biology 2000;296(1):281-294.
26. Grigoryan G, Zhou F, Lustig SR, Ceder G, Morgan D, Keating AE. Ultra-fast evaluation of protein energies directly from sequence. PLoS Comput Biol 2006;2(6):e63.

Curriculum Vitae

James R. Apgar

Education

Massachusetts Institute of Technology, Cambridge, MA — *Sept 2002 to May 2008 (expected)*
Candidate for Doctorate of Philosophy Degree in Chemistry
Relevant Course Work:
Chemistry: Quantum Mechanics, Statistical Thermodynamics, Chemical Kinetics
Biology: Introduction to Computational and Systems Biology
Cumulative GPA: 4.8/5

Williams College, Williamstown, MA — *Sept 1997 to June 2001*
Bachelor of Arts degree, Cum Laude with Honors in Chemistry
Relevant Course Work:
Chemistry: Organic I and II, Inorganic, Bioinorganic, Physical I and II, Quantum, Instrumental Methods
Biology: Genetics, Evolution, Biochemistry I and II
Physics: Material Science, Modern Physics, Electricity and Magnetism
Math: Multivariable Calculus, Linear Algebra, Differential Equations, Group Theory
Computer Science: Advanced Data Structures using Java.
Cumulative GPA: 3.6/4; Dean's List 1998, 1999, 2000, 2001

Research Experience

MIT Department of Chemistry, Cambridge MA — *Jan 2003 to present*
Graduate Research in Computational Biophysics with Advisor Prof. Amy Keating

- Computational design of novel helical Bcl-2 ligands
  - Developed normal mode-based strategy for sampling backbone structure and applied this method for use in flexible backbone design.
  - Implemented a two tier design method to generate novel BH3 sequences, many of which have been shown to have strong binding affinity to Bcl-x_L receptors (collaboration w/ X. Fu).
  - Generated an improved binding model and scoring function to better predict experimental binding data.
- Prediction of binding orientation for dimeric coiled coils
  - Extended parallel coiled-coil parameterization method to model antiparallel coiled coils.
  - Developed an effective structure based model to differentiate the binding orientation preference of dimeric coiled coils.
- Cluster expansion of sequences on multiple backbones
  - Evaluated the energies of sequences on ensembles of structural templates.
  - Used cluster expansion to convert structure based energies from multiple backbones to a sequence based function.
  - Showed that the performance of cluster expansion was often improved, as compared to use with a single backbone.

Transform Pharmaceuticals, Lexington, MA — *July 2001 to August 2002*
Research Associate in Solid State Chemistry

- Optimization of the physical forms of drug molecules for improved solubility, stability and bioavailability.
- High throughput crystallization of polymorphs and salts
- Analytical analysis of solid drug forms using powder X-ray diffraction, TGA, DSC and Raman spectroscopy

Williams College Department of Chemistry, Williamstown, MA *Summer*
Research Student in Physical Chemistry with Advisor Prof. John Thoman Jr., *2000 to*
- Determination of overtone spectra of OH and CH bonds using long pathlength absorption spectroscopy and cavity ringdown spectroscopy *Spring 2001*
- Computational prediction of overtone spectra using Gaussian 94/98 and GAMESS

Williams College Department of Chemistry, Williamstown, MA *Summer*
Lab Technician in Biochemistry with Advisor Prof. Deborah Weiss *1999 and*
- Measured cytokine expression levels using radio labeled RT–PCR. *Spring 2000*

Teaching Experience

Teaching Assistant: Thermodynamics and Kinetics, MIT, Cambridge, MA *Spring 2003*
Teaching Assistant: Principles of Chemical Science, MIT, Cambridge, MA *Fall 2002*
Teaching Assistant: Organic Chemistry, Williams College, Williamstown, MA *Fall 1998*

Publications

1. Apgar, J. R., Gutwin, K. N., Keating, A. E., "Prediction of helix orientation for coiled-coil dimers", (2008) *Proteins*, In Press.
2. Fu, X.,[#] Apgar, J. R.,[#] Keating, A. E., "Modeling backbone flexibility to achieve sequence diversity: The design of novel alpha-helical ligands for Bcl-x_L" (2007) *Journal of Molecular Biology, 371, 1099.*

# contributed equally to this work

Meeting Abstracts

1. James R. Apgar, Xiaoran Fu, Karl N. Gutwin, Amy E. Keating "Modeling the flexibility of α-helices in protein interfaces: Structure based design and prediction of helix mediated protein-protein interactions" Computers in Chemistry Poster Session, ACS National Meeting, Boston, August 2007
2. James R. Apgar,[#] Xiaoran Fu,[#] Amy E. Keating, "Modeling backbone flexibility to achieve sequence diversity: The design of novel alpha-helical ligands for Bcl-x_L", Protein Society National Meeting, Boston, July 2007
3. James R. Apgar,[#] Xiaoran F. Stowell,[#] Amy E. Keating, "From Native Structure to Novel Sequences: Computaional Design using Normal Mode Generated Flexible Backbones", Protein Society National Meeting, Boston, July 2005
4. Jason S. Leith, Adam H. Steeves, James R. Apgar, Brian G. Saar, Saroj Bhattarai, John W. Thoman, Jr., "Vibrational overtone transitions of hydrofluoropropanes and ethanes" Physical Chemistry Poster Session, ACS National Meeting, New Orleans, March 2003
5. Sherry L. Morissette, Michael J. Read, Stephen Soukasene, Michael K. Tauber, Lisa A. Scoppettuolo, James R. Apgar, Hector R. Guzman, John-Michael Sauer, David S. Collins, Prabhakar K. Jadhav, Thomas Engler, and Colin G. Gardner. "High throughput crystallization of polymorphs and salts: Applications in early lead optimization" Division of Medicinal Chemistry Poster Session, ACS National Meeting, New Orleans, March 2003

# contributed equally to this work