Computational Prediction of Coiled-coil Interaction Structure Specificity

by

Karl N. Gutwin

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2009

Author .....
Department of Biology
May 22, 2009

Certified by .....
Amy E. Keating
Associate Professor
Thesis Supervisor

Accepted by .....
Tania Baker
Program Director, Committee on Graduate Students

Computational Prediction of Coiled-coil Interaction Structure Specificity

by

Karl N. Gutwin

Submitted to the Department of Biology
on May 22, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The alpha-helical coiled coil is a protein sequence and structural motif that consists of two or more helices in a parallel or antiparallel orientation supercoiling around a central axis. Coiled coils have been observed in a wide range of protein families, and many studies have focused on their sequence and structural diversity over the past half-century. In particular, the observation that coiled coils can be involved in determining protein-protein interactions and protein architectures has prompted the developments of methods to predict the structure of a coiled-coil complex from sequence information alone. In this thesis, I discuss the development of a structurally annotated database of coiled-coil sequence useful for training statistics-based methods of coiled-coil structure prediction. This database was used to retrain and stringently cross-validate the Multicoil method of predicting coiled-coil oligomerization state. In addition, I describe recent work using implicit and explicit structure models to predict dimeric coiled-coil orientation and alignment. Improvements to existing models, insight into coiled-coil structure determinants, and the future of coiled-coil prediction are also discussed.

Thesis Supervisor: Amy E. Keating
Title: Associate Professor

Acknowledgements

This thesis would not have been possible without the help of so many people. First and foremost, I would like to thank my parents, Paul and Sharon, who have encouraged me all the way through over two decades of my education. They have given me the best opportunities possible, and have kept me going when things got tough. I wouldn't have been able to come this far without their tireless support of me and my dreams. I would also like to thank my sisters, Rebecca and Anna, who have given me many good times and many great stories. Finally, I must thank my grandparents, especially Otto and Leona, who have always been an inspiration to me as well as great friends.

My advisor, Amy Keating, has been instrumental to my career here at MIT. She has been an advocate, mentor and friend, and I have always appreciated her persistent encouragement to grow as a scientist. I have learned far more than I ever expected as a member of her lab, and I am grateful to the discussions and insights that she has invested in my work. In addition, each and every member of the Keating lab has been a great friend and colleague, and I have thoroughly enjoyed working with you all. Each of you has been been helpful and kind, and I have benefited from our scientific and personal discussions. In particular, I would like to thank James Apgar for our fruitful collaboration, without whom I would still be searching for successful models. I would also like to thank Dr. William Cutter and the MIT Chamber Chorus for being a welcome distraction from the scientific rigors of MIT.

I would also like to thank my thesis committee members, Tania Baker, Chris Burge, and Jonathan King, for their support and their suggestions. I have particularly appreciated their encouragement and their advice on scientific and career matters.

Finally, I cannot help but thank my wife, Rebecca, for everything. She has been my strongest ally and my best friend. Throughout the trials and triumphs of our life together, she has never failed to encourage and strengthen me. Her spirit of dedication has inspired me to always work harder, and her enjoyment of the simple pleasures of life has enriched my soul. I never would have made it had it not been for her help, with matters as mundane as the dishes and as profound as the questions of life. As we embark on our next adventure, I thank God that we will be working together as partners and friends.

S.D.G.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Protein structure-function relationships and the study of conserved structural motifs

Significant effort in current biomedical research is devoted to understanding how proteins function and malfunction. Proteins are known to have highly specific yet diverse structures, and in each instance, structure and function are closely intertwined. Many significant advancements in molecular biology have been made through studying protein structure, making methods of experimental structural characterization and computational structure prediction important for modern biology[1].

Studies across the large number of solved protein structures have revealed that proteins often contain conserved structural domains[2]. These domains, despite being present in diverse proteins, often share evolutionary history, significant sequence identity, and basic functions. Understanding the structures and functions of commonly recurring

conserved domains is one strategy adopted by protein scientists to broaden our understanding of structure-function relationships. Here, I illustrate this by briefly reviewing a few of the most common and most studied domains along with progress on predicting their occurrences and annotating their functions.

One large class of protein domains is involved in mediating protein-peptide associations. This class includes, among many others, the SH2, SH3 and PDZ domains. SH2 and SH3 domains have been primarily identified in signal transduction processes, in which SH2 domains preferentially bind phosphotyrosine-containing peptides, while SH3 domains bind proline-rich peptides and are often associated with protein kinases[3]. PDZ domains have been found to form scaffolding interactions, also primarily in signal transduction[4]. Because these domains mediate many critical protein-protein interaction networks, much work has been done to characterize the structure and specificity of these domains[5,6]. However, while identifying domains is relatively straightforward using sequence comparison methods, identifying putative interaction peptides and measuring or predicting the interaction specificity of these domains is an area of active research[7,8,9,10].

Another important class of domains are the zinc fingers. Each of these short (20-30 residue) domains coordinates a zinc ion through combinations of cysteine and histidine residues. Although found primarily as specific DNA binding domains in a wide array of transcription factors and other nucleic acid binding proteins[11], zinc fingers have also been identified as protein-lipid and protein-protein interaction domains[12]. Zinc fingers can be recognized by their sequence similarity and are extremely common, being identified in approximately 2% of all human proteins[11]. In their function as DNA

recognition domains, chains of zinc fingers have shown combinatorial specificity for DNA sequence[13]. This combinatorial aspect has important implications for evolutionary mechanisms of DNA-binding specificity and has been exploited in protein engineering [14]. As with the protein-peptide interaction domains, the prediction of DNA-binding specificity is important for understanding transcriptional regulation networks, and has been addressed using machine learning approaches[15].

A highly conserved catalytic structure is that of the protein kinase catalytic domains, which preferentially phosphorylate key serine, threonine or tyrosine residues in their specific substrates. These kinase domains generally share similar structures, which are conserved from yeast to humans[16]. However, despite this domain conservation, kinases are known to have diverse substrate specificity[17]. This specificity plays an important role in determining signal transduction pathways that are critical to all biological processes. Therefore, many diverse approaches to predicting the specificity of kinases and substrates have been developed[17,18,19].

The conserved structural domain that forms the subject of this thesis is the alpha-helical coiled coil. Many proteins, a few of which are highlighted below, have been identified to contain coiled coils, and a recent survey of 22 proteomes using a coiled-coil detection method has estimated that between 2% and 8% of all proteins in any given proteome contain coiled-coil motifs[20]. Unlike other protein domains, however, the function of coiled coils is not always easily inferred from sequence similarity. Many diverse coiled coils share very low levels of sequence similarity, and these have a wide range of functions. Coiled-coil-containing fibrous proteins are involved in cellular architecture, shorter coiled coils act as critical structural and mechanical elements of

globular proteins, and other coiled coils mediate protein-protein interactions to regulate key cellular functions[21]. Because each coiled coil has a function that is influenced strongly by its structure, understanding the functions of novel coiled coils would be significantly enhanced through methods of predicting coiled-coil structure.

In this chapter, I describe some of the history and structure of the coiled coil, how prediction of coiled-coil structure can play an important part in understanding protein function, and how coiled-coil modeling is related to modeling of other types of protein folds. Compared to predicting the structures of proteins generally, there are both significant advantages as well as particular challenges involved in the prediction of coiled-coil structure. Statistical approaches, previously demonstrated for secondary structure prediction, have been successfully used in the prediction of coiled-coil structure. However, coiled-coil databases used for training such methods, particularly those annotated with structural information, have lagged behind the growth of sequence and structure databases. In addition, structural approaches such as fold recognition have been widely used in the prediction of protein structure, and the application of such techniques to the coiled-coil geometry is discussed. Finally, this chapter summarizes the recent advancements in coiled-coil structure prediction as presented in this thesis.

1.2 The history of coiled-coil structure

The coiled coil was first proposed in 1952 by Francis Crick as a solution to the structure of certain fibrous proteins such as keratin[22]. Earlier models and experiments had previously established the alpha helix as a key protein structural element[23];

however, X-ray fiber diffraction data of keratin did not exactly fit the alpha-helix model. Crick's proposal was that the alpha helix, with only minor distortion, could be twisted into a supercoil and pack against other alpha-helices having the same supercoil. This packing was suggested to occur via a "knobs-into-holes" mechanism, where a side chain protruding from one helix (the knob) packs into the space between four side chains from the opposing helix (the hole)[24]. Figure 1-1 illustrates this basic structure. Later analyses of alpha-helix interactions have shown that knobs-into-holes packing is the most common of several classes of possible helix packing modes, including ridges-into-grooves and knobs-onto-knobs[25,26].

Crick and others also noted that a seven-residue periodicity overlaid on the helices involved in a coiled-coil interaction would result in a consistent set of residues present in the core and peripheral positions[24,27]. This "heptad" periodicity, denoted by the letters (**a-b-c-d-e-f-g**), placed hydrophobic residues primarily at the **a** and **d** positions, creating a hydrophobic core that stabilized the interaction. This was first confirmed through sequence analysis of the coiled-coil region of tropomyosin which showed a seven-residue periodicity of hydrophobic residues[28,29].

Crick recognized that this simple combination of knobs-into-holes interaction and heptad repeat could be achieved through the interaction of two, three or more helices together in parallel or antiparallel relative orientations[24]. But it was not until 1963 that tropomyosin was demonstrated to form a dimer[30], and the parallel orientation of myosin was not confirmed until 1967[31]. These and further studies on fibrous proteins led to the idea that the coiled coil was most often long, parallel and dimeric[32]. However, it soon became clear that significant coiled-coil structural diversity could be

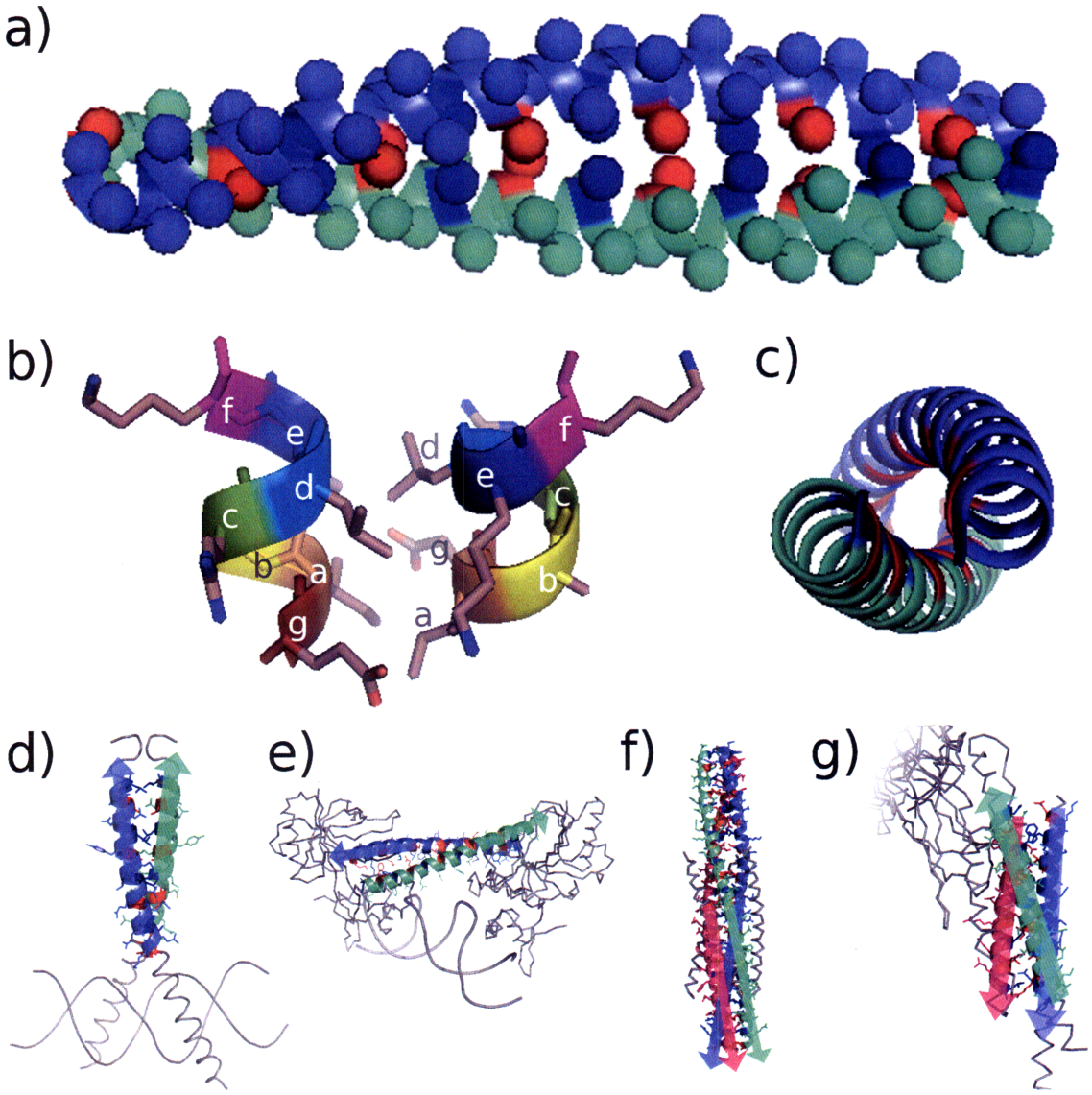


Figure 1-1. Illustration of coiled-coil structure. (a) Side view of a parallel dimeric coiled coil. Spheres denote location of Cβ atoms. Red and blue spheres represent **a** and **d** position residues, respectively. (b) Structure of the heptad repeat in a parallel dimeric configuration. Note hydrophobic interactions of **a-a**' and **d-d**' residues, as well as Glu-Lys **g-e**' charge pair interaction. (c) Top view of parallel dimer, showing consistency of hydrophobic interaction (red and blue segments) along the supercoil. (d-g) Examples of various coiled-coil structures. Colored cartoon region is coiled-coil assigned by SOCKET[33]. Arrows denote N→C sequence polarity. (d) Parallel dimer (e) Antiparallel dimer (f) Parallel trimer (g) Antiparallel trimer)

found in a wide range of protein families. The crystal structure of influenza hemagglutinin was published in 1981, showing a three-stranded coiled coil as the core of this globular protein[34]. ROP, a protein involved in plasmid replication in *E. coli*, was shown to consist entirely of a dimer of antiparallel-associated helices, creating a four-stranded coiled coil[35]. In addition, crystal structures of *E. coli* seryl-tRNA synthetase clearly showed a five-heptad-long antiparallel dimeric coiled coil, possibly involved in binding tRNA[36]. Crystallographic studies of the leucine-zipper-containing protein GCN4 demonstrated that it contained a two-stranded, parallel coiled coil[37], settling a standing debate over the structure of the leucine zipper motif[38].

1.3 Computational prediction of coiled-coil structures

Since the original discovery of the coiled coil, increasing numbers of coiled-coil domains have been found to be relevant for a wide array of protein structures and interactions[39]. For example, the gp41 protein from the HIV virus contains a coiled-coil trimer at its core that is crucial for viral-membrane fusion[40]. The SNAREs are a class of yeast and mammalian membrane fusion proteins that associate as hetero-tetrameric coiled-coil complexes comprised of three distinct proteins[41]. Motor proteins such as myosin and kinesin use coiled coils as rigid rods to transmit force between load and substrate[21]. The alphavirus capsid consists of capsid proteins that are known to dimerize through a coiled-coil domain[42]. This dimerization is crucial for proper capsid assembly, as mutations to the coiled-coil domain that promote trimerization disrupt proper assembly[43]. In all of these examples, the structure of the coiled coil is critical to

the overall activity of the protein. However, over the past several decades, the growth of sequence databases has significantly outpaced that of structure databases[44]. There remain many coiled-coil-containing proteins with unsolved structures and important functions. For example, the yeast spindle pole body is hypothesized to contain many coiled coils[45], some of which have already been implicated in determining the architecture of the large complex[46]. Therefore, computational methods that could predict the structure of a coiled coil from its sequence alone would significantly enhance our understanding of many aspects of biology.

The prediction of coiled-coil structure is a subproblem of the general protein structure prediction problem. Many approaches have been devised to predict protein structure generally. The first methods focused solely on predicting secondary structure through simple statistical models[47,48]. Later, the concept of "fold recognition" was introduced as an inverse protein folding problem; that is, the problem of finding a sequence compatible with an observed fold[49,50]. Such approaches were the first reasonably successful general protein structure prediction methods, but they can only make accurate predictions for sequences that adopt a previously crystallized fold. Despite this constraint, the consistent growth in protein structure databases, as well as developments in structure evaluation potentials, have contributed to the continuing success of fold recognition[51].

For structures without appropriate templates, modern advancements in structure sampling algorithms, scoring functions and computational power have enabled the development of *ab initio* or "free modeling" prediction methods that do not require a complete template model[52,53]. These approaches have shown great promise on certain

prediction problems, such as high-resolution structure prediction[54,55,56], improving homology models[57], and flexible-backbone protein-protein docking[58]. However, such methods are still relatively unreliable, computationally expensive and not yet available for genome-scale applications[51].

Given the many advancements in general protein structure prediction, it would be natural to conclude that such approaches may be directly useful for predicting the structure of the coiled coil. While this is a possibility, it is more likely that methods specifically designed to utilize features of the coiled coil, such as the regular supercoil and the heptad repeat, will reduce the need for crystal-based templates or extensive structural sampling, and may be more accurate in their predictions. However, there are unique challenges to predicting coiled-coil structure, such as fewer topological restraints when predicting coiled-coil interactions, and many candidate template structures with very similar energetics, as discussed in Section 1.6. These challenges have hindered the development of methods that are able to completely predict coiled-coil structure. However, much work has been done to address various subproblems of this goal, which is summarized below.

1.4 Statistical models for coiled-coil structure prediction

One of the key challenges in coiled-coil structure prediction is identifying regions of sequence that have the propensity to form coiled coils. This is similar to approaches that assign secondary structure propensity to un-annotated sequence. Chou and Fasman first developed such a method, which was trained using statistical information derived

from crystallographic observations[47]. This was later improved through the GOR method, which considers pairwise residue information[48]. While these initial methods were not highly accurate, modern approaches can achieve up to 70-80% prediction accuracy through sequence-profile information and neural network approaches[59,60].

The first practical method of coiled-coil propensity prediction was suggested by Parry, who proposed an algorithm based on the geometrical average of heptad-position-specific residue frequencies determined from a small collection of diverse coiled-coil sequences[61]. This approach was extended by Lupas et al., who suggested the use of a maximum-over-window function to smooth scores and calibrated the results to a large database of protein sequences in order to estimate the probability that a certain sequence corresponds to a coiled-coil structure[62]. The Lupas approach, as originally captured in the program COILS, is highly effective in many situations, but it is vulnerable to false positives. The Paircoil algorithm extends the heptad-position-specific approach to include pairwise heptad positions, similar to how the GOR method considers the effect of neighboring residues on secondary structure propensity[48]. Paircoil was shown to be more specific and yields fewer false positives[63]; however, the corresponding drastic increase in parameter space leaves open the question of whether or not the available sequence information is sufficient for training[64].

The success of the Paircoil method inspired several extensions of its approach to different problems in coiled-coil structure prediction. The Learncoil method[65], which applies Paircoil in an iterative training process to improve detection of under-represented families, was successfully used to predict coiled coils found in viral membrane fusion proteins [66] and histidine kinases[67]. In addition, the recognition that known coiled-

coil dimers and trimers exhibit significantly different residue preferences[68] prompted development of the Multicoil method[69], which uses pairwise residue correlations from a large dimer and relatively small trimer database to predict the propensity of sequences to form one of those two structures. Finally, recognition of the significant growth in protein sequence databases prompted updating the original Paircoil training set, resulting in Paircoil2[70] which showed improved performance.

One aspect of all previously mentioned approaches is that they rely on a window-based method to smooth scores for more accurate predictions. The Marcoil method does not use such a window; instead, it is based on a hidden Markov model that uses residue state transition probabilities to model the coiled-coil region[71]. Eliminating the window makes it possible to predict much shorter coiled coils. A recent review of the above methods suggests that while performance has improved significantly from the original, simple approaches, it is still not possible to recognize all structurally confirmed coiled coils with high confidence[64].

1.5 Coiled-coil sequence databases

All existing coiled-coil detection methods require a database of known coiled-coil sequence for training purposes, and all except Marcoil require accurate heptad annotations. The size and composition of the training databases plays an important role in the ability of any given method to detect both known and previously unknown coiled-coil sequence. As seen in Figure 1-2, the amount of sequence annotated with coiled-coil structure initially grew to match the pace of growth in all known genomic sequence;

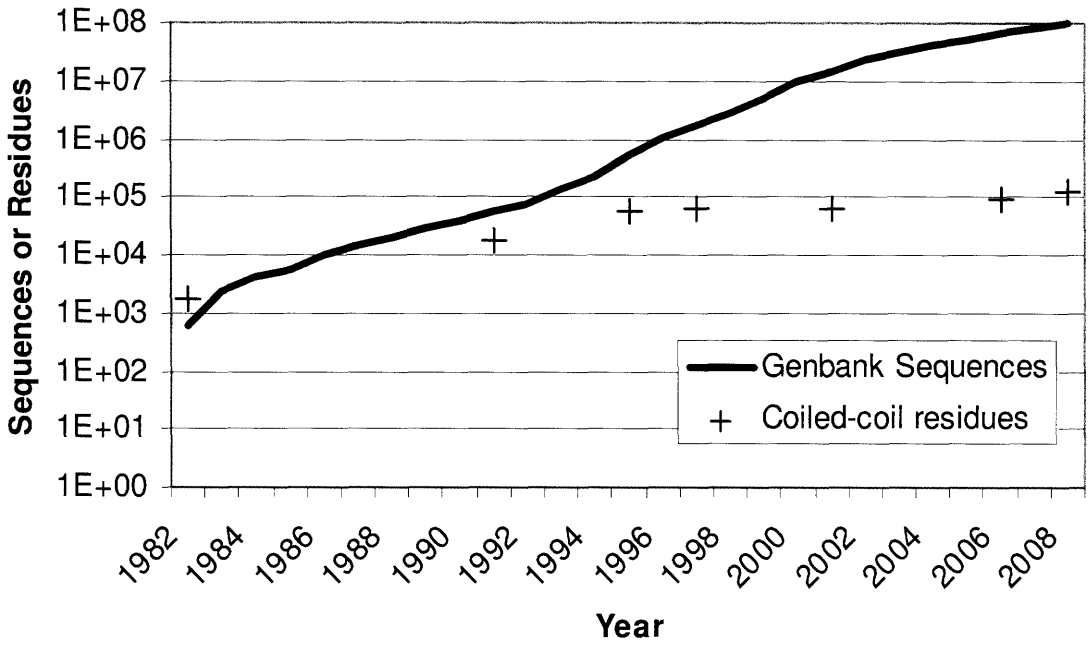


Figure 1-2. Growth of known genomic and coiled-coil sequence databases. Genbank sequence size from [73]. Coiled coil database sizes from [61,62,63,69,70,74] and Chapter 2.

however, after a certain amount of time, that growth slowed significantly. This is likely due to the exponential growth in sequence information overwhelming the predominantly manual approaches used for curating such databases. The vast majority of sequences in the original databases were collected from long fibrous coiled coils such as myosin, tropomyosin and keratin[62]. Recently, the development of the SOCKET algorithm[33], which detects the characteristic knobs-into-holes packing of coiled coils in PDB structures, has opened the possibility of including many previously poorly annotated short coiled coils in sequence databases[72]. However, the relatively small size of the protein structure database has minimized the impact of such tools on general coiled-coil sequence databases. Currently, the most common way to obtain new coiled-coil sequence

is to use homology searches from existing annotated sequence and to individually discover and characterize new families using experimental data[70].

Predicting additional features of the coiled coil, such as dimerization versus trimerization preference, requires databases that are annotated beyond the heptad repeat. Early coiled-coil sequence databases consisted of sequences with uniform (parallel two-stranded) structure. Subsequently, Woolfson and Alber studied a small database of two- and three-stranded coiled coils (~2000 residues each) and identified patterns of residue preferences that could distinguish between these two sets[68]. These successes prompted the development of the Multicoil method and database[69], which contained 6,319 three-stranded coiled-coil residues along with the 58,191-residue two-stranded database created for Paircoil[63]. However, as discussed in Chapter 2, this three-stranded database does not capture enough diversity to accurately predict some coiled coils. The primary aim of the work described in Chapter 2 was to increase these database sizes and characterize performance of the Multicoil method under more rigorous validation standards than were possible when the method was first developed.

1.6 Structural models for coiled-coil prediction

The previously described prediction methods, with the exception of Multicoil, only predict the propensity of a sequence to form generic coiled-coil structure. However, as discussed above, coiled coils have been found to adopt a wide variety of structures[75], including variations in number of helices, helix orientation, alignment and partnering preference (Figure 1-1). Because protein sequence encodes structure, given the

proper model, it should be possible to predict structure from sequence. However, in some cases, closely related sequences have been observed to form different structures, and the energetic threshold between these structures can be low. For example, studies of point mutations in the GCN4 leucine zipper (bZIP) domain showed variation not only in interaction stability but in helix number, helix orientation and partnering preference[76,77]. A solvent-exposed mutation in a GCN4 variant was shown to specify the formation of parallel or antiparallel tetramers[78]. Studies of a model parallel dimeric coiled coil demonstrated the formation of antiparallel dimers simply by moving the position of one core asparagine residue[79], and further investigation showed the overall contribution of the resulting core polar interaction was roughly equivalent to that of a single interhelical electrostatic interaction, approximately 2.1 kcal/mol[80]. Therefore, in order to accurately predict what type of structure a sequence will form, accurate structural models that consider the entire complex are likely necessary.

1.6.1 Simple implicit structure models

One of the earliest innovations in the prediction of coiled-coil structure was the recognition that a key set of interactions between specific positions of the heptad (summarized in Table 1-1) are most important in determining structural preference[81]. This led to the development of a simple "charge-patterning" model where charge-charge interactions between **g** and **e'** residues were scored according to their complementarity[82]. A similar charge-patterning model was successful at predicting the hetero-trimerization of laminins[83]. This rational approach of distilling complex

Name	Parallel positions	Antiparallel positions	Predominant environment
Core	**a-a'**, **d-d'**	**a-d'**	hydrophobic
Edge	**g-e'**	**e-e'**, **g-g'**	polar, charged
Core-to-edge	**g-a'**, **d-e'**	**a-e'**, **g-d'**	all types
Vertical	**a-d'**, **d-a'**	**a-a'**, **d-d'**	hydrophobic

Table 1-1. Key coiled-coil interactions. Interacting positions (such as **g-e'**) denote an interaction between a **g** position residue on one helix and an **e** position residue on the other (prime) helix.

interactions into simple integer scores was later extended to include simple core patterning terms, and was successful at predicting the association preference of a set of bZIP proteins[84]. A similar rational model which considered charge and core patterning along with a helix-propensity term was recently shown to correlate well with a set of melting temperatures of bZIPs[85]. However, these types of models are extremely low-resolution and tend to assign the same interaction weights to a group of related residues, even when experiments have suggested that this is not strictly appropriate.

In order to refine such simplistic association models, the technique of double-mutant thermodynamic cycle analysis was used. Experimental coupling energies between residues commonly found at both **e-g'**[86,87] and **a-a'**[88,89] heptad pair positions (Table 1-1) were measured in a model system derived from the avian bZIP coiled coil VBP. A small set of antiparallel coupling energies have also been derived using synthetic peptides[90,91]. The advantage of these coupling energies is that the thermodynamic cycle allows for isolation and quantification of the interaction between a specific pair of residues[92]. On a test of predicting a set of known bZIP associations, coupling energies were shown to consistently perform better than a set of simple charge-patterning weights[93]. While this approach appears to provide useful data, its major disadvantage is

the large number of experiments necessary to create and characterize the mutant proteins. This has limited the number of residue pairs for which data is available.

Due to the significant amount of coiled-coil sequence available, machine learning approaches are an attractive alternative to extensive experimental characterization. Singh and Kim used a support vector machine (SVM) approach to determine residue pair weights for core, edge and core-to-edge interactions (Table 1-1, [74]). The major advantage of the SVM framework is that it allows for learning from heterogeneous data derived from both sequence and experimental observations. The resulting weights were shown to give good performance in predicting dimeric coiled-coil alignment, as well as partner prediction among both the keratin[74] and bZIP families[93]. However, the resulting weights are not always physically interpretable, as they are optimized only to provide the greatest separation between "interacting" and "non-interacting" datasets.

Despite their diversity of construction, each of the previously described methods shares a common trait: pairwise residue interactions used in evaluating models are predefined according to knowledge about the structural relationships in canonical coiled coils. These interactions are assumed to be independent at the pairwise level, and their contributions to the stability of the complex are assumed to be additive. Therefore, stability can be calculated simply by summing the scores assigned to all relevant interactions. This class of models effectively has structure information "baked in" in an implicit form, hence the name "implicit structure model" (ISM).

1.6.2 Statistical contact potential-based implicit structure models

As mentioned previously, fold recognition techniques have been widely used in the prediction of protein structures. The basic process of fold recognition involves collecting a set of structural templates, then "threading" a sequence to be predicted onto those templates using a sequence/structure alignment protocol[94]. A key distinction among fold recognition methods is the scoring function used to evaluate candidate structural solutions. One common class of scoring functions is the statistical residue-based contact potentials. Potentials such as these are constructed from the frequencies of pairwise residue "contacts" (residue-residue distances below a given cutoff) in a large database of protein structures, that are normalized to generate additive scores[95]. Other, more detailed functions have been developed and tested, and have shown improved performance at the expense of computational complexity[96].

The growing number of protein-protein interactions observed in structural data has enabled the development of multimeric threading approaches that predict both protein tertiary and quaternary structure. The MULTIPROSPECTOR method considers as templates protein pairs that are observed to interact; candidate sequences are threaded first onto individual components of the complex and then onto the complex itself. At the complex stage, models are scored using a statistical contact potential specifically derived from observed protein-protein interfaces. This method, along with others, has been shown to recapitulate pieces of the protein interaction networks observed through experimental data[97,98].

This work on structural prediction of protein-protein interactions using fold-recognition methods was used as a starting point for predicting coiled-coil association, as discussed in Chapters 3 and 4. We developed a statistical contact potential, known as RISP, that is similar to those discussed above but is derived exclusively from heterotypic protein-protein interfaces. While we first tested an approach where interhelical contacts were defined using distance constraints derived from 3D models, as is common in fold recognition, we found that the implicit structure modeling framework used in conjunction with residue contact potentials was not only less computationally intensive but also more accurate. Implicit structure models in the form of position-specific scoring matrices derived from statistical potentials have been used previously to predict SH2-phosphopeptide associations[8]. Unfortunately, the major drawback to such approaches is the effort needed to develop new models for each possible structural variant. This process is not always straightforward, as discussed in Chapter 4, where we observed that different structures and even different sequence families are predicted best with different pairwise heptad interactions. In addition, the representation of residue interactions as a single value approximates the "true" potential, which can include multi-body effects such as rotamer selection, packing strain and desolvation[99].

1.6.3 Explicit structure models

In recognition of the limitations of the implicit-structure approach, an alternative class of structure modeling considers the full 3D representation of the interaction at the atomic level. This detailed model allows for consideration of multi-body effects, as well

as having much finer granularity of interatomic distances. Such models were originally developed for detailed modeling and refinement of protein structures[100,101], and modern high-resolution scoring functions form the basis of the *ab initio* structure prediction methods described above[102]. These potentials are applicable to the coiled-coil structure as long as a suitable structural modeling framework is used.

One challenge with traditional fold recognition approaches is that coiled coils have significant diversity in their backbone structures, not only in gross architectural parameters but also in fine details such as superhelical pitch, radius and interhelical displacement[103]. Instead of the fragment library approach used by methods such as ROSETTA[54], the mathematical description of the coiled coil originally proposed by Crick prompted the development of parameterized models of coiled-coil backbones. These form the basis of most current coiled-coil explicit structure models. The first example of this work was in the prediction of the coiled-coil structure of GCN4 to within 1.75 Å, several months before the crystal structure was published[104,105]. Harbury and coworkers used structure-energy minimization techniques to predict the rotamer preferences of GCN4[106], as well as to design a coiled coil with a slight right-handed superhelical twist[107]. Keating et al. later extended this technique to predict the association preferences of core-position mutants in designed parallel heterodimers[108].

In contrast to the above approaches, which use parameterized constraints during structure minimization to refine each modeled structure, Grigoryan and Keating modeled a large number of native bZIP sequences on a fixed coiled-coil backbone[109]. This approach, after correcting for concerns with the unfolded reference state and poorly

modeled core-position interactions, was shown to be superior to implicit-structure models derived from coupling energies[109].

In Chapters 3 and 4 of this thesis, we have investigated the combination of modeling backbone flexibility using parameterized backbones along with evaluating the resulting structures with a variety of previously published energy functions. Our results indicate that careful modeling can elucidate some of the determinants of structural specificity; however, issues with structural sampling as well as reference state models can diminish performance significantly.

While explicit structure models show some predictive ability, these models suffer from several important limitations. First, the computational time necessary to build and evaluate such detailed models makes this approach prohibitive for genome-scale prediction. Computational time constraints also limit the amount of structural sampling necessary to achieve an optimal model; several studies have demonstrated the detrimental effect of even small inaccuracies in an evaluated structure[110,111]. However, cluster expansion methods[112,113], which can be used to derive a set of sequence-based weights from explicitly computed energies, effectively build a bridge between the explicit and implicit structure models, making it possible to efficiently use explicit structure models in computationally intensive tasks. For example, this method has recently proven highly useful for designing novel heterospecific inhibitors of bZIP interactions[114].

1.7 Summary of thesis work

Progress has been made in the field of coiled-coil structure prediction, yet several major issues remain. This thesis describes our recent work in updating statistical coiled-coil prediction methods using new sequences, introducing more stringent tests of method performance and exploring the behavior of a range of both implicit and explicit structure models for two important coiled-coil structure prediction problems. Dramatic growth of protein sequence and structure databases prompted us to develop updated coiled-coil sequence databases. These databases, as described in Chapter 2, are organized by structure and useful for training and validating prediction methods. The significantly larger amount of available sequence expands the types of tests that can be used for method validation. In addition, two major aspects of coiled-coil structure – helix orientation and helix alignment – have yet to be comprehensively treated. Chapters 3 and 4 address these two problems, respectively, and show that useful prediction performance can be achieved using both simple implicit structure models as well as detailed explicit structure models. Finally, Chapter 5 discusses the progress made in coiled-coil structure prediction and proposes future approaches towards the ultimate goal of a unified, accurate coiled-coil structure prediction framework.

1.8 References

1. Cohen C (2007) Seeing and Knowing in Structural Biology. J Biol Chem 282: 32529-32538.
2. Kohn W, Mant C, Hodges R (1997) alpha -Helical Protein Assembly Motifs. J Biol Chem 272: 2583-2586.
3. Engen JR, Wales TE, Hochrein JM, Meyn MA, Banu Ozkan S, et al. (2008) Structure and dynamic regulation of Src-family kinases. Cellular and molecular life sciences : CMLS 65: 3058-3073.
4. Ranganathan R, Ross E (1997) PDZ domain proteins: Scaffolds for signaling complexes. Current Biology 7: R770-R773.
5. Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, et al. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. Journal of molecular biology 343: 703-718.
6. Tonikian R, Zhang Y, Sazinsky S, Currell B, Yeh J-H, et al. (2008) A Specificity Map for the PDZ Domain Family. PLoS Biology 6: e239.
7. Li L, Wu C, Huang H, Zhang K, Gan J, et al. (2008) Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. Nucleic acids research 36: 3263-3273.
8. Sánchez IE, Beltrao P, Stricher F, Schymkowitz J, Ferkinghoff-Borg J, et al. (2008) Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. PLoS computational biology 4.
9. Hou T, Xu Z, Zhang W, McLaughlin W, Case D, et al. (2008) Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. Mol Cell Proteomics: M800450-MCP800200.
10. Fernandez-Ballester G, Beltrao P, Gonzalez JM, Song Y-H, Wilmanns M, et al. (2009) Structure Based Prediction of the S. cerevisiae SH3-Ligand Interactions. Journal of molecular biology.
11. Matthews JM, Sunde M (2002) Zinc fingers--folds for many occasions. IUBMB life 54: 351-355.
12. Brayer KJ, Segal DJ (2008) Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. Cell biochemistry and biophysics 50: 111-131.
13. Choo Y (1997) Physical basis of a protein-DNA recognition code. Current Opinion in Structural Biology 7: 117-125.
14. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, et al. (1997) Gene families: the taxonomy of protein paralogs and chimeras. Science 278: 609-614.
15. Persikov AV, Osada R, Singh M (2008) Predicting DNA recognition by Cys2His2 zinc finger proteins. Bioinformatics (Oxford, England).
16. Hanks SK, Quinn AM, Hunter T (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 241: 42-52.

17. Kobe B, Kampmann T, Forwood JK, Listwan P, Brinkworth RI (2005) Substrate specificity of protein kinases and computational prediction of substrates. Biochim Biophys Acta 1754: 200-209.
18. Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, et al. (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic acids research 35.
19. Miller ML, Blom N (2009) Kinase-specific prediction of protein phosphorylation sites. Methods in molecular biology (Clifton, NJ) 527.
20. Rose A, Schraegle SJ, Stahlberg EA, Meier I (2005) Coiled-coil protein composition of 22 proteomes--differences and common themes in subcellular infrastructure and traffic control. BMC evolutionary biology 5.
21. Rose A, Meier I (2004) Scaffolds, levers, rods and springs: diverse cellular functions of long coiled-coil proteins. Cellular and molecular life sciences : CMLS 61: 1996-2009.
22. Crick FH (1952) Is alpha-keratin a coiled coil? Nature 170: 882-883.
23. Pauling L, Corey R, Branson HR (1951) The Structure of Proteins. Proceedings of the National Academy of Sciences of the United States of America 37: 205-211.
24. Crick FHC (1953) The packing of alpha-helices: simple coiled-coils. Acta Crystallographica 6: 689-697.
25. Chothia C, Levitt M, Richardson D (1977) Structure of proteins: packing of alpha-helices and pleated sheets. Proc Natl Acad Sci U S A 74: 4130-4134.
26. Walther D, Eisenhaber F, Argos P (1996) Principles of helix-helix packing in proteins: the helical lattice superposition model. J Mol Biol 255: 536-553.
27. Pauling L, Corey R (1953) Compound Helical Configurations of Polypeptide Chains: Structure of Proteins of the [alpha]-Keratin Type. Nature 171: 59-61.
28. Parry DA (1975) Analysis of the primary sequence of alpha-tropomyosin from rabbit skeletal muscle. Journal of molecular biology 98: 519-535.
29. Hodges RS, Sodek J, Smillie LB, Jurasek L (1973) Tropomyosin: Amino Acid Sequence and Coiled-Coil Structure. Cold Spring Harbor Symposia on Quantitative Biology 37: 299-310.
30. Cohen C, Holmes KC (1963) X-ray diffraction evidence for alpha-helical coiled-coils in native muscle. Journal of molecular biology 6: 423-432.
31. Slayter H, Lowey S (1967) Substructure of the Myosin Molecule as Visualized by Electron Microscopy. Proceedings of the National Academy of Sciences of the United States of America 58: 1611-1618.
32. Caspar DL, Cohen C, Longley W (1969) Tropomyosin: crystal structure, polymorphism and molecular interactions. Journal of molecular biology 41: 87-107.
33. Walshaw J, Woolfson DN (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J Mol Biol 307: 1427-1450.
34. Wilson IA, Skehel JJ, Wiley DC (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 [angst] resolution. Nature 289: 366-373.
35. Banner D (1987) Structure of the ColE1 Rop protein at 1.7 [angs] resolution. Journal of Molecular Biology 196: 657-675.

36. Cusack S, Berthet-Colominas C, Härtlein M, Nassar N, Leberman R (1990) A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 A. Nature 347: 249-255.
37. O'Shea EK, Klemm JD, Kim PS, Alber T (1991) X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science 254: 539-544.
38. Landschulz W, Johnson P, McKnight S (1988) The Leucine Zipper: A Hypothetical Structure Common to a New Class of DNA Binding Proteins. Science 240: 1759-1764.
39. Grigoryan G, Keating A (2008) Structural specificity in coiled-coil interactions. Current Opinion in Structural Biology 18: 477-483.
40. Chan D, Fass D, Berger J, Kim P (1997) Core Structure of gp41 from the HIV Envelope Glycoprotein. Cell 89: 263-273.
41. Sutton B, Fasshauer D, Jahn R, Brunger A (1998) Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. Nature 395: 347-353.
42. Perera R, Owen K, Tellinghuisen T, Gorbalenya A, Kuhn R (2001) Alphavirus Nucleocapsid Protein Contains a Putative Coiled Coil {alpha}-Helix Important for Core Assembly. J Virol 75: 1-10.
43. Perera R, Navaratnarajah C, Kuhn RJ (2003) A heterologous coiled coil can substitute for helix I of the Sindbis virus capsid protein. Journal of virology 77: 8345-8353.
44. Westbrook J, Feng Z, Chen L, Yang H, Berman HM (2003) The Protein Data Bank and structural genomics. Nucleic Acids Res 31: 489-491.
45. Zizlsperger N, Malashkevich VN, Pillay S, Keating AE (2008) Analysis of coiled-coil interactions between core proteins of the spindle pole body. Biochemistry 47: 11858-11868.
46. Kilmartin JV, Dyos SL, Kershaw D, Finch JT (1993) A spacer protein in the Saccharomyces cerevisiae spindle poly body whose transcript is cell cycle-regulated. The Journal of cell biology 123: 1175-1184.
47. Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13: 222-245.
48. Garnier J, Osguthorpe D, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology 120: 97-120.
49. Ponder J, Richards F (1987) Tertiary templates for proteins *1Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of Molecular Biology 193: 775-791.
50. Bowie JU, Lüthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164-170.
51. Zhang Y (2008) Progress and challenges in protein structure prediction. Current Opinion in Structural Biology 18: 342-348.
52. Das R, Baker D (2008) Macromolecular modeling with rosetta. Annual review of biochemistry 77: 363-382.
53. Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biology 5: 17.
54. Bradley P, Misura K, Baker D (2005) Toward High-Resolution de Novo Structure Prediction for Small Proteins. Science 309: 1868-1871.

55. Das R, Qian B, Raman S, Vernon R, Thompson J, et al. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 69: 118-128.
56. Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins: Structure, Function, and Bioinformatics 69: 108-117.
57. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci U S A 103: 5361-5366.
58. Wang C, Bradley P, Baker D (2007) Protein-Protein Docking with Backbone Flexibility. Journal of Molecular Biology 373: 503-519.
59. Rost B, Sander C (1993) Prediction of Protein Secondary Structure at Better than 70% Accuracy. Journal of Molecular Biology 232: 584-599.
60. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202.
61. Parry DA (1982) Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. Bioscience reports 2: 1017-1024.
62. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science (New York, NY) 252: 1162-1164.
63. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, et al. (1995) Predicting coiled coils by use of pairwise residue correlations. Proceedings of the National Academy of Sciences of the United States of America 92: 8259-8263.
64. Gruber M, Söding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. J Struct Biol 155: 140-145.
65. Berger B, Singh M (1997) An iterative method for improved protein structural motif recognition. Journal of computational biology : a journal of computational molecular cell biology 4: 261-273.
66. Singh M, Berger B, Kim PS (1999) LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins. Journal of molecular biology 290: 1031-1041.
67. Singh M, Berger B, Kim P, Berger J, Cochran A (1998) Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. Proceedings of the National Academy of Sciences of the United States of America 95: 2738-2743.
68. Woolfson DN, Alber T (1995) Predicting oligomerization states of coiled coils. Protein science : a publication of the Protein Society 4: 1596-1607.
69. Wolf E, Kim PS, Berger B (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein science : a publication of the Protein Society 6: 1179-1189.
70. McDonnell AV, Jiang T, Keating AE, Berger B (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 22: 356-358.
71. Delorenzi M, Speed T (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18: 617-625.
72. Testa OD, Moutevelis E, Woolfson DN (2008) CC+: a relational database of coiled-coil structures. Nucleic acids research.

73. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. Nucleic acids research 37.
74. Singh M, Kim P. Towards predicting coiled-coil protein interactions; 2001. ACM. pp. 279-286.
75. Moutevelis E, Woolfson D (2009) A Periodic Table of Coiled-Coil Protein Structures. Journal of Molecular Biology 385: 726-732.
76. Harbury PB, Zhang T, Kim PS, Alber T (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science 262: 1401-1407.
77. Zeng X, Herndon A, Hu J (1997) Buried Asparagines Determine the Dimerization Specificities of Leucine Zipper Mutants. Proceedings of the National Academy of Sciences of the United States of America 94: 3673-3678.
78. Yadav M, Leman L, Price D, Brooks C, Stout D, et al. (2006) Coiled Coils at the Edge of Configurational Heterogeneity. Structural Analyses of Parallel and Antiparallel Homotetrameric Coiled Coils Reveal Configurational Sensitivity to a Single Solvent-Exposed Amino Acid Substitution. Biochemistry 45: 4463-4473.
79. Oakley MG, Kim PS (1998) A buried polar interaction can direct the relative orientation of helices in a coiled coil. Biochemistry 37: 12603-12610.
80. McClain DL, Binfet JP, Oakley MG (2001) Evaluation of the energetic contribution of interhelical Coulombic interactions for coiled coil helix orientation specificity. J Mol Biol 313: 371-383.
81. Cohen C, Parry D (1990) alpha-Helical coiled coils and bundles: How to design an alpha-helical protein. Proteins: Structure, Function, and Genetics 7: 1-15.
82. Vinson CR, Hai T, Boyd SM (1993) Dimerization specificity of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design. Genes & development 7: 1047-1058.
83. Beck K, Dixon TW, Engel J, Parry DA (1993) Ionic Interactions in the Coiled-coil Domain of Laminin Determine the Specificity of Chain Assembly. Journal of Molecular Biology 231: 311-323.
84. Vinson C, Myakishev M, Acharya A, Mir AA, Moll JR, et al. (2002) Classification of human B-ZIP proteins based on dimerization properties. Mol Cell Biol 22: 6321-6335.
85. Mason JM, Schmitz MA, Müller KM, Arndt KM (2006) 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 103: 8989-8994.
86. Krylov D, Mikhailenko I, Vinson C (1994) A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. The EMBO journal 13: 2849-2861.
87. Krylov D, Barchi J, Vinson C (1998) Inter-helical interactions in the leucine zipper coiled coil dimer: pH and salt dependence of coupling energy between charged amino acids. J Mol Biol 279: 959-972.
88. Acharya A, Ruvinov SB, Gal J, Moll JR, Vinson C (2002) A heterodimerizing leucine zipper coiled coil system for examining the specificity of a position interactions: amino acids I, V, L, N, A, and K. Biochemistry 41: 14122-14131.

89. Acharya A, Rishi V, Vinson C (2006) 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 45: 11324-11332.
90. Hadley E, Gellman S (2006) An Antiparallel [alpha]-Helical Coiled-Coil Model System for Rapid Assessment of Side-Chain Recognition at the Hydrophobic Interface. Journal of the American Chemical Society 128: 16444-16445.
91. Hadley E, Testa O, Woolfson D, Gellman S (2008) Preferred side-chain constellations at antiparallel coiled-coil interfaces. Proceedings of the National Academy of Sciences: 0709068105.
92. Horovitz A (1996) Double-mutant cycles: a powerful tool for analyzing protein structure and function. Folding and Design 1: R121-R126.
93. Fong JH, Keating AE, Singh M (2004) Predicting specificity in bZIP coiled-coil protein interactions. Genome Biol 5.
94. Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358: 86-89.
95. Vajda S, Sippl M, Novotny J (1997) Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 7: 222-228.
96. Lazaridis T (2000) Effective energy functions for protein structure prediction. Current Opinion in Structural Biology 10: 139-145.
97. Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins 49: 350-364.
98. Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci U S A 99: 5896-5901.
99. Boas FE, Harbury PB (2007) Potential energy functions for protein design. Curr Opin Struct Biol.
100. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, et al. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry 4: 187-217.
101. Cornell W, Cieplak P, Bayly C, Gould I, Merz K, et al. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. Journal of the American Chemical Society 117: 5179-5197.
102. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, et al. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34: 82-95.
103. Crick FHC (1953) The Fourier Transform of a Coiled-Coil. Acta Crystallography 6: 685-689.
104. Nilges M, Brünger AT (1991) Automated modeling of coiled coils: application to the GCN4 dimerization region. Protein Eng 4: 649-659.
105. Nilges M, Brünger AT (1993) Successful prediction of the coiled coil geometry of the GCN4 leucine zipper domain by simulated annealing: comparison to the X-ray structure. Proteins 15: 133-146.
106. Harbury PB, Tidor B, Kim PS (1995) Repacking protein cores with backbone freedom: structure prediction for coiled coils. Proc Natl Acad Sci U S A 92: 8408-8412.

107. Harbury P, Plecs J, Tidor B, Alber T, Kim P (1998) High-Resolution Protein Design with Backbone Freedom. Science 282: 1462-1467.
108. Keating AE, Malashkevich VN, Tidor B, Kim PS (2001) Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. Proc Natl Acad Sci U S A 98: 14825-14830.
109. Grigoryan G, Keating AE (2006) Structure-based prediction of bZIP partnering specificity. J Mol Biol 355: 1125-1142.
110. Grigoryan G, Ochoa A, Keating AE (2007) Computing van der Waals energies in the context of the rotamer approximation. Proteins 68: 863-878.
111. Kono H, Saven JG (2001) Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. J Mol Biol 306: 607-628.
112. Grigoryan G, Zhou F, Lustig SR, Ceder G, Morgan D, et al. (2006) Ultra-fast evaluation of protein energies directly from sequence. PLoS Comput Biol 2.
113. Apgar JR, Hahn S, Grigoryan G, Keating AE (2009) Cluster expansion models for flexible-backbone protein energetics. Journal of computational chemistry.
114. Grigoryan G, Reinke A, Keating A (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature 458: 859-864.

Chapter 2

Discriminating coiled-coil dimers vs. trimers using an annotated sequence database and Multicoil2

Author Contributions

This work was done in collaboration with Andrew V. McDonnell, Bonnie Berger and Amy E. Keating. A.V.M. reimplemented the Multicoil2 algorithm in Java and contributed cross-validation code. A.V.M. and B.B. contributed Paircoil2 training databases.

2.1 Abstract

The alpha-helical coiled coil can adopt a variety of topologies, among the most common of which are parallel and antiparallel dimers and trimers. In order to facilitate computational approaches to predicting the structural specificity of coiled coils, we

constructed a database of coiled-coil-forming sequences. This database, comprised of 2,105 sequences containing 124,088 residues, contains highly reliable structural annotations that are based on experimental data in the literature. Sequences are annotated with respect to oligomerization state, helix orientation and coiled-coil heptad positions. We used the database to train Multicoil2, an updated version of the Multicoil program. Multicoil2 predicts both the location and oligomerization state (two versus three helices) of coiled coils in protein sequences. We tested Multicoil2 using a variety of cross-validation methods, including a stringent leave-family-out framework that reflects expected performance on challenging new prediction targets that have minimal sequence similarity to known coiled-coil families. Multicoil2 shows enhanced performance over both Multicoil and Paircoil2. The training database, the Multicoil2 program, and scripts that can be used to run various cross-validation tests will be available for download.

2.2 Introduction

The alpha-helical coiled coil is a protein motif characterized by superhelical twisting of two or more alpha helices around one another. The structure of the coiled coil is characterized by a regular, repeating backbone geometry and characteristic side-chain interactions termed "knobs-into-holes" packing. Coiled coils are remarkably prevalent in protein structures, and they adopt a wide range of structural topologies with variations in helix orientation and oligomerization state. Structurally characterized examples of native and designed coiled coils range from two to seven helices, with dimers and trimers most common[1]. Knowledge of coiled-coil architecture is important for understanding the

overall structure and function of coiled-coil-containing proteins, e.g. for inferring oligomerization stoichiometries[2], for determining whether attendant domains are close or distant in space[3], and for reasoning about mechanism in molecular machines, signaling cascades and motors[4].

Coiled-coil structures are encoded by a seven-residue heptad pattern of the form $(HPPHPPP)_n$, where H positions are predominantly hydrophobic and P positions are predominantly polar. The positions in the repeat are denoted by the letters **a** – **g**, with **a** and **d** hydrophobic. The repeating sequence motif makes the coiled-coil structure amenable to prediction, and several algorithms have been developed to detect the presence of coiled-coil-forming segments in protein sequence[5]. More complete annotation of structure, however, requires predicting the number of helices participating in a coiled-coil bundle, as well as the axial alignment and orientation of all helices. Among these aspects of coiled-coil structure, the prediction of oligomerization state has so far received the most attention, though work on other aspects of structural specificity is becoming tractable as the number of solved coiled-coil structures grows[6].

In 1997, the Multicoil algorithm was introduced for predicting coiled-coil dimer vs. trimer propensities[7]. It showed outstanding performance at the time, and after 12 years remains the only widely used method for predicting coiled-coil oligomerization state. The algorithm has been used extensively and successfully to predict the propensity of coiled-coil sequences to form dimers or trimers and has been cited over 450 times. Multicoil is based on the Paircoil algorithm, which uses a probabilistic framework to detect coiled-coil-forming segments in proteins, based on residue-pair frequencies in known coiled coils. Multicoil uses a pair of sequence databases constructed from both

authentic dimers and trimers to derive pairwise residue frequency tables, which are then used to derive both dimer and trimer propensities. Coiled-coil dimer versus trimer prediction has also been attempted with the approach of Woolfson and Alber[8]. Both approaches rely only on sequence-level features to make predictions.

At the time of initial development, relatively few sequences were available to train the Multicoil program. With only 6,300 coiled-coil-trimer residues in the original training database[7], it is unclear whether enough data were available to adequately describe sequence features that determine oligomerization state for coiled coils broadly. In addition, the limited amount of data also restricted the types of validation tests that could be run. However, significant amounts of new sequence are now available. Genome databases have grown larger, with 780% growth since 1997[9]. Many more protein structures are available, and the SOCKET algorithm[10] can now be used to automatically detect coiled-coil sub-structures in the Protein Data Bank. Finally, many new coiled-coil-containing protein families have been experimentally characterized and described in the literature. This has increased the number as well as the diversity of known coiled coils.

The availability of new data motivated us to construct a database of coiled-coil sequences useful for training as well as testing coiled-coil structure prediction methods. Here, we describe this database, which is annotated with structural details and grouped by biologically relevant families for validation purposes. We also present the results of re-training Multicoil and illustrate how the validation method used affects perceived performance. We discuss the impact of the larger dataset on the performance of the method, identify areas of improvement, and suggest future directions for the prediction of

coiled-coil oligomerization state. We are releasing the database as well as the source code and executables for the updated prediction software. We are also releasing validation scripts that can be used with the database to evaluate other methods. We hope that the availability of the data and testing framework may motivate further improvements in, and novel approaches to, the prediction of coiled-coil structures.

2.3 Method

2.3.1 Database construction

The coiled-coil database was derived from three sources: the Paircoil2 training set, coiled coils detected in the PDB using SOCKET, and new coiled-coil families described in the literature.

The Paircoil2 training database consists primarily of manually annotated sequences from long coiled coils (i.e. myosins, tropomyosin, intermediate filaments, viral coat proteins, cortexillin, SNAREs) as well as many examples of shorter coiled coils (i.e. bZIPs, flagellin, hemagglutinin) [11]. These sequences were not processed prior to the global filtering step described below.

Structure-derived training examples resulted from application of SOCKET [10] to a version of the PQS database[12] downloaded on September 3, 2008. SOCKET was run with a distance cutoff of 7.0 Å to reduce false positives. Skips and stutters were eliminated by removing 10 residues on either side of any heptad discontinuity. Sequences shorter than 21 residues were discarded, and the remaining sequences were filtered for

coiled-coil sequence identity no greater than 90%. Sequence identity filtering was performed using BLAST-discovered alignments between coiled-coil regions only. Contiguous clusters of sequences linked by edges representing >90% identity were removed and replaced with the longest constituent coiled-coil domain. Structure-derived sequences were grouped into families using information from the SCOP database[13] by pooling sequences sharing the same SCOP superfamily.

Coiled-coil families designated as "new" were not present in either the Paircoil2 or SOCKET-derived sets of sequences. These families have no representation in the structural database, but have strong experimental evidence to support the formation of either a parallel dimeric or parallel trimeric coiled coil. Seed sequences were downloaded from the NCBI[9] and the heptad register was assigned using Paircoil2. These sequences were then used as BLAST[14] queries against a recent copy of the UniRef100 protein sequence database[15]. BLAST results were filtered to exclude hits with E-value greater than $1x10^{-15}$. Hits were also excluded if the BLAST-provided alignment did not fully align the coiled-coil region from the query to the subject. Heptad assignment for hit sequences was copied from the query, based on the BLAST alignment, and was accepted if the Paircoil2 P-score of the given heptad was < 0.20. The resulting sequence set was subsequently filtered for coiled-coil sequence identity no greater than 90%.

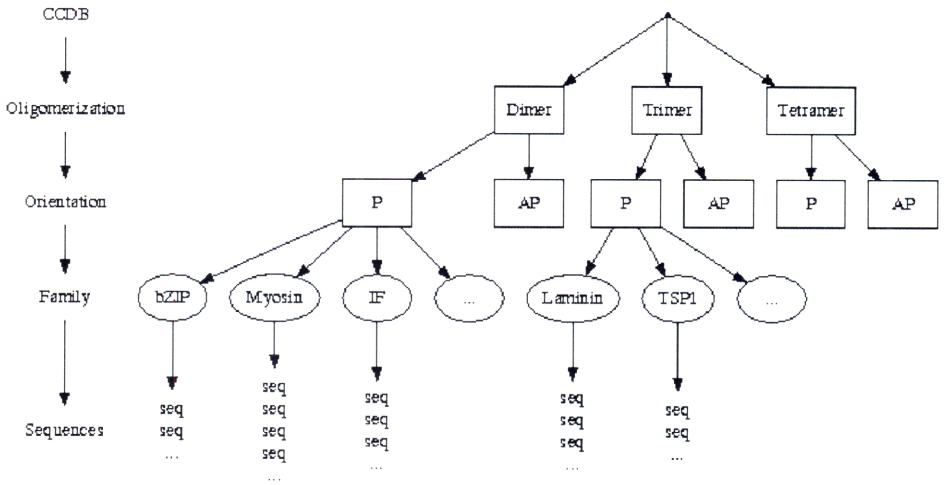
The complete database was named NPS (for **N**ew families, **P**aircoil2, **S**OCKET). To construct it, sequences from the three sources were pooled and filtered for coiled-coil sequence identity no greater than 90%. Entries in the database were annotated based on oligomerization state (dimer, trimer or tetramer; no other oligomerization states were represented) and orientation (parallel vs. antiparallel). Orientation was defined as parallel

if all helices were oriented the same direction, and antiparallel otherwise. Finally, within each annotation group, families originating from different primary sources were combined using family information from the SYSTERS database[16]. Family identification was determined by using BLAST to compare individual protein sequences to the SYSTERS non-redundant database, which is annotated with SYSTERS family IDs. Clusters of families sharing a common SYSTERS family assignment were combined into a single family. In particular, TPR (from the new-family source) clustered together with MLP (from the Paircoil2 database), which has been previously discussed[17].

2.3.2 Database format

The database is organized hierarchically according to oligomerization state, helix orientation, protein family and sequence, as depicted in Figure 2-1. Each family is contained within one text file, with each sequence represented by a four-line record. The first line contains the protein name or PDB-id and BLAST E-value to the query (plus query name), where appropriate. The second line contains structural descriptors drawn from a standardized vocabulary, such as "long parallel homo dimer". In the last two lines, each sequence is annotated with its coiled-coil domain using heptad-register notation (**a**-**g**). Flanking un-annotated sequence is also included, although this may not span the entire protein, e.g. when entries were taken from the PDB or from the Paircoil2 (PC2) training set. The flanking sequence may or may not form a coiled-coil structure, and our database is not authoritative for coiled-coil domain boundaries. The NPS database is available on our website.

(a)



(b)

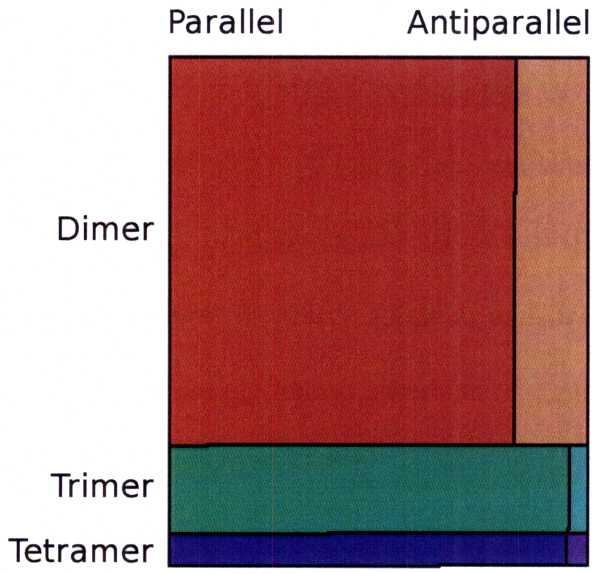


Figure 2-1. Overview of the NPS coiled-coil database. (a) Structure of the database. Sequences are grouped hierarchically by oligomerization state, orientation, and family. (b) The size of the database by oligomerization state and orientation. Block areas are drawn to scale to represent the number of sequences present in each structural class. Block colors correspond to parallel dimers (red), antiparallel dimers (brown), parallel trimers (green), antiparallel trimers (cyan), parallel tetramers (blue), antiparallel tetramers (violet).

2.3.3 *Database Analysis*

Residue frequencies were computed as the fraction of occurrences of a given residue at a certain heptad position relative to all residues at the same heptad position. Statistical significance was determined using a two-tailed exact binomial test for each residue at each heptad position, with observed frequencies from the parallel trimer database and expected frequencies from the parallel dimer database. To reduce family-size biases, statistical significance was computed independently for each family and P-values for each residue and heptad position were combined across families using Fisher's method.

Sequence similarity tests were computed using BLAST-derived alignments, and reported as the maximum sequence identity among all sequences between families.

2.3.4 *Multicoil in brief*

Multicoil is based on the probabilistic framework of the Paircoil method. Paircoil can detect the presence of coiled-coil forming segments in protein sequence data by using residue-pair frequencies derived from a single database of known coiled-coil sequences[18]. Multicoil uses frequencies derived from two databases containing known dimers and known trimers. For each database, residue-pair frequencies P, along with background single-residue frequencies P_{bkg}, are used to calculate residue propensities for residue r_i according to the formula

$$R(r_i, r_{i+d}, h_i, h_{i+d}) = \ln \frac{P(r_i, r_{i+d}, h_i, h_{i+d})}{P_{bkg}(r_i, r_{i+d}, d)} - \ln \frac{P(r_i, r_{i+d})}{P_{bkg}(r_i)}.$$

Here i is an index to the sequence, d is a distance in the range 1-7, r_i is the residue at position i, and h_i is the heptad assignment of residue i.

To score residues in a sequence, overlapping windows of length 28 are defined and all residue propensities under those windows are summed to produce a set of window scores. The final score of a given residue is the maximum of all window scores containing that residue. This process is repeated for both dimer and trimer frequency tables, over all possible heptad register assignments h_i and pairwise distance values d. The result of this calculation is a 14-dimensional raw-score vector for each residue position and possible heptad assignment in the sequence (7 values of d used with each of the dimer and trimer databases). In practice, to reduce noise, this vector is reduced to a 6-dimensional vector by choosing three distances for each of the two frequency tables. Based on testing the performance of the method with all possible 6-dimensional vectors, we did not observe significant sensitivity in performance with respect to this parameter. Therefore, we selected those distances (dimer: 2, 3 and 4; and trimer: 3, 4 and 5) used previously[7].

Using the reduced raw score vectors, probabilities are calculated using a set of predetermined Gaussian functions along with estimated prior probabilities for each possible class (dimer, trimer and non-coiled-coil). This is done using Bayes' theorem[19], written as (e.g. computing dimeric probability)

$$P(di \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid di)P(di)}{P(\mathbf{X})}$$

where $\mathbf{X}$ is the score vector of a given residue, P($\mathbf{X}$|di) is the value of the Gaussian

function evaluated at the point **X**, P(di) is the prior probability of any given residue being found as a dimer, and P(**X**) is the cumulative probability of score vector **X** being any one of the three classes, written as

$$P(\mathbf{X}) = P(\mathbf{X} \mid di)P(di) + P(\mathbf{X} \mid tri)P(tri) + P(\mathbf{X} \mid noncc)P(noncc).$$

The Gaussians are determined by computing means and covariances from the distributions of raw scores over a training set. More details about the training sets used are in the Validation section. The final residue probabilities and predicted heptad assignment according to each class are taken as the maximum probability over all possible heptad assignments.

2.3.5 Multicoil rewrite

Multicoil was rewritten in Java using the BioJava libraries[20]. The algorithm remains the same as previously published[7]. Included in the distribution is a set of scripts, written in Perl and Python, useful for testing Multicoil with the various validation tests described here. The JAR archive is freely available on our website and source code is available upon request.

2.3.6 Multicoil2 training database

The Multicoil2 training database was derived from the NPS database with minor modifications. For cross-validation testing, we combined all families containing four sequences or fewer into a "Miscellaneous" family. In addition, Multicoil2 is unable to

score sequences containing non-canonical residues such as B, J, X or Z; therefore, sequences containing such residues were omitted. This criterion excluded only one sequence from training.

2.3.7 Estimating prior class probabilities

The conversion of scores to probabilities requires a conditional probability distribution for coiled-coil scores and an estimate of the prior probability of a residue being found in a coiled coil. Given a representative database, this estimate can be made by determining the 14-dimensional raw score vector of every residue and subsequently finding the values of P_{dim}, P_{trim} and P_{noncc} that maximize the likelihood function

$$\sum_{\mathbf{X}=scores} \log[P_{\dim} V_{\dim}(\mathbf{X}) + P_{trim} V_{trim}(\mathbf{X}) + P_{noncc} V_{noncc}(\mathbf{X})]$$

where V_{dim}, V_{trim} and V_{noncc} are multi-dimensional Gaussian functions previously fit to leave-family-out score distributions, and $\mathbf{X}$ is iterated over each raw score vector. We performed this analysis with non-redundant protein sequence databases from *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Escherischia coli* K12 downloaded from the NCBI on November 8, 2008. Score vectors were calculated for all residues, over all 7 residue-residue distances and over both training databases. To find maximal values of the prior probabilities, we performed a two-dimensional grid search, varying P_{dim} and P_{trim} from 0.0 to 0.2 with a step size of 0.001.

2.3.8 *Assessing performance using cross-validation*

Multicoil2 testing was performed under three cross-validation frameworks: leave-family-out, leave-N-percent-identity-out, and leave-sequence-out. Training under cross-validation required two steps. First, residue frequencies were tallied. Second, three Gaussian functions were fit to the distributions of dimer, trimer and non-coiled-coil raw scores. Raw scores and Gaussian fits were derived under the appropriate validation protocol. For example, under leave-family-out validation, the raw scores used to fit the Gaussians were generated through a leave-family-out protocol. A flow chart describing this process is shown in Figure 2-2. This is different from the testing protocol in the previous version of Multicoil, where Gaussians were fit to the average of raw scores from leave-sequence-out and leave-family-out tests, due to the much smaller amount of available sequence data. Non-coiled-coil raw scores were determined from PDB-minus, a database of protein sequences known to not contain coiled coils[11]. This database contains 6363 sequences, totaling 1,480,158 residues.

The cross-validation frameworks differ in the way they divide the training database. Leave-family-out validation uses family definitions from the training database as biologically relevant sequence groupings. For each family, training sets were prepared that excluded that family, and then prediction performance was evaluated for all family sequences. Leave-N-percent-identity-out used individual sequences as the testing unit. For each sequence, training sets excluded that sequence along with all other sequences with sequence identity greater than or equal to the predefined cutoff. At the 100% identity cutoff, only the sequence under test was excluded from the training databases, which we

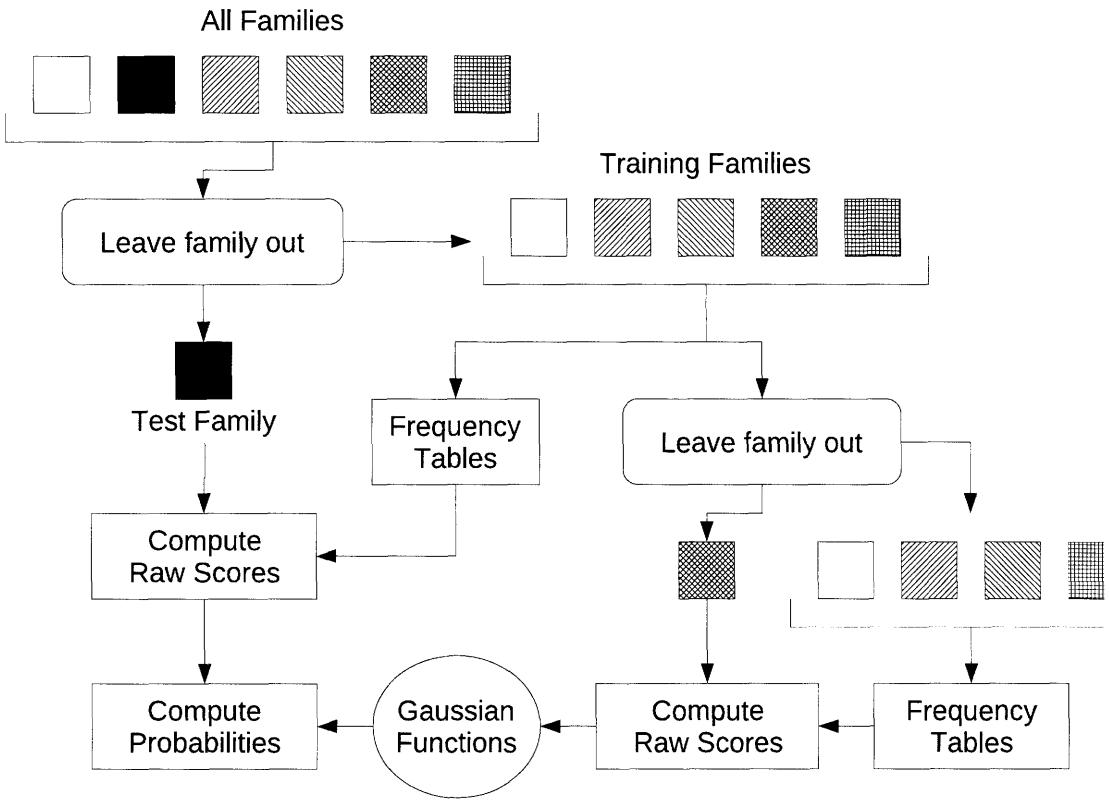


Figure 2-2. Flow chart of validation method. Rounded rectangles ("leave family out") indicate iteration over all input families. This figure depicts leave-family-out cross-validation; leave-sequence-out and leave-N%-out validation is similar but with sequences or N%-identical clusters as the training unit, respectively.

refer to as leave-sequence-out. Finally, leave-nothing-out was used for comparison. In this case, the testing database was prepared containing all available sequences and no sequences were removed for training.

Performance was evaluated at the sequence level. For each sequence, a prediction was labeled as correct if the sum over all residues in the known coiled-coil region of the probability of the correct class was greater than the same sum over the probability of the incorrect class, as described previously[7]. Performance was reported as the fraction of sequences correct out of all tested sequences.

2.4 Results

2.4.1 A database of structurally annotated coiled-coil sequences

Probabilistic methods such as Paircoil and Multicoil (also COILS, Marcoil) rely on discerning discriminating features of coiled coils using sequences of known structural classification[7,11,21,22]. For the problem of distinguishing coiled coils from non-coiled coils, many training examples exist. However, structurally annotated data that could be informative about features such as coiled-coil oligomerization state are less abundant. To address this, we assembled structurally annotated coiled-coil sequences from three sources: sequences that form coiled coils in crystal structures, sequences recently reported to fold as coiled coils based on experimental data, and sequences previously curated for the Paircoil2 database. We excluded sequences shorter than 21 residues, due to ambiguity about whether these regions are truly coiled coils[10].

The Paircoil2 training database[11] has been recently updated and contains 1,382 coiled-coil sequences (94,876 coiled-coil residues) longer than 21 residues each. To collect coiled coils from crystal structures, the SOCKET program[10] was run on the most recent release of the PQS database[12]. Filtering at the 90% sequence level gave 231 coiled coils contributing 340 coiled-coil sequences (10,605 residues). From the literature, we compiled twelve protein families recently reported to contain coiled coils that were characterized experimentally using electron microscopy, deletion analysis, analytical ultracentrifugation and/or cross-linking studies: astrin[23,24], fer[2],

hsfbp1[25], 11orf1[26], matrilin[27], nemo[28], numa[29], snv_n[30], spc110p[3,31], tenascin[32], tpr[33] and tsp1[34]. Four of these families (astrin, numa, spc110p and tpr) are reported to be parallel dimers; the remaining eight families are described as parallel trimers. Coiled-coil domains in these proteins were manually assigned with the assistance of the Paircoil2 program at a P-score cutoff of 0.05. Seed sequences for each of these families were used to search the UniRef100 protein sequence database[15] for additional family members, resulting in a total of 561 coiled-coil sequences (26,065 residues).

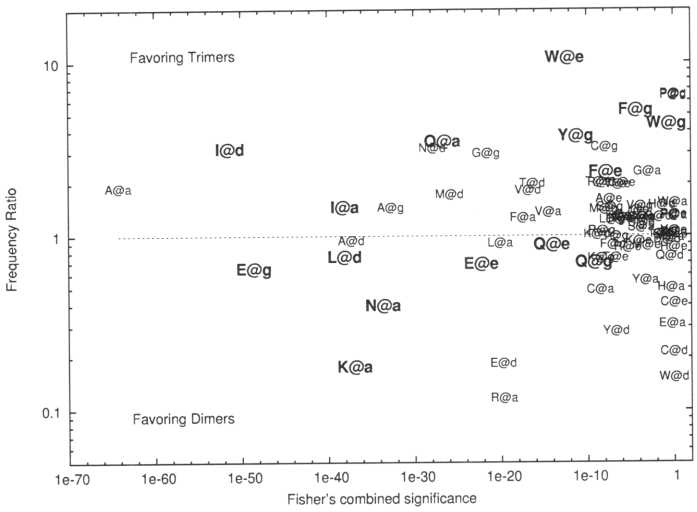
We combined sequences from these three sources and then re-filtered to 90% sequence identity to form the NPS (**N**ew families, **P**aircoil, **S**OCKET) database, resulting in a total of 2,105 sequences containing 124,088 residues. The database was organized according to the structural features of each sequence, as depicted in Figure 2-1a. Sequences were grouped by oligomerization state (dimer, trimer, tetramer) and by orientation (parallel, antiparallel). Figure 2-1b illustrates the size of the NPS database, broken down by structural class. Despite adding a significant amount of trimeric sequence, the size of the dimer fraction outweighs the trimer and tetramer fractions in both databases. There are very few antiparallel trimer and tetramer sequences in the NPS database.

Within each structural class, sequences were grouped by family, in order to enable leave-family-out validation methods. We found that the most effective way to produce meaningful family definitions was to preserve the family assignments made within each sequence source, and then to combine families upon merging sources using a previously developed family database[16]. Due to the low complexity of the coiled-coil motif, families defined solely by sequence identity were not able to recapitulate known families.

2.4.2 *Features of the dimer and trimer sequences*

Using the NPS database, we looked for differences in residue frequencies among known dimer and trimer sequences. Previous studies have made similar investigations, albeit on much smaller databases[8,35]. Tables of residue frequencies as a function of heptad position in the different structural classes are given in the supplementary material. While these may be biased due to family composition, many trends can still be observed from features that are common among families. Figure 2-3a shows frequency ratios for coiled-coil core (**a**, **d**) and edge (**e**, **g**) heptad positions, plotted as a function of the statistical significance of that frequency observation given our database. Experimental data are available to support those observations marked in bold[8]. Other significant observations, such as a preference in trimers for Ala@**a**, Asn@**d**, Met@**d** and Gly@**g**, are variously distributed over families. For example, Ala@**a** is found mostly in laminins, and Gly@**g** is found mostly in the viral coat family. Given the relative rarity of these features over all families, it is difficult to determine whether they indicate a strong influence on the oligomerization state or whether they are conserved for other reasons. Other residues that are enriched in trimers relative to dimers are more distributed across families, e.g. Met@**d** is found in five families ($P < 0.001$), and Asn@**d** is found in four families ($P < 0.001$). This suggests a possible role for these residues in determining coiled-coil oligomerization state specificity.

(a)



(b)

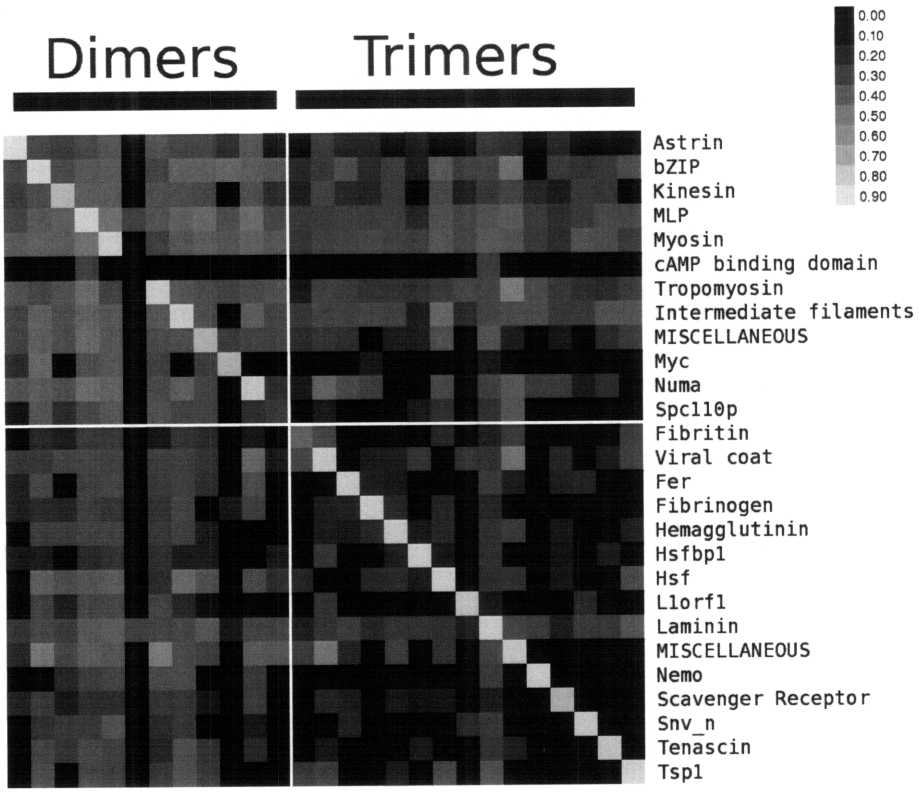


Figure 2-3. Characteristics of the NPS coiled-coil database. (a) Residue frequency comparisons. The horizontal axis gives the statistical significance of a given observation as a P-value computed using Fisher's method. The vertical axis represents the ratio of observed trimer vs. dimer frequencies. Highlighted residues indicate residue/position pairs previously determined experimentally to favor either dimer or trimer formation. [8] (b) Maximum sequence similarities among dimer and trimer families.

We examined the sequence similarity within and between families of coiled coils in our database. Figure 2-3b shows the maximum sequence identity among all parallel dimer and trimer families in the NPS database. The brightly colored diagonal shows that for nearly all families, sequence identity reaches the homology filter limit of 90%. However, the cross-family similarities are much lower, with a maximum value of 50-60%. The dimer families appear to have higher similarity to other dimers than the trimer families have to other trimers, suggesting that the trimers may be more diverse than the dimers. Interestingly, the dimer-to-trimer similarity is relatively high, indicating that separating these populations using sequence-based methods is non-trivial.

Kammerer et al. have discussed the significance of a certain sequence motif in the folding of short trimeric coiled coils[36]. We searched the NPS database for the presence of this "trimerization motif", denoted by the PROSITE pattern R-[ILVM]-X-X-[ILV]-E. When looking at all sequences 50 residues or shorter, we found the motif in 17.0% of parallel trimer sequences and 13.2% of parallel dimer sequences. Interestingly, when considering all sequences, the motif was found in 19.7% of parallel dimer sequences while it was only present in 16.4% of all parallel trimer sequences. These slight and contradictory differences suggest that the motif, while possibly exerting a significant influence on structure in certain cases, is not likely to be a widespread determinant of coiled-coil oligomerization, and cannot be used for high-confidence prediction.

2.4.3 Retraining Multicoil to Multicoil2

Multicoil2 is designed to discriminate between two coiled-coil structural classes.

The NPS parallel dimer and parallel trimer structural classes comprise 1,686 sequences, or 80% of the total NPS database, and we selected these as the relevant classes for training and testing.

Calculating coiled-coil propensity values in the Multicoil framework requires prior probabilities for the distribution of dimers, trimers and non-coiled-coils. We re-implemented the previously described method[7] of maximum log likelihood as described in the Methods to determine these probabilities for a series of representative protein sequence databases. In the original paper, the OWL database was used as a representative set of proteins. However, the selection of an appropriate representative database can be important, as there may be variations in the background coiled-coil probabilities among organisms. Here, we used a variety of complete genomes to examine this possibility. The prior probabilities that maximize the log likelihood are shown in Table 2-1. These values must be regarded as rough estimates, because the sensitivity of the likelihood to the prior probabilities is low, as originally demonstrated by Wolf et al[7]. It is also likely that these are underestimates of the true number of coiled coils, given that the method often does not assign high raw scores to coiled-coil families that are only

Organism	Size (residues)	Dimer prior	Trimer prior
E. coli K12	1,315,392	0.002	0.004
S. cerevisiae	2,914,765	0.005	0.017
C. elegans	10,043,780	0.009	0.016
D. melanogaster	11,824,157	0.011	0.020
M. musculus	15,624,175	0.017	0.017
H. sapiens	17,175,172	0.019	0.018

Table 2-1. Estimates of the coiled-coil dimer and trimer content of various genomes. Dimer and trimer priors were calculated using the maximum log-likelihood method described in Methods. See the main text for caveats.

distantly related to those in the training set (see below). In the validation tests described below, we used the value 0.02 for both P_{dim} and P_{trim}.

2.4.4 Validation

Testing should ideally be designed to approximate the accuracy of a method that can be anticipated under typical use. Cross-validation testing, in which the sequence to be tested and its close relatives are omitted from the training set, is commonly used for this purpose. There are many ways to set up validation tests. The number and identity of sequences omitted during training are important considerations that can dramatically affect perceived performance. Here we use validation tests that successively remove every member of the training set. This is in contrast to approaches that reserve a single portion of the known space for testing purposes[5]. Due to the diversity of coiled coils, it is important to characterize differences in performance over many test sequences, which are difficult to capture in a static test set. We consider and compare leave-nothing-out, leave-sequence-out, leave-family-out and leave-percent-identity-out. The raw score distributions from leave-nothing-out, where all sequences including the one being tested are used for training, are shown in Figure 2-4a. Prediction performance is given in Table 2-2. Leave-nothing-out is obviously a poor estimator of actual expected performance on new sequences.

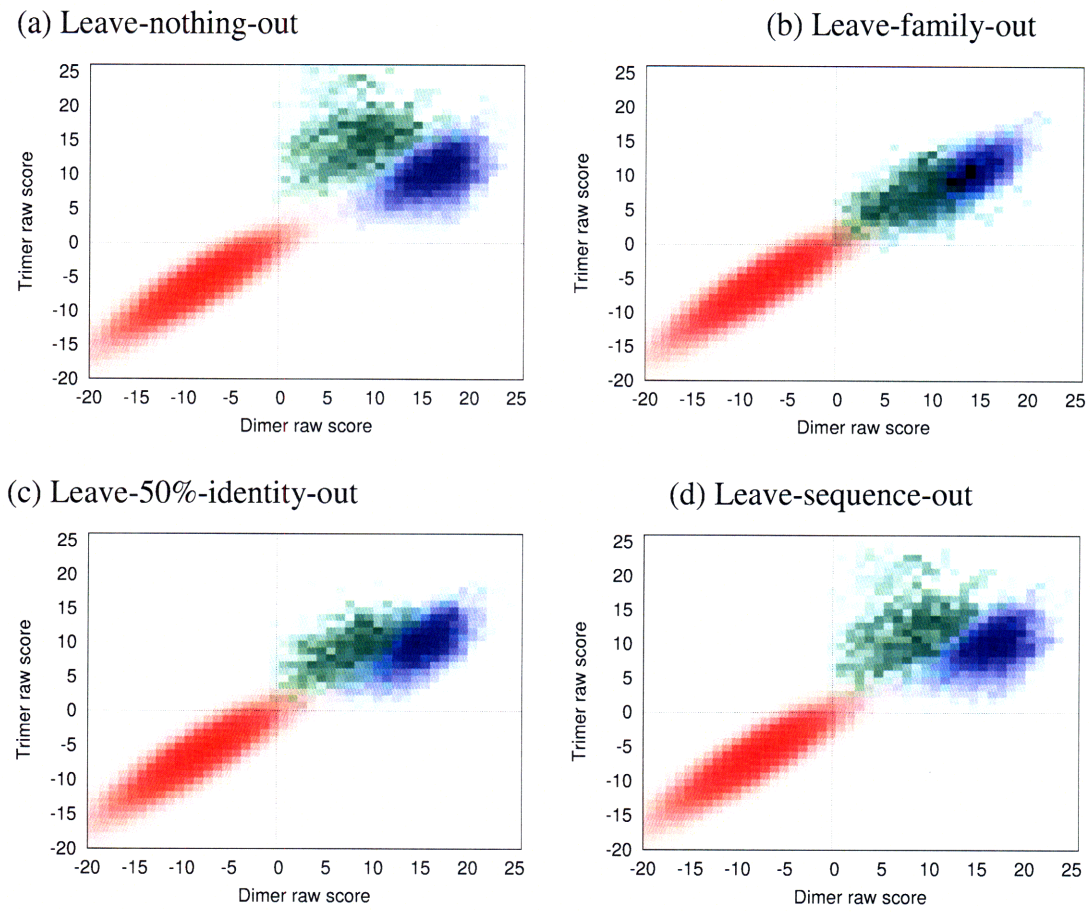


Figure 2-4. Distributions of raw scores resulting from cross-validation testing. Axes describe raw score averages over the three dimer (horizontal) and trimer (vertical) scoring distances. Scores for parallel trimers are in green (upper center); parallel dimers are in blue (center right) and non-coiled coils from PDB-minus are in red (lower left).

2.4.4.1 Leave-family-out testing

A particularly stringent validation test involves successively omitting biologically defined families of sequences from the training set. This simulates the behavior of the method on never-before-encountered coiled-coil families. We used family definitions from the NPS database. A plot of the average raw scores generated is shown in Figure 2-4b, and contrasted with the raw scores generated through the leave-nothing-out test (Figure 2-4a). A striking observation is that there is significantly less separation observed

Family	Total Seqs	Leave-nothing-out Seqs correct	Leave-nothing-out % correct	Leave-family-out Seqs correct	Leave-family-out % correct	Leave-50%-sequence-identity-out Seqs correct	Leave-50%-sequence-identity-out % correct	Leave-sequence-out Seqs correct	Leave-sequence-out % correct
Parallel Dimers									
Astrin	27	26	96.3	16	59.3	19	70.4	21	77.8
bZIP	114	112	98.3	103	90.4	109	95.6	109	95.6
Kinesin	32	31	96.9	31	96.9	30	93.8	30	93.8
MLP	344	332	96.5	309	89.8	297	86.3	319	92.7
Myosin	220	219	99.6	212	96.4	213	96.8	215	97.7
cAMPbd	5	0	0.0	0	0.0	0	0.0	0	0.0
Tropomyosin	67	66	98.5	65	97.0	65	97.0	64	95.5
IF	371	370	99.7	319	86.0	360	97.0	369	99.5
Myc	8	8	100.0	8	100.0	6	75.0	8	100.0
Numa	87	79	90.8	80	92.0	74	85.1	77	88.5
Spc110p	8	8	100.0	6	75.0	6	75.0	6	75.0
Miscellaneous	49	37	75.5	24	49.0	27	55.1	27	55.1
Parallel Trimers									
Fibritin	5	5	100.0	3	60.0	4	80.0	4	80.0
Viral coat	71	71	100.0	69	97.2	71	100.0	71	100.0
Fer	24	24	100.0	10	41.7	11	45.8	21	87.5
Fibrinogen	23	22	95.7	18	78.3	20	87.0	20	87.0
Hemagglutinin	14	14	100.0	12	85.7	14	100.0	14	100.0
Hsfbp1	9	9	100.0	6	66.7	9	100.0	9	100.0
Hsf	32	26	81.3	8	25.0	15	46.9	24	75.0
L1orf1	14	14	100.0	11	78.6	12	85.7	12	85.7
Laminin	85	85	100.0	47	55.3	68	80.0	76	89.4
Nemo	13	12	92.3	3	23.1	2	15.4	10	76.9
Scavenger receptor	6	6	100.0	6	100.0	6	100.0	6	100.0
Snv_n	7	7	100.0	4	57.1	5	71.4	6	85.7
Tenascin	7	7	100.0	3	42.9	5	71.4	6	85.7
Tsp1	19	19	100.0	14	73.7	19	100.0	19	100.0
Miscellaneous	24	18	75.0	13	54.2	13	54.2	14	58.3
Total Dimer	1332	1288	96.7	1173	88.1	1206	90.5	1245	93.5
Total Trimer	353	339	96.0	227	64.3	274	77.6	312	88.4
Total Seqs	1685	1627	96.6	1400	83.1	1480	87.8	1557	92.4

Table 2-2. Multicoil2 prediction performance for all families under different testing protocols. Families taken from NPS parallel dimer and trimer sets. Miscellaneous includes all families with four or fewer sequences. Prior probabilities were 0.02 for both dimers and trimers.

between the distributions of known dimers and trimers. In particular, the raw score values for the trimer sequences drop significantly upon leave-family-out testing, from a mean of 15.14, when nothing is left out, to 7.25 when the family being tested is left out. Dimer raw scores also drop, but much less, from a mean of 16.62 to a mean of 13.57. As

expected, the consequence is a significant decrease in predictive ability, with the greatest decrease along the trimer score dimension.

Performance figures per-family are shown in Table 2-2. Across all families, dimer prediction performance is consistently higher than trimer prediction performance (88% vs. 64% overall). Despite the lower trimer prediction rate, many families are predicted well, including the Viral Coat family, with performance greater than 90%. However, four trimer families are predicted correctly at rates less than 50%. These include three of the new families added to the database: Fer[2], Nemo[28] and Tenascin[32], as well as the heat shock factor[37] proteins, which were included in the original Multicoil training set. The poor performance of these families can be understood by observing their distributions of raw scores, shown in Figure 2-5. In each case, the raw scores for many members of these families lie close to the center of the dimer raw score distribution (see Discussion).

2.4.4.2 Leave-percent-identity-out testing

An alternative validation protocol involves leaving out all sequences more than N% identical to the test sequence during training. This method is particularly appropriate in the real-world case where an unknown sequence can be determined to have a certain percent identity to sequences in the training set. We tested prediction performance using N = 50%, which is close to both the median and average maximum-identity of all families to the training database (median similarity 45%, average similarity 47%).

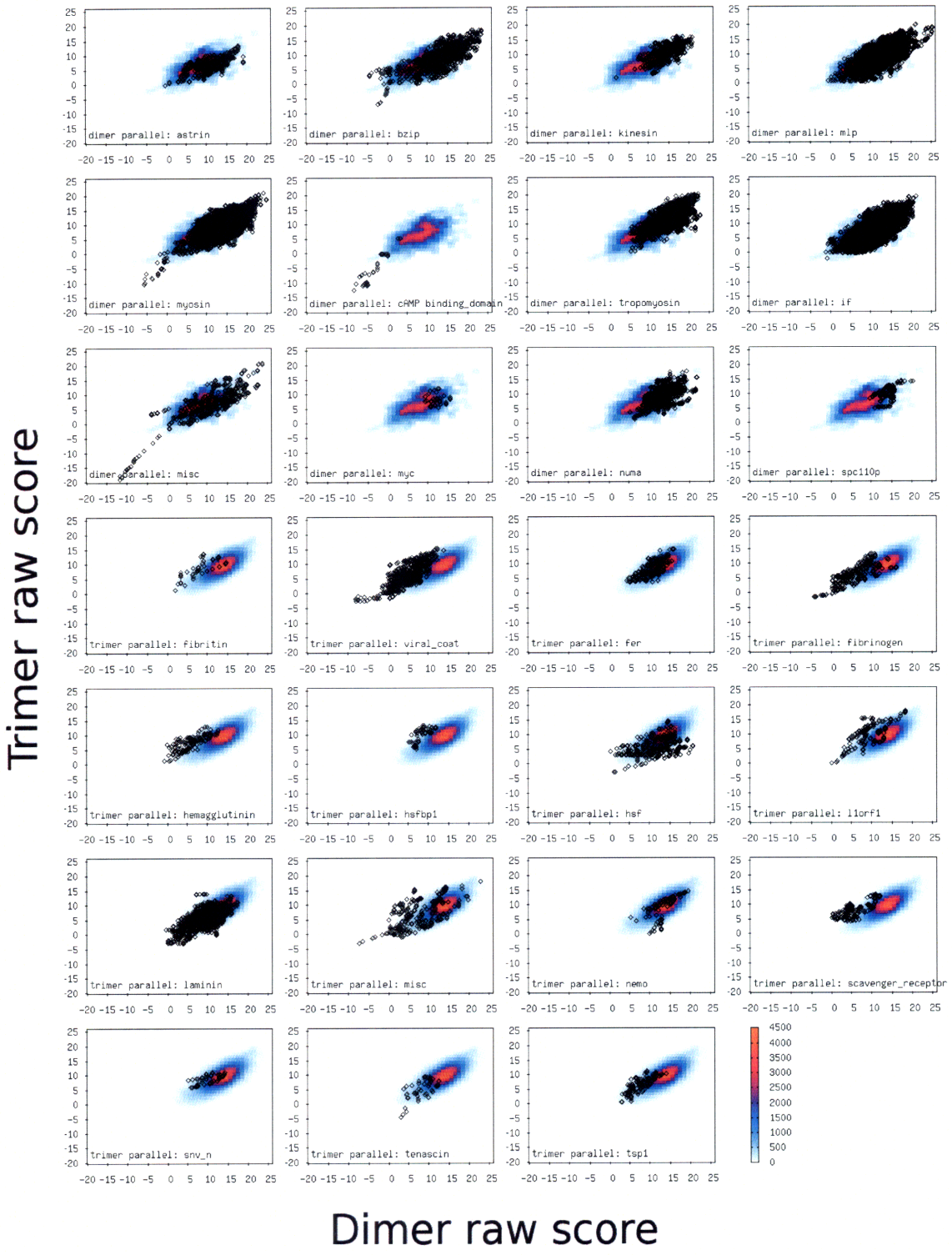


Figure 2-5. Leave-family-out cross-validated raw score plots per family. Raw score averages as in Figure 2-4. Black diamonds indicate per-residue scores of the tested family, and underlying heatmap indicates distribution of contrasting test class (for dimer parallel families, distribution is of parallel trimers and vice versa).

Raw score distributions from the leave-50%-identity-out calculations are shown in Figure 2-4c, with performance figures in Table 2-2. Leave-50%-identity-out testing gives prediction performance distinctly better than the stringent leave-family-out method (90.5% for dimers, 77.6% for trimers). This is expected, given that median percent-identities within families are mostly lower than 50%. We also observe a smooth increase in average performance as the N% identity threshold is increased (Figure 2-6). This seems to be a result of increasing the size of the training set, with more observations of less common residue pairs.

2.4.4.3 Leave-sequence-out testing

A much less stringent validation method involves omitting each single sequence in turn from the training set, when predicting the oligomerization state of that example. Although the maximum sequence identity between any two families is usually no larger than 55%, sequence identity within families can be as large as 90%. The distribution of raw scores achieved under leave-sequence-out validation is shown in Figure 2-4d, and is highly similar to the results of leave-nothing-out tests. The much greater separation of dimer vs. trimer scores is reflected in very high prediction performance (93.5% correct for dimers, 88.4% correct for trimers), shown in Table 2-2. Many of the families that performed poorly under leave-family-out testing do much better under leave-sequence-out. Leave-sequence-out performance is nearly as good as leave-nothing-out performance, consistent with the small differences between sequences in many families. Leave-sequence-out testing was used to test the original Multicoil application, because

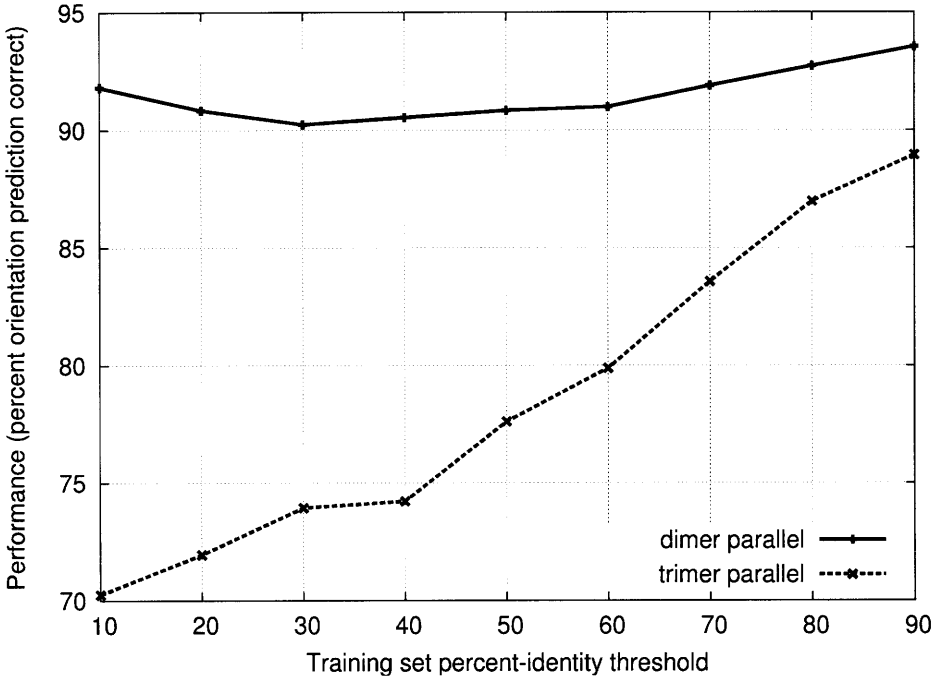


Figure 2-6. Effect of leave-N%-identity-out threshold on prediction performance. Curves represent average performance of all parallel dimers (solid) and parallel trimers (dashed) as a function of the percent-identity cutoff.

relatively few sequences were available at the time it was developed. However, now that more data are available, comparisons with more stringent cross-validation protocols make it clear that this testing may overestimate the performance that can be expected for some new sequences.

2.4.5 Improvement over Multicoil (1997)

To compare the effectiveness of the original and the re-trained versions of Multicoil on the dimer vs. trimer recognition problem, we assessed performance on new families not included in the 1997 training set. The original Multicoil program was run

and compared to the performance of Multicoil2 under leave-family-out cross validation. Overall, Multicoil (1997 version) correctly predicts the oligomerization state of 560 out of 784 test sequences (71% correct). Upon retraining to give Multicoil2, this increases to 637 correct predictions (81%). Interestingly, Multicoil2 appears to have a bias favoring the prediction of dimers, which is not observed with original Multicoil (data not shown).

Two effects contribute to changes in performance. One is the larger size of the new training database. The other is the fitting procedure used to derive the Gaussian functions that reflect the expected distributions of dimer, trimer and non-coiled-coil scores. In 1997, the prior probabilities for the three classes were fit using scores that were an average of leave-sequence-out and leave-family-out testing. Here, prior probabilities were fit using scores computed under the relevant validation protocol (i.e. for leave-family-out testing, score distributions came from leave-family-out raw scores). To separate these two effects, we trained Multicoil2 (using the new Gaussian-fitting protocol) on those families in the NPS database that were represented in training Multicoil (1997), and tested its performance on the newly identified families. This gave overall performance intermediate between the old and new methods (78% of all sequences correct). Interestingly, however, dimer prediction performance in this test was almost the same as the fully re-trained Multicoil2, whereas trimer prediction performance was much worse. This suggests that in terms of training set size, the original database was in fact adequate for dimers but not yet large enough for strong trimer prediction. In fact, this trend seems to hold even now, in 2009. Another conclusion is that the more stringent procedure of fitting the Gaussians according to the expected score distribution under cross-validation improves performance under cross-validation.

2.5 Discussion

Predicting the oligomerization state of a coiled coil from its sequence is a challenging problem that requires discriminating between closely similar structures. Efforts to discover simple motifs that specify coiled-coil oligomerization, while successful in specific circumstances[36], have not been generally applicable[38,39]. The Multicoil program, first published over 10 years ago, has proven valuable for this purpose and remains widely used. Many more coiled-coil sequences with known structure have now been annotated than were previously available, and we have used such examples to assemble a database of 124,088 structurally annotated coiled-coil residues. We report an updated version of Multicoil, Multicoil2, which is trained using this data. We show that the method exhibits enhanced ability to distinguish dimeric from trimeric structures. Validated performance on most families is high, even when assessed using stringent leave-family-out cross-validation, although some families, particularly certain trimers, show reduced performance under such a stringent test. This indicates that the Multicoil framework is appropriate and powerful, and that its predictive performance will continue to improve as more trimer families are discovered and added to the training set.

With a larger database available, we were able to test and compare several different validation protocols. The choice of method has a significant effect on perceived performance. Leave-sequence-out performance is not an accurate predictor of performance for a new sequence lacking close homology to the training set; it is nearly indistinguishable from non-cross-validated leave-nothing-out performance. Also, while

not tested here, methods in which a random N% of sequences are reserved as a test set also fail to control for the high degree of homology among training sequences in the same family[22]. We tested leave-family-out and leave-N%-identity-out methods as examples of scenarios more likely to be encountered in real prediction situations. Our results indicate that the expected accuracy of Multicoil2 for new sequences is ~88-90% for dimers and ~64-77.6% for trimers. To facilitate testing of other (old or new) methods under these various protocols, we are providing validation scripts in conjunction with our new database. We hope that future studies will report performance under some of the stricter validation protocols that we recommend.

Our validation tests suggest that the trimer database is still limiting prediction performance. Although trimers behave very well under leave-sequence-out and leave-nothing-out cross-validation, performance degrades in leave-family-out and leave-50-percent-identity-out tests. Thus, these sequences are "predictable" as long as the right training data is present. However, for many of the trimeric examples, the right training data is only found within a protein's own family. This indicates a lack of inter-family redundancy, which is critical to the ability of Multicoil to categorize new families correctly.

It is possible that coiled-coil trimers exist not as one generic structural class but as multiple classes, each with distinct sequence patterns determining trimerization. This may be true; however, our method of combining all known trimer-forming sequences into one class should not be detrimental unless aggregating such classes actually degrades performance. We do not have any evidence to suggest that this is the case; instead, when we remove from training those families with very poor performance (such as Nemo and

Fer), the prediction rates for the remaining families are unchanged. Thus, distant families in the current database do not detract from overall performance.

We considered some of the probable causes of performance differences among families. Each family has differences in the number of sequences and number of residues, as well as in residue composition and identity to other families in the training set. This makes it difficult to determine why certain families perform better or worse than others. However, we can eliminate some potential problems as unlikely. The training database was carefully prepared and thoroughly checked against available structural information and the literature; therefore, we expect that incorrect predictions are not due to errors in the training or test set annotations. For example, the cAMP binding domain, while predicted uniformly as likely to be trimeric, is observed to form dimers according to crystal structures[40]. Also, the CC2 domain of Nemo, while predicted to be a dimer, has been confirmed through a variety of experiments to form coiled-coil trimers in solution[28].

Low prediction performance is likely attributable to other factors. First, poorly predicted families could have unique sequence features, not shared by other families, that determine their oligomerization state. In addition, some families may have sequence features typical of both training databases. This could happen, e.g., if a sequence can form both a dimeric and a trimeric coiled coil. In such cases, the incorrect database may provide stronger scores than the correct database. This may be true for Nemo, where it is predicted that the LZ domain packs against the trimeric CC2 domain in an antiparallel fashion[28]. This complex structure likely impacts the residue distribution of the family, causing it to be poorly predicted. Finally, the largest families may have poor cross-

validated performance simply due to the large reduction in the size of the training set that results from omitting them from the training database. This may be true for the laminin family, which contributes 24% of sequences to the trimer database and shows very poor leave-family-out performance. However, this is not uniformly the case, as the viral coat family (20% of the trimer database) performs very well under all validation tests.

We expect that the most straightforward route to improving the performance of Multicoil2 is to continue to increase the size of the training databases. One important recent advancement is the CC+ database, which is regularly updated with coiled coils detected using the SOCKET method[39]. We have strongly considered the use of homology-search methods to increase the size of the known families; however, we must express caution, given that sequence homology does not always imply structure conservation[41], particularly in the case of coiled coils, where point mutations have been observed to significantly change structural preferences[42]. Also, the greatest improvement in leave-family-out performance will result from discovering new families that share sequence features with known families that now perform poorly; simply adding homologous sequences to existing families will likely not lead to significant improvements. Finally, the development of structure-based methods, which rely less on sequence-based training sets, provides an alternative route forward [6,43], that has not yet been extensively tested.

2.6 References

1. Parry DAD, Fraser RDB, Squire JM (2008) Fifty years of coiled-coils and alpha-helical bundles: A close relationship between sequence and structure. Journal of structural biology.
2. Craig AW, Zirngibl R, Greer P (1999) Disruption of coiled-coil domains in Fer protein-tyrosine kinase abolishes trimerization but not kinase activation. The Journal of biological chemistry 274: 19934-19942.
3. Kilmartin JV, Dyos SL, Kershaw D, Finch JT (1993) A spacer protein in the Saccharomyces cerevisiae spindle poly body whose transcript is cell cycle-regulated. The Journal of cell biology 123: 1175-1184.
4. Rose A, Meier I (2004) Scaffolds, levers, rods and springs: diverse cellular functions of long coiled-coil proteins. Cellular and molecular life sciences : CMLS 61: 1996-2009.
5. Gruber M, Söding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. J Struct Biol 155: 140-145.
6. Apgar JR, Gutwin KN, Keating AE (2008) Predicting helix orientation for coiled-coil dimers. Proteins 72: 1048-1065.
7. Wolf E, Kim PS, Berger B (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein science : a publication of the Protein Society 6: 1179-1189.
8. Woolfson DN, Alber T (1995) Predicting oligomerization states of coiled coils. Protein science : a publication of the Protein Society 4: 1596-1607.
9. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. Nucleic acids research 37.
10. Walshaw J, Woolfson DN (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J Mol Biol 307: 1427-1450.
11. McDonnell AV, Jiang T, Keating AE, Berger B (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 22: 356-358.
12. Henrick K, Thornton JM (1998) PQS: a protein quaternary structure file server. Trends Biochem Sci 23: 358-361.
13. Hubbard T, Murzin A, Brenner S, Chothia C (1997) SCOP: a structural classification of proteins database. Nucleic Acids Research 25: 236-239.
14. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, et al. (2005) Protein database searches using compositionally adjusted substitution matrices. FEBS J 272: 5101-5109.
15. Suzek B, Huang H, McGarvey P, Mazumder R, Wu C (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23: 1282-1288.
16. Krause A, Stoye J, Vingron M (2005) Large scale hierarchical clustering of protein sequences. BMC Bioinformatics 6.
17. Strambio-de-Castillia C, Blobel G, Rout MP (1999) Proteins connecting the nuclear pore complex with the nuclear interior. The Journal of cell biology 144: 839-855.
18. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, et al. (1995) Predicting coiled coils by use of pairwise residue correlations. Proceedings of the National Academy of Sciences of the United States of America 92: 8259-8263.
19. James M (1985) Classification algorithms. New York: John Wiley and Sons.

20. Holland RCG, Down T, Pocock M, Prlic A, Huen D, et al. (2008) BioJava: an Open-Source Framework for Bioinformatics. Bioinformatics (Oxford, England).
21. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science (New York, NY) 252: 1162-1164.
22. Delorenzi M, Speed T (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18: 617-625.
23. Mack G, Compton D (2001) Analysis of mitotic microtubule-associated proteins using mass spectrometry identifies astrin, a spindle-associated protein. Proceedings of the National Academy of Sciences of the United States of America 98: 14434-14439.
24. Gruber J, Harborth J, Schnabel J, Weber K, Hatzfeld M (2002) The mitotic-spindle-associated protein astrin is essential for progression through mitosis. J Cell Sci 115: 4053-4059.
25. Tai LJ, McFall SM, Huang K, Demeler B, Fox SG, et al. (2002) Structure-function analysis of the heat shock factor-binding protein reveals a protein composed solely of a highly conserved and dynamic coiled-coil trimerization domain. The Journal of biological chemistry 277: 735-745.
26. Martin SL, Li J, Weisz JA (2000) Deletion analysis defines distinct functional domains for protein-protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1. Journal of molecular biology 304: 11-20.
27. Frank S, Schulthess T, Landwehr R, Lustig A, Mini T, et al. (2002) Characterization of the matrilin coiled-coil domains reveals seven novel isoforms. The Journal of biological chemistry 277: 19071-19079.
28. Agou F, Ye F, Goffinont S, Courtois G, Yamaoka S, et al. (2002) NEMO trimerizes through its coiled-coil C-terminal domain. The Journal of biological chemistry 277: 17464-17475.
29. Harborth J, Weber K, Osborn M (1995) Epitope mapping and direct visualization of the parallel, in-register arrangement of the double-stranded coiled-coil in the NuMA protein. The EMBO journal 14: 2447-2460.
30. Alfadhli A, Steel E, Finlay L, Bächinger HP, Barklis E (2002) Hantavirus nucleocapsid protein coiled-coil domains. The Journal of biological chemistry 277: 27103-27108.
31. Wigge PA, Jensen ON, Holmes S, Souès S, Mann M, et al. (1998) Analysis of the Saccharomyces spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. The Journal of cell biology 141: 967-977.
32. Kammerer RA, Schulthess T, Landwehr R, Lustig A, Fischer D, et al. (1998) Tenascin-C hexabrachion assembly is a sequential two-step process initiated by coiled-coil alpha-helices. The Journal of biological chemistry 273: 10602-10608.
33. Hase ME, Kuznetsov NV, Cordes VC (2001) Amino acid substitutions of coiled-coil protein Tpr abrogate anchorage to the nuclear pore complex but not parallel, in-register homodimerization. Molecular biology of the cell 12: 2433-2452.
34. Misenheimer TM, Huwiler KG, Annis DS, Mosher DF (2000) Physical characterization of the procollagen module of human thrombospondin 1 expressed in insect cells. The Journal of biological chemistry 275: 40938-40945.
35. Conway J, Parry D (1991) Three-stranded α-fibrous proteins: the heptad repeat and its implications for structure. International Journal of Biological Macromolecules 13:

14-16.
36. Kammerer R, Kostrewa D, Progias P, Honnappa S, Avila D, et al. (2005) A conserved trimerization motif controls the topology of short coiled coils. Proceedings of the National Academy of Sciences of the United States of America 102: 13891-13896.
37. Wu C (1995) Heat Shock Transcription Factors: Structure and Regulation. Annual Review of Cell and Developmental Biology 11: 441-469.
38. Lee DL, Lavigne P, Hodges RS (2001) Are trigger sequences essential in the folding of two-stranded alpha-helical coiled-coils? Journal of molecular biology 306: 539-553.
39. Testa OD, Moutevelis E, Woolfson DN (2008) CC+: a relational database of coiled-coil structures. Nucleic acids research.
40. Passner JM, Schultz SC, Steitz TA (2000) Modeling the cAMP-induced allosteric transition using the crystal structure of CAP-cAMP at 2.1 A resolution. J Mol Biol 304: 847-859.
41. Galkin V, Yu X, Bielnicki J, Heuser J, Ewing C, et al. (2008) Divergence of Quaternary Structures Among Bacterial Flagellar Filaments. Science 320: 382-385.
42. Harbury PB, Zhang T, Kim PS, Alber T (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science 262: 1401-1407.
43. Ramos J, Lazaridis T (2006) Energetic Determinants of Oligomeric State Specificity in Coiled Coils. Journal of the American Chemical Society 128: 15499-15510.

Chapter 3

Predicting helix orientation for coiled-coil dimers

Author Contributions

This work was prepared in collaboration with James R. Apgar and Amy E. Keating. J.R.A. prepared explicit structure models, implemented the antiparallel Crick parameterization and computed structure-based statistics.

Attribution

This chapter is republished with permission from John Wiley & Sons, Inc. from: James R. Apgar, Karl N. Gutwin and Amy E. Keating. *Proteins* **72**, 1048-1065 (2006).

3.1 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

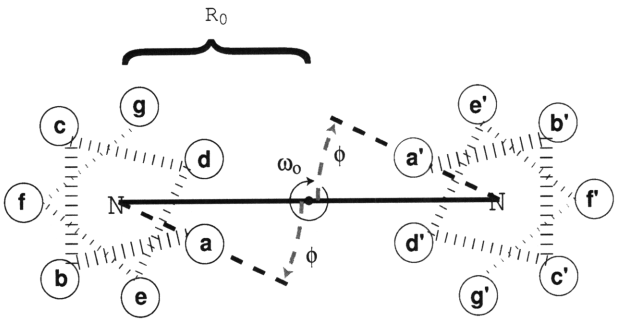
alignment. Because of the near degeneracy of many of these variants, coiled coils pose a challenge to fold recognition methods for structure prediction. Whereas distinctions between some protein folds can be discriminated on the basis of hydrophobic/polar patterning or secondary structure propensities, 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 asses the ability of five energy functions to recognize the correct fold. We also developed and tested three sequenced-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.

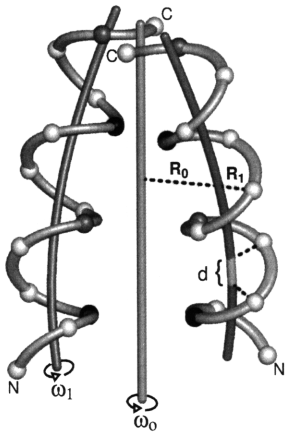
3.2 Introduction

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 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 (Figure 3-1a 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.[1,2,3,4] However, few methods exist to predict the variety of topologies found in coiled-coil structures.[4,5,6] 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]

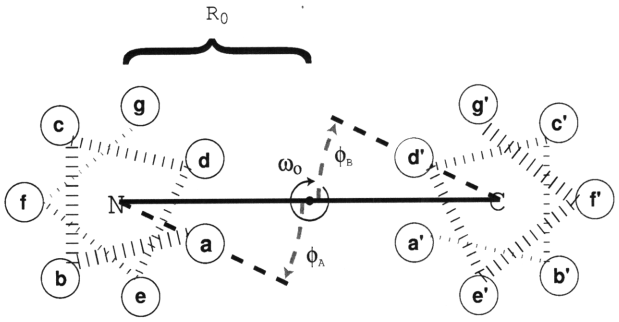
The difficulty of predicting coiled-coil structure lies in differentiating what can be subtle distinctions in interactions. For example, it has been reported for several designed coiled coils that changing a single **a**- or **d**-position residue can lead to a change or loss of oligomerization specificity.[8,9,10] 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

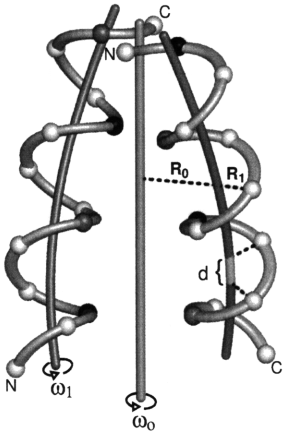
(a)

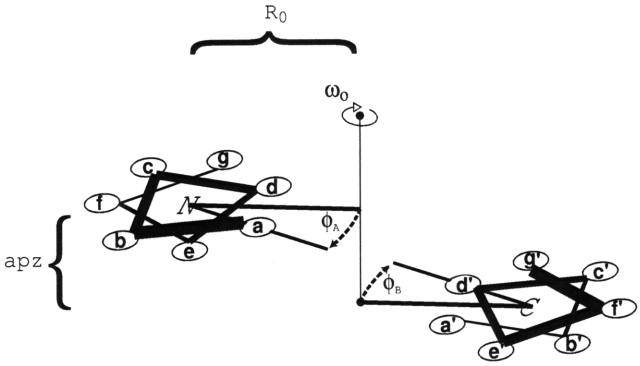

R0
g
c
d
e'
b'
f
N
ω0
ϕ
a'
f'
a
d'
b
e
g'
c'


C
C
R0
R1
d
N
N
ω1
ω0

(b)


R0
g
c
d
g'
c'
f
N
ω0
ϕB
d'
C
f'
a
ϕA
a'
b
e
e'
b'


C
N
R0
R1
d
C
N
ω1
ω0


R0
ω0
apz
c
g
d
f
N
a
ϕA
b
e
ϕB
g'
c'
d'
C
f'
a'
e'
b'

(c)

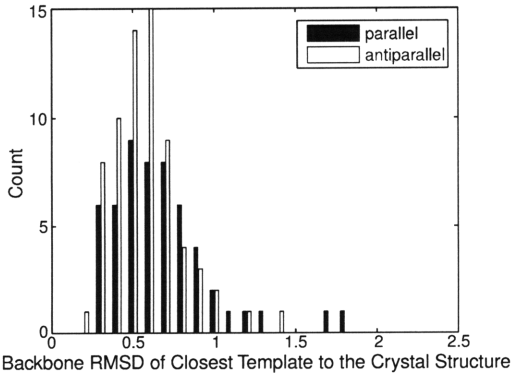

15
10
5
0
Count
parallel
antiparallel
0
0.5
1
1.5
2
2.5
Backbone RMSD of Closest Template to the Crystal Structure

Figure 3-1. 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. (c) 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.

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.[6,18,19,20,21] 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.[11,22,23,24,25,26,27] 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 paper, 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.[1,29,30,31,32] 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.

3.3 Methods

3.3.1 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.[34] 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.[35] 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 can be determined by sequence alignment[21,36]: 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.

3.3.2 Crick Parameterization

To describe and generate parallel coiled-coil dimer backbones, we used the parameterization originally proposed by Crick and subsequently implement by Harbury et al. as a user routine in CHARMM.[41,42] This parameterization has been shown to closely mimic the geometry of several parallel coiled coils.[41] 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 (supporting data in Appendix B, Figure B-1).

We modified the Crick/Harbury approach to describe and generate antiparallel coiled-coil backbones. As in the fitcc program (Personal Communication Tom Alber; Author Mark Sales http://ucxray.berkeley.edu/~mark/fitcc.html), 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. We modified the coiled-coil parameterization to account 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. We re-write the parameterization for antiparallel coiled coils as:

$$CC(\tau) = EC'(\tau) + H(\tau)$$

$$E(\omega_0\tau, \alpha, 0) = \begin{pmatrix} \cos(\omega_0\tau) & -\sin(\omega_0\tau)\cos(\alpha) & 0 \\ \sin(\omega_0\tau) & \cos(\omega_0\tau)\cos(\alpha) & 0 \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

$$C' = \begin{pmatrix} R_1\cos(\omega_1\tau + \phi_i) \\ R_1\sin(\omega_1\tau + \phi_i) \\ apz_i \end{pmatrix}$$

$$H(\tau) = \begin{pmatrix} R_0\cos(\omega_0\tau) \\ R_0\sin(\omega_0\tau) \\ d\cos(\alpha) \end{pmatrix}$$

$$where \quad \sin(\alpha) = \frac{R_0\omega_0}{d}$$

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 α-helix radius and ω_1 is the α-helix frequency. As described above, apz_i is an axial helical 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.

We used the Crick parameterization both to fit idealized backbones to native structures and to generate *de novo* backbones. To fit a native structure, we 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, we 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[65] 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.

3.3.3 Generation of 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.[45] Parameters R_1, ω_1 and d, which describe α-helix geometry, were set to 2.26 Å, $4\pi/7$ radians per residue and 1.52 Å respectively.[41] 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 Appendix B, Figures B-2 and B-3. ϕ_A, ϕ_B values were sampled as pairs due to correlations between these in native structures (Appendix B, Figure B-4).

3.3.4 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.

3.3.5 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.[66] 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.[47,67] 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,[68] 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:

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

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.

3.3.6 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 3-2. Energy components used in Figure 3-3 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.

3.4 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.[14,37,38,39,40] 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 3-1 and a detailed list is available in Appendix B, Table B-1.

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

Table 3-1. Test set of coiled-coil dimers of known orientation.

	Sequence Pairs	**Avg. Length (range) in residues**	**Number of intra-molecular 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. [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 Appendix B, Figure B-1.

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 (Appendix B, Figures B-1 through B-4). 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 3-1c).

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 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.

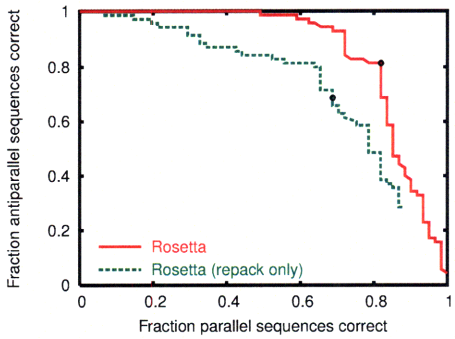
Finally, ESMs are more generalizable in that they can potentially be applied equally to any structure; ISMs must be created specifically for the structure to be modeled.

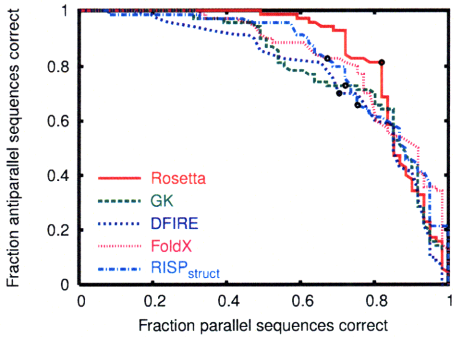
3.4.1 Performance of explicit structure models

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). [43,44,45] The third step gave rise to differences between models, with each ESM named according to the energy function used at this stage.

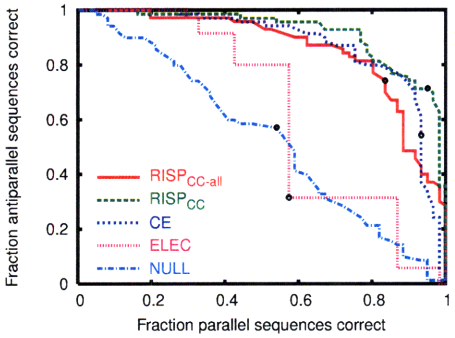
In a preliminary set of calculations, we tested two structure-prediction methods for use in step 2. Initially, Rosetta was simply used to place side chains into preferred conformations on each of 81 parallel and 120 antiparallel idealized Crick backbones. When Rosetta was used to select the lowest-energy 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 3-2a (left panel). Results are plotted as the fraction of antiparallel sequences

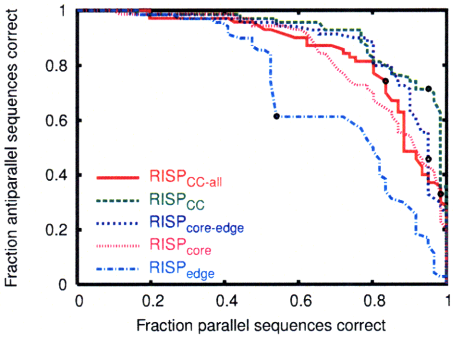
(a)


Fraction antiparallel sequences correct
Fraction parallel sequences correct
Rosetta
Rosetta (repack only)

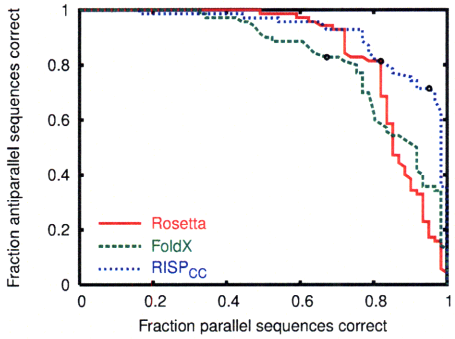

Fraction antiparallel sequences correct
Fraction parallel sequences correct
Rosetta
GK
DFIRE
FoldX
RISP$_{struct}$

(b)

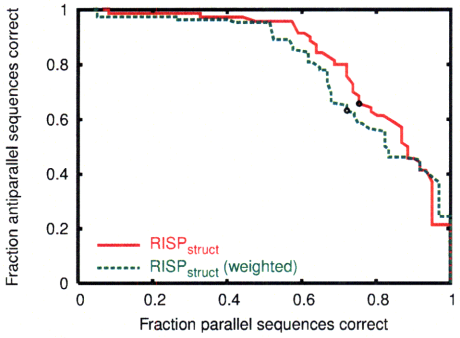

Fraction antiparallel sequences correct
Fraction parallel sequences correct
RISP$_{CC-all}$
RISP$_{CC}$
CE
ELEC
NULL

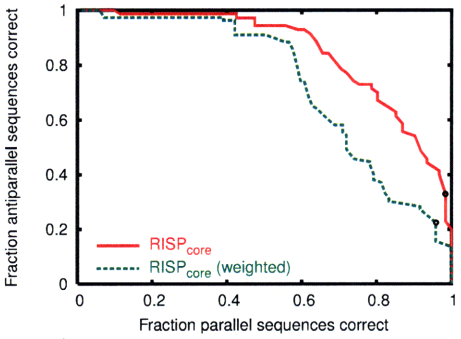

Fraction antiparallel sequences correct
Fraction parallel sequences correct
RISP$_{CC-all}$
RISP$_{CC}$
RISP$_{core-edge}$
RISP$_{core}$
RISP$_{edge}$

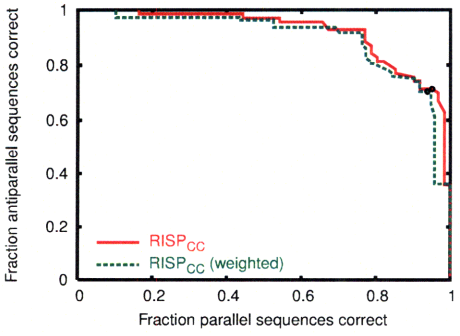
(c)


Fraction antiparallel sequences correct
Fraction parallel sequences correct
Rosetta
FoldX
RISP$_{CC}$

(d)


Fraction antiparallel sequences correct
Fraction parallel sequences correct
RISP$_{struct}$
RISP$_{struct}$ (weighted)


Fraction antiparallel sequences correct
Fraction parallel sequences correct
RISP$_{core}$
RISP$_{core}$ (weighted)


Fraction antiparallel sequences correct
Fraction parallel sequences correct
RISP$_{CC}$
RISP$_{CC}$ (weighted)

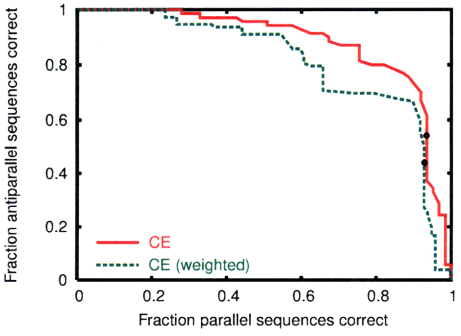

Fraction antiparallel sequences correct
Fraction parallel sequences correct
CE
CE (weighted)

Figure 3-2. 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 (repacked, min) 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 structure (Appendix B, Table B-2).

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 3-2a).

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 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 3-2a).

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 3-2). Figure 3-2d 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.

3.4.2 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 structure. Heptad assignment 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 3-2. To approximate the $RISP_{struct}$ method using an

Table 3-2: 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; n 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.

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 3-2b). 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 3-2c).

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[50,51,52,53] 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'** pairs (antiparallel), giving model $RISP_{core}$, or only edge **g-e'** or **g-g'** (parallel) or **e-e'** (antiparallel) pairs, giving $RISP_{edge}$, degraded performance (Figure 3-2b).

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 3-2b). The performance of two control models is also shown in Figure 3-2b. Model ELEC scores only the **e-** and **g-**position electrostatic complementarity and did not provide good parallel vs. antiparallel 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 3-2d). 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 3-2d). 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 unequal numbers of weights available for scoring homo vs. heterodimers.

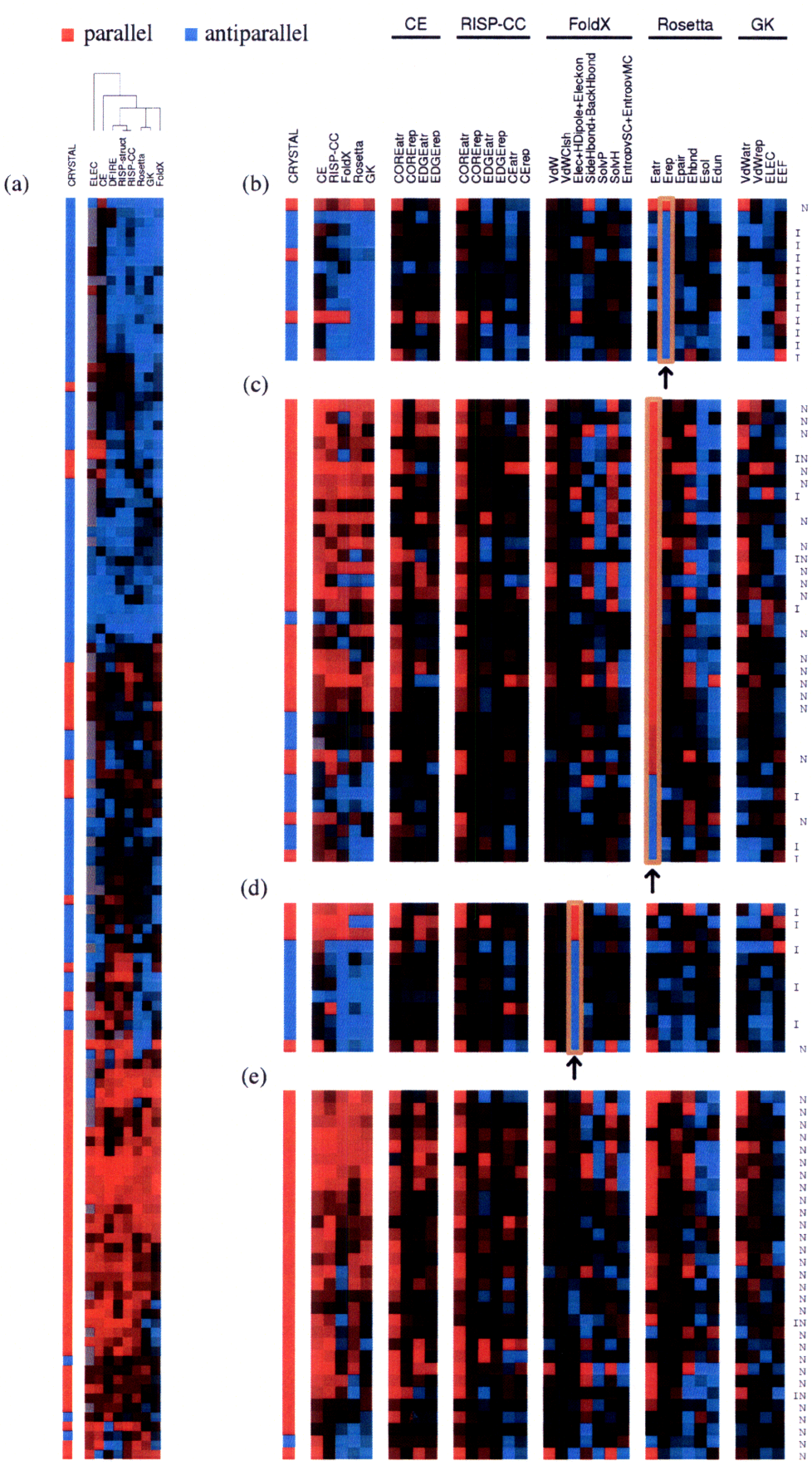
3.4.3 Analysis

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 3-3a 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 3-2b). Figure 3-3a shows that

Figure 3-3. 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 Appendix B, Table B-3.


parallel
antiparallel
(a)
CRYSTAL
ELEC
CE
DFIRE
RISP-Struct
RISP-CC
Rosetta
GK
FoldX
(b)
CRYSTAL
CE
RISP-CC
FoldX
Rosetta
GK
CE
RISP-CC
FoldX
Rosetta
GK
COREatr
CORErep
EDGEatr
EDGErep
CEatr
CErep
VdW
VdWClsh
Elec+HDipole+Eleckon
SideHbond+BackHbond
SolvP
SolvH
EntropySC+EntropyMC
Eatr
Erep
Epair
Ehbnd
Esol
Edun
VdWatr
VdWrep
ELEC
EEF
(c)
(d)
(e)

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 prediction rate it was similar to the very good $RISP_{CC}$ (also an ISM). Yet, the clustering in Figure 3-3a 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 (Appendix B, Figure B-5a). 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 Appendix B, Figure B-5b). 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 3-4

(descriptions of components are included in Appendix B, Table B-3). Figure 3-4 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 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.

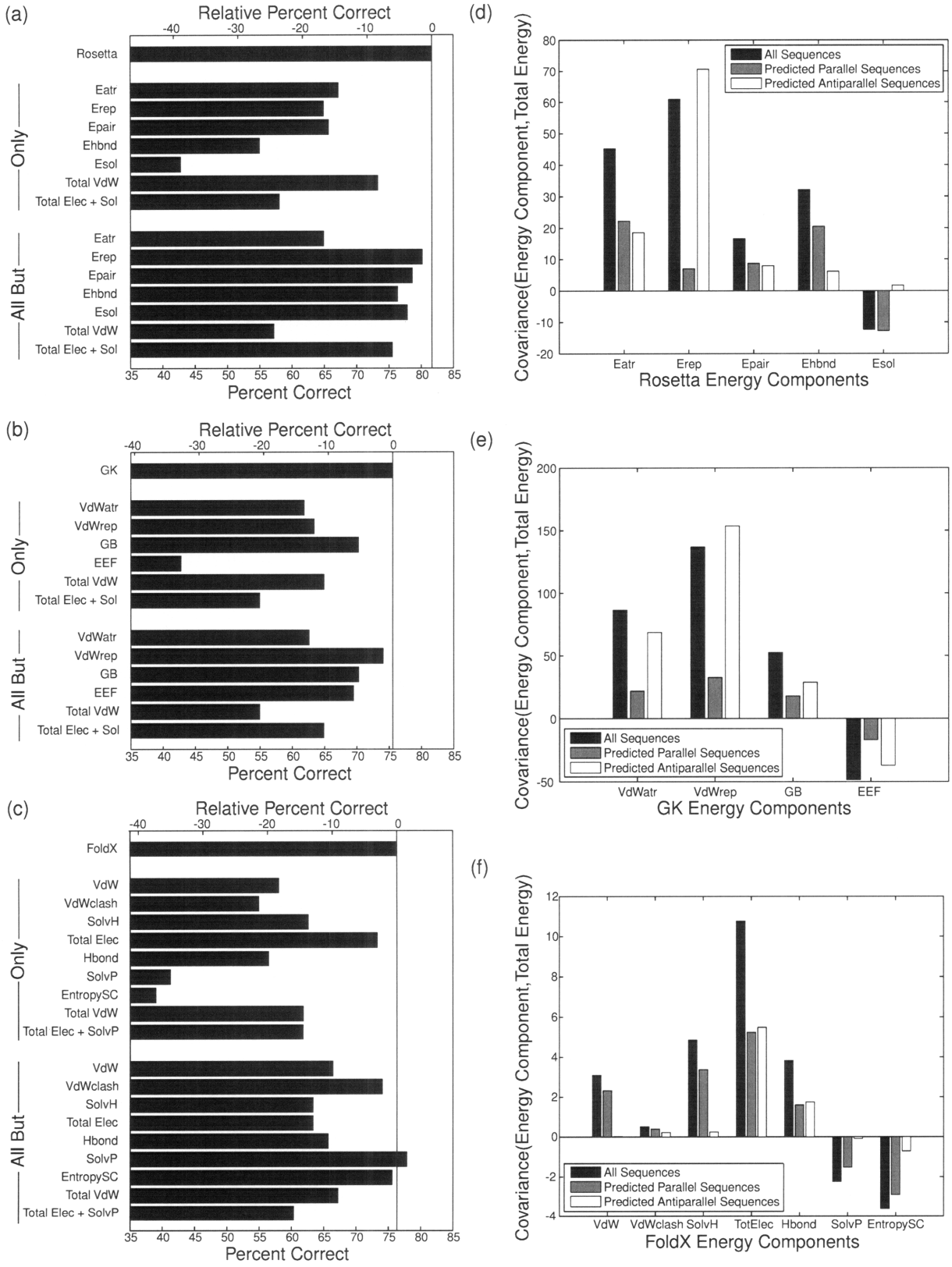

(a)
Relative Percent Correct
-40 -30 -20 -10 0
Rosetta
Only
Eatr
Erep
Epair
Ehbnd
Esol
Total VdW
Total Elec + Sol
All But
Eatr
Erep
Epair
Ehbnd
Esol
Total VdW
Total Elec + Sol
35 40 45 50 55 60 65 70 75 80 85
Percent Correct
(b)
Relative Percent Correct
-40 -30 -20 -10 0
GK
Only
VdWatr
VdWrep
GB
EEF
Total VdW
Total Elec + Sol
All But
VdWatr
VdWrep
GB
EEF
Total VdW
Total Elec + Sol
35 40 45 50 55 60 65 70 75 80 85
Percent Correct
(c)
Relative Percent Correct
-40 -30 -20 -10 0
FoldX
Only
VdW
VdWclash
SolvH
Total Elec
Hbond
SolvP
EntropySC
Total VdW
Total Elec + SolvP
All But
VdW
VdWclash
SolvH
Total Elec
Hbond
SolvP
EntropySC
Total VdW
Total Elec + SolvP
35 40 45 50 55 60 65 70 75 80 85
Percent Correct
(d)
Covariance(Energy Component,Total Energy)
80 70 60 50 40 30 20 10 0 -10 -20
All Sequences
Predicted Parallel Sequences
Predicted Antiparallel Sequences
Eatr Erep Epair Ehbnd Esol
Rosetta Energy Components
(e)
Covariance(Energy Component,Total Energy)
200 150 100 50 0 -50
All Sequences
Predicted Parallel Sequences
Predicted Antiparallel Sequences
VdWatr VdWrep GB EEF
GK Energy Components
(f)
Covariance(Energy Component,Total Energy)
12 10 8 6 4 2 0 -2 -4
All Sequences
Predicted Parallel Sequences
Predicted Antiparallel Sequences
VdW VdWclash SolvH TotElec Hbond SolvP EntropySC
FoldX Energy Components

Figure 3-4. 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 + Eref, 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.

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 3-4d 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 3-4f), consistent with the analysis of Figure 3-4c. 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 3-3 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 3-3b 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 3-3b 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 3-5). 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 3-3c 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 3-4f). Figure 3-3d shows examples that had large contributions from

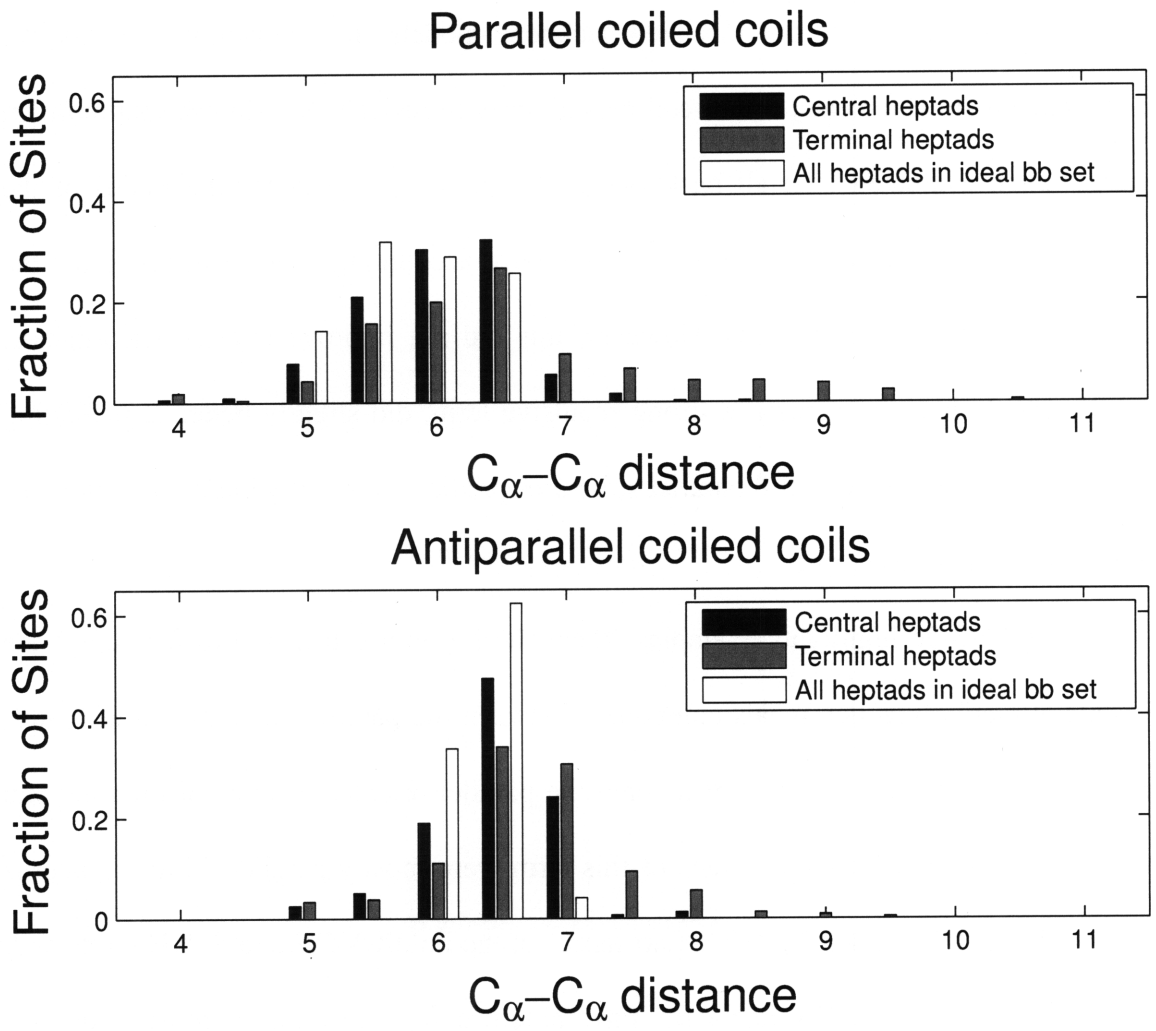


Figure 3-5. Distribution of C_α-C_α distances for core residues in parallel and antiparallel coiled coils. All C_α-C_α distances between core residues (**a-a'**, **d-d'** in parallel and **a-d'** in antiparallel) were binned by distance. For the test-set structures, residues were divided into two sets: Central heptads (black) include positions that are not the first or last seven residues of a coiled-coil helix, and terminal heptads (gray) include residues that are 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.

FoldX electrostatics (Elec, HDipole and Eleckon); these terms more often favored antiparallel structures. The GK potential 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 3-4d 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 3-3e). 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.

3.4.4 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 examples unclassified. Figure 3-6 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.

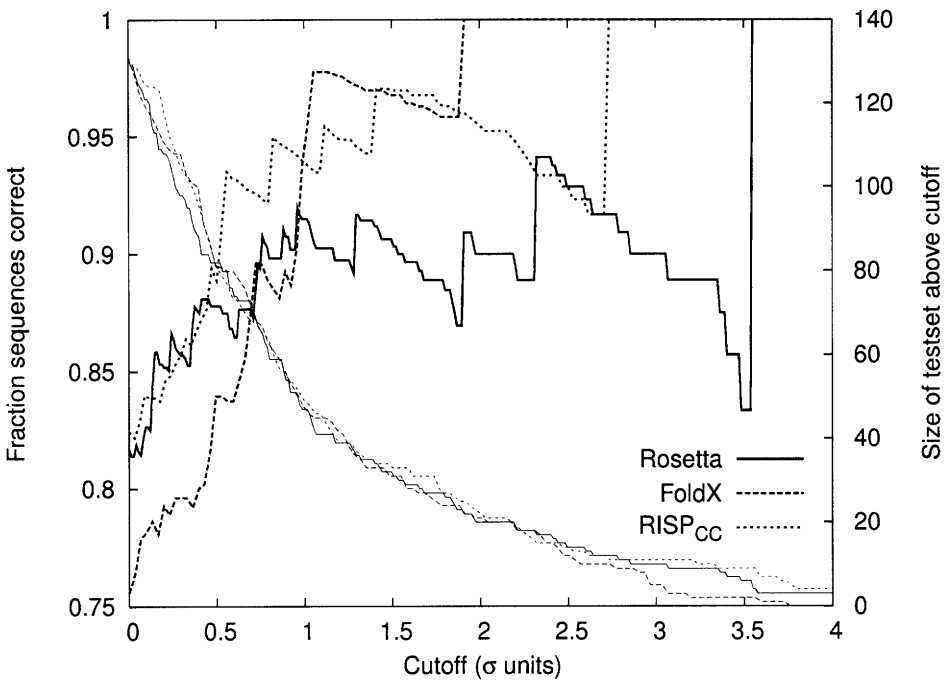


Figure 3-6. Performance as a function of increasing the gap requirement. Performance was evaluated only for those examples with |Eparallel –Eantiparallel| > x*σ and is plotted (thick lines, left axis) as a function of x. The size of the test set at each value of x is plotted using thin lines and the right axis.

3.5 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 3-2a) 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 3-5). 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 3-3b). 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 3-3b). 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 3-3 and 3-4). 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,[50,51,52,53] 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 (Appendix B, Figure B-5a). Interestingly, including the 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,[60,61,62] 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 stoichiometry. 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.

3.6 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.

3.7 References

1. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, et al. (1995) Predicting coiled coils by use of pairwise residue correlations. Proceedings of the National Academy of Sciences of the United States of America 92: 8259-8263.
2. Delorenzi M, Speed T (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18: 617-625.
3. McDonnell AV, Jiang T, Keating AE, Berger B (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 22: 356-358.
4. Wolf E, Kim PS, Berger B (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci 6: 1179-1189.
5. Woolfson DN, Alber T (1995) Predicting oligomerization states of coiled coils. Protein Sci 4: 1596-1607.
6. Fong JH, Keating AE, Singh M (2004) Predicting specificity in bZIP coiled-coil protein interactions. Genome Biol 5.
7. Lupas AN, Gruber M (2005) The structure of alpha-helical coiled coils. Adv Protein Chem 70: 37-78.
8. Tripet B, Wagschal K, Lavigne P, Mant CT, Hodges RS (2000) 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". J Mol Biol 300: 377-402.
9. Wagschal K, Tripet B, Lavigne P, Mant C, Hodges RS (1999) 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 8: 2312-2329.
10. Liu J, Zheng Q, Deng Y, Kallenbach NR, Lu M (2006) Conformational transition between four and five-stranded phenylalanine zippers determined by a local packing interaction. J Mol Biol 361: 168-179.
11. Oakley MG, Kim PS (1998) A buried polar interaction can direct the relative orientation of helices in a coiled coil. Biochemistry 37: 12603-12610.
12. Lumb KJ, Kim PS (1995) A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. Biochemistry 34: 8642-8648.
13. Harbury PB, Zhang T, Kim PS, Alber T (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science 262: 1401-1407.
14. Taylor CM, Keating AE (2005) Orientation and oligomerization specificity of the Bcr coiled-coil oligomerization domain. Biochemistry 44: 16246-16256.
15. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, et al. (2003) An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 53: 76-87.
16. Kihara D, Lu H, Kolinski A, Skolnick J (2001) TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A 98: 10125-10130.
17. Vieth M, Kolinski A, Brooks CL, 3rd, Skolnick J (1995) Prediction of quaternary structure of coiled coils. Application to mutants of the GCN4 leucine zipper. J Mol Biol 251: 448-467.

18. Grigoryan G, Keating AE (2006) Structure-based prediction of bZIP partnering specificity. J Mol Biol 355: 1125-1142.
19. Mason JM, Schmitz MA, Muller KM, Arndt KM (2006) 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 103: 8989-8994.
20. Fassler J, Landsman D, Acharya A, Moll JR, Bonovich M, et al. (2002) B-ZIP proteins encoded by the Drosophila genome: evaluation of potential dimerization partners. Genome Res 12: 1190-1200.
21. Vinson C, Myakishev M, Acharya A, Mir AA, Moll JR, et al. (2002) Classification of human B-ZIP proteins based on dimerization properties. Mol Cell Biol 22: 6321-6335.
22. Gurnon DG, Whitaker JA, Oakley MG (2003) Design and characterization of a homodimeric antiparallel coiled coil. J Am Chem Soc 125: 7518-7519.
23. McClain DL, Woods HL, Oakley MG (2001) Design and characterization of a heterodimeric coiled coil that forms exclusively with an antiparallel relative helix orientation. J Am Chem Soc 123: 3151-3152.
24. Monera OD, Kay CM, Hodges RS (1994) Electrostatic interactions control the parallel and antiparallel orientation of alpha-helical chains in two-stranded alpha-helical coiled-coils. Biochemistry 33: 3862-3871.
25. Monera OD, Zhou NE, Lavigne P, Kay CM, Hodges RS (1996) Formation of parallel and antiparallel coiled-coils controlled by the relative positions of alanine residues in the hydrophobic core. J Biol Chem 271: 3995-4001.
26. Myszka DG, Chaiken IM (1994) Design and characterization of an intramolecular antiparallel coiled coil peptide. Biochemistry 33: 2363-2372.
27. Schnarr NA, Kennan AJ (2004) Strand orientation by steric matching: a designed antiparallel coiled-coil trimer. J Am Chem Soc 126: 14447-14451.
28. Gernert KM, Surles MC, Labean TH, Richardson JS, Richardson DC (1995) The Alacoil: a very tight, antiparallel coiled-coil of helices. Protein Sci 4: 2252-2260.
29. Berger B, Singh M (1997) An iterative method for improved protein structural motif recognition. J Comput Biol 4: 261-273.
30. Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, et al. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J Mol Biol 343: 703-718.
31. Obenauer JC, Cantley LC, Yaffe MB (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31: 3635-3641.
32. Brannetti B, Via A, Cestra G, Cesareni G, Helmer-Citterich M (2000) SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. J Mol Biol 298: 313-328.
33. McClain DL, Binfet JP, Oakley MG (2001) Evaluation of the energetic contribution of interhelical Coulombic interactions for coiled coil helix orientation specificity. J Mol Biol 313: 371-383.
34. Walshaw J, Woolfson DN (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J Mol Biol 307: 1427-1450.
35. Walshaw J, Woolfson DN (2003) Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. J Struct Biol 144: 349-361.

36. Newman JR, Keating AE (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science 300: 2097-2101.
37. Supekar VM, Bruckmann C, Ingallinella P, Bianchi E, Pessi A, et al. (2004) Structure of a proteolytically resistant core from the severe acute respiratory syndrome coronavirus S2 fusion protein. Proc Natl Acad Sci U S A 101: 17958-17963.
38. Strelkov SV, Schumacher J, Burkhard P, Aebi U, Herrmann H (2004) Crystal structure of the human lamin A coil 2B dimer: implications for the head-to-tail association of nuclear lamins. J Mol Biol 343: 1067-1080.
39. Oakley MG, Kim PS (1997) Protein dissection of the antiparallel coiled coil from Escherichia coli seryl tRNA synthetase. Biochemistry 36: 2544-2549.
40. Lumb KJ, Carr CM, Kim PS (1994) Subdomain folding of the coiled coil leucine zipper from the bZIP transcriptional activator GCN4. Biochemistry 33: 7361-7367.
41. Harbury PB, Tidor B, Kim PS (1995) Repacking protein cores with backbone freedom: structure prediction for coiled coils. Proc Natl Acad Sci U S A 92: 8408-8412.
42. Crick FH (1953) The Fourier Transform of a Coiled-Coil. Acta Cryst 6: 685-689.
43. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97: 10383-10388.
44. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302: 1364-1368.
45. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, et al. (1983) CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comp Chem 4: 187-217.
46. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. Proteins 35: 133-152.
47. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320: 369-387.
48. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714-2726.
49. Lu H, Lu L, Skolnick J (2003) Development of unified statistical potentials describing protein-protein interactions. Biophys J 84: 1895-1901.
50. Krylov D, Mikhailenko I, Vinson C (1994) A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. The EMBO journal 13: 2849-2861.
51. Moitra J, Szilak L, Krylov D, Vinson C (1997) Leucine is the most stabilizing aliphatic amino acid in the d position of a dimeric leucine zipper coiled coil. Biochemistry 36: 12567-12573.
52. Acharya A, Ruvinov SB, Gal J, Moll JR, Vinson C (2002) A heterodimerizing leucine zipper coiled coil system for examining the specificity of a position interactions: amino acids I, V, L, N, A, and K. Biochemistry 41: 14122-14131.
53. Acharya A, Rishi V, Vinson C (2006) 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 45: 11324-11332.

54. Oakley MG, Hollenbeck JJ (2001) The design of antiparallel coiled coils. Curr Opin Struct Biol 11: 450-457.
55. McClain DL, Gurnon DG, Oakley MG (2002) Importance of potential interhelical salt-bridges involving interior residues for coiled-coil stability and quaternary structure. J Mol Biol 324: 257-270.
56. Campbell KM, Sholders AJ, Lumb KJ (2002) Contribution of buried lysine residues to the oligomerization specificity and stability of the fos coiled coil. Biochemistry 41: 4866-4871.
57. Hadley EB, Gellman SH (2006) An antiparallel alpha-helical coiled-coil model system for rapid assessment of side-chain recognition at the hydrophobic interface. J Am Chem Soc 128: 16444-16445.
58. Harbury PB, Zhang T, Kim PS, Alber T (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science 262: 1401-1407.
59. Hadley EB, Testa OD, Woolfson DN, Gellman SH (2008) Preferred Side-chain Constellation at Antiparallel Coiled-Coil Interfaces. Proc Natl Acad Sci U S A 105: 530-535.
60. Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci U S A 99: 5896-5901.
61. Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 19: 161-162.
62. Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins 49: 350-364.
63. Grigoryan G, Zhou F, Lustig SR, Ceder G, Morgan D, et al. (2006) Ultra-fast evaluation of protein energies directly from sequence. PLoS Comput Biol 2: e63.
64. Zhou F, Grigoryan G, Lustig SR, Keating AE, Ceder G, et al. (2005) Coarse-graining protein energetics in sequence variables. Phys Rev Lett 95: 148103.
65. (2005) Matlab R14. The MathWorks, Inc.
66. Beroza P, Fredkin DR (1996) Calculation of amino acid pK(a)s in a protein from a continuum electrostatic model: Method and sensitivity analysis. Journal of Computational Chemistry 17: 1229-1244.
67. Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, et al. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proc Natl Acad Sci U S A 102: 10147-10152.
68. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2: e155.

Chapter 4

Structure-based approaches to the prediction of coiled-coil alignment

4.1 Introduction

One important aspect of the folding of fibrous proteins with periodic repeats is the axial registration or alignment of the component strands. The repeating sequence pattern in these proteins implies that many possible relative alignments will have similar structures and thus similar stabilities. However, these repetitive proteins usually form one stable, specific alignment, which is likely to be determined by some part of the protein's sequence. Therefore, much work has been done to study the determinants of proper alignment in this class of proteins.

Collagen consists of a three-stranded rope of chains containing a repetitive pattern of proline and glycine residues[1]. These chains normally assemble in a fully-overlapping

register, which is defined primarily by their C-terminal globular domains, as removal of these domains causes non-specific chain registration and the formation of a gel[2]. Interestingly, in the collagens, there appears to be very little specificity for proper alignment within the fibrous regions themselves[2].

Coiled coils are another large class of repetitive proteins observed in a wide range of lengths and structures. Tropomyosin, a long two-stranded parallel coiled coil, has been shown to fold via a two-state mechanism in which binding in the proper alignment occurs rapidly[3]. Certain subregions of sequence have been shown to be important for this process and for determining partnering specificity[4,5]. However, many more coiled coils exist for which the relative alignment is not well known. Therefore, methods of predicting coiled-coil alignment would be quite valuable.

Many approaches have been developed to predict various aspects of the structural specificity of coiled-coil-forming sequence, including predictions of oligomerization state[6], helix orientation[7] and partnering preference[8]. However, the prediction of relative helix alignment has not yet been systematically addressed. In most cases, there are no clear rules that dictate coiled-coil alignment, as the energetic contributions of many residues are expected to contribute to structural specificity, making this a complex problem[9].

There is very little prior work on predicting coiled-coil alignment. Parry et al. used a rational charge-patterning model to predict the helix alignment and connectivity of the spectrin superfamily by evaluating all possible topologies of the three known component helices[10]. This work was later experimentally verified by Yan et al.[11] Also, as part of validating their machine-learning-based model for predicting parallel

dimeric coiled coils, Singh and Kim tested its ability to recognize the alignment of keratins, intermediate filaments, tropomyosins and myosins[12]. Their results indicated over 90% of sequences were aligned correctly according to their most stringent criteria. However, both of these attempts each tested only one approach, and did not consider a structurally diverse test set. Encouraged by our recent progress in predicting helix orientation in coiled-coil dimers[7], we extended our structure-based modeling approach to the prediction of helix alignment in coiled-coil dimers.

Several models to describe coiled-coil stability have been previously developed. These fall into two major classes: explicit structure models (ESMs) and implicit structure models (ISMs)[7]. The ESMs rely on all-atom three-dimensional models of the coiled-coil interaction, while the ISMs leverage a reduced representation that considers only pairwise contacts between interfacial residues. Intriguingly, the ISMs have shown significant prediction performance in tests of helix orientation and coiled-coil partnering, despite a significant reduction in model detail[7,12,13]. However, previous tests tended to be very simple, with a small number of structural states under consideration. As the number of considered states increases, reduced representations may fail to capture the nuances of structural preference. Therefore, in addition to a diverse set of ISMs, we also considered some of the more detailed ESMs, in the hopes of improving performance at the expense of speed.

We tested pre-existing methods using several diverse test sets of known coiled-coil dimers and various performance metrics. Analysis of these results led us to develop new hybrid models that showed improved performance. We describe trends observed in the performance of different models that highlight sequence features that may be

important for determining coiled-coil alignment, and suggest further directions to improve coiled-coil scoring models.

4.2 Methods

4.2.1 Framework

Our structural approach to predicting coiled-coil alignment involves modeling a series of possible alignments as blunt-ended coiled coils, scoring each alignment and making a prediction based on the best scoring (lowest energy) alignment. Figure 4-1 illustrates this process. In order to make the problem practical, we considered a one-dimensional search whereby one of the two helices in the test interaction is held fixed at its known extent, while the partnering sequence is shifted in 7-residue steps relative to the first. At each step, the overlap between the two sequences is taken to be the possible alignment. Overlaps shorter than the length of the known helix are discarded. This approach avoids comparing sequences of differing lengths, which can be a challenge for many scoring models. Each heterodimeric sequence pair from the test set results in two alignment test cases, while homodimers only yield one. Each alignment search is bounded according to boundaries defined independently for each sequence, as described below.

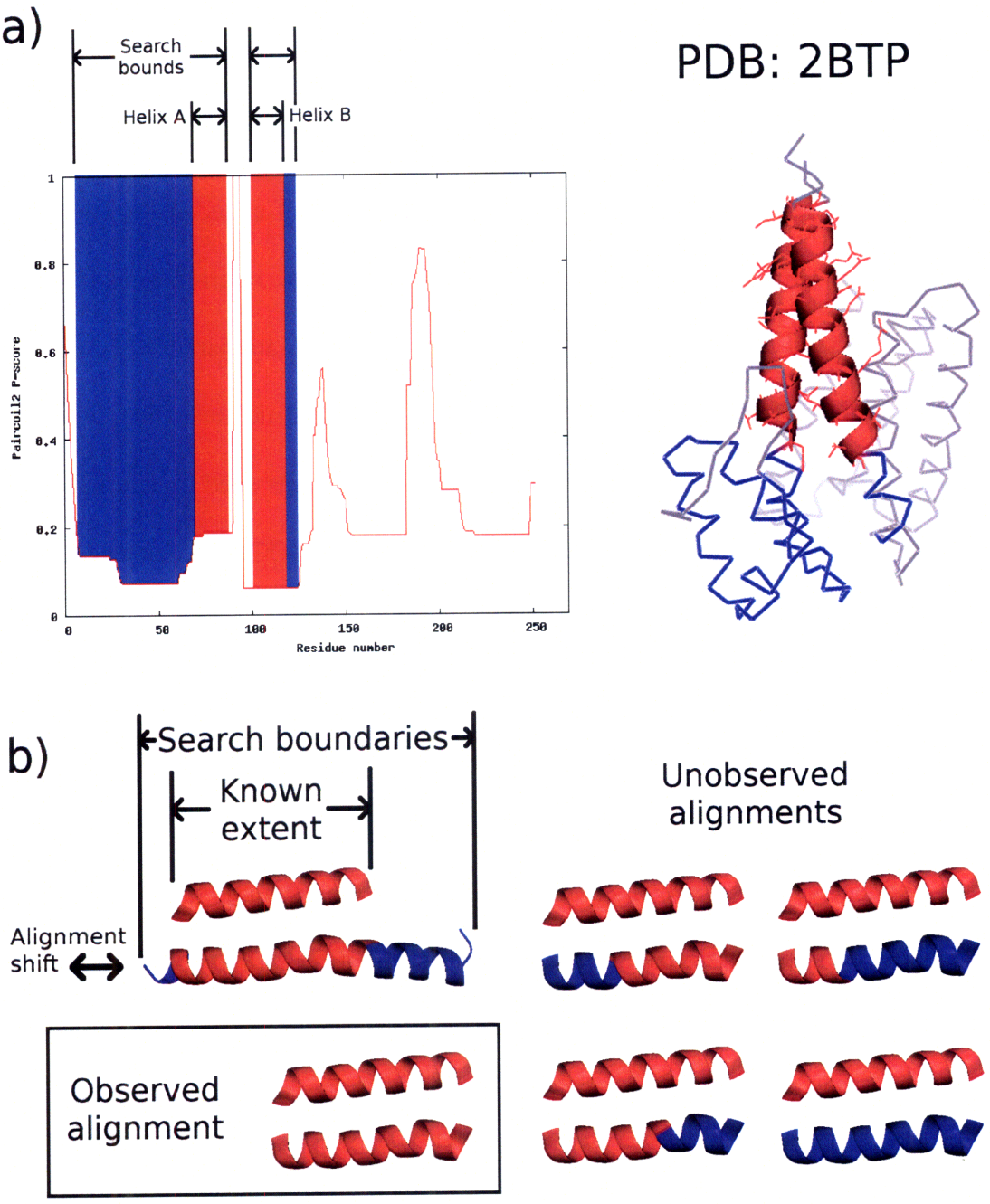


Figure 4-1. Alignment prediction framework. (a) Use of SOCKET and Paircoil2 to assign known extent (red) and search boundaries (blue), respectively. Graph depicts Paircoil2 P-score across the entire protein chain. (b) Illustration of alignment search protocol. Upper helix is fixed to known extent while lower helix shifts in seven-residue increments and blunt-ended overlaps are scored. Blue regions denote sequence not assigned by SOCKET as coiled coil.

4.2.2 Test sets

We constructed four test sets that covered a wide range of coiled coils. The sizes of each test set are provided in Table 4-1. The first two test sets, referred to as "crystal parallel" and "crystal antiparallel", consisted of parallel and antiparallel sequences, respectively, derived from coiled coils detected in crystal structures. The program SOCKET, which detects coiled coils within crystal structures, was used to automatically assign alignments and known extents from crystal structures. SOCKET distance cutoff was 7.0 Å. Test sets were collected according to the protocol described in Chapter 3, without the requirement that the sequence aligns in a blunt-ended fashion in both orientations. Alignment search boundaries were defined by scoring each sequence with Paircoil2[14] and defining the ends of the candidate coiled-coil region as the N- and C-terminal points nearest to the known extent where the P-score increases faster than 0.055 per residue. This cutoff was defined as the 1σ point of the distribution of all per-residue score deltas from a set of known coiled coils. The third test set consisted of sequences from the bZIP family of coiled coils. All bZIP sequences, alignments and partnering data were derived from Newman and Keating [15]. Alignment search boundaries were defined

Name	**Sequence Pairs**	**Test Cases**	**Median (range) alignments per test case**	**Coiled-coil residues**
Crystal parallel	84	149	4 (1-28)	8,977
Crystal antiparallel	70	84	4 (1-36)	9,576
bZIPs	76	130	2 (2-3)	6,666
Parallel heterodimers	9	18	4 (2-7)	1,074

Table 4-1. Summary of alignment test sets.

from the N-terminal coiled-coil **f** position through the C-terminus. In order to increase the number of tested alignments, known extents were defined as the alignment search boundaries but with seven residues deleted from the C-terminus. A final test set, known as parallel heterodimers, consisted of five sequences from crystal structures, two from bZIPs and two from the known heterodimeric keratins[16]. Keratin alignments were derived from [12]. All sequences for all test sets are included in Appendix C.

4.2.3 Scoring models

Both ISMs and ESMs were tested. The ISMs consist of weights applied to interhelical residue pair interactions that are specific to certain heptad pair positions (e.g. **a-a', e-g'**). These weights can be derived from three sources: (1) rationally, choosing pairs and weights based on hypotheses or literature consensus; (2) experimentally, from experimentally-measured residue coupling energies or otherwise; (3) computationally, from either statistical methods or machine learning methods. We used five previously developed ISMs[7] that differ significantly in their weights and positions scored. A summary of the ISMs can be found in Table 4-2. Model ELEC is a rational-based model that scores edge (parallel: **e-g'**; antiparallel: **e-e'** + **g-g'**) interactions with a simple electrostatic patterning score, with +1 for same-charge pairs and -1 for opposite-charge pairs. Model CE is based on experimentally determined coupling energies at edge as well as core positions (parallel: **a-a'** + **d-d'**; antiparallel: **a-d'**). The RISP models are based on applying the RISP potential of Apgar et al.[7] (Chapter 3) to various sets of heptad pair positions: $RISP_{core,edge}$ has only core and edge weights; $RISP_{CC}$ contains all weights from

Term	P	AP	ELEC	CE	SVM	$RISP_{core,edge}$	$RISP_{CC}$	$RISP_{CCall}$
CORE	a-a' d-d'	a-d'		X	X	X	X	X
EDGE	g-e'	g-g' e-e'	X	X	X	X	X	X
COR-EDG	a-g' d-e'	a-e' d-g'			X		X	X
VERT	a-d'	a-a' d-d'						X

Table 4-2. Pair terms used in ISMs. All models except FKS from [7]. FKS model from [17].

$RISP_{core,edge}$ plus core-to-edge weights (parallel: **a-g'** + **d-e'**; antiparallel: **a-e'** + **d-g'**); and $RISP_{CCall}$ contains all weights from $RISP_{CC}$ plus vertical core weights (parallel: **a-d'**; antiparallel: **a-a'** + **d-d'**). We also applied a model (FKS) based on a set of weights learned via a support vector machine from long parallel coiled coil dimers[12]. The FKS model contains terms representing core, edge and core-to-edge interactions. Because the FKS model was not trained on antiparallel dimers, the weights derived for parallel examples were applied to the approximately equivalent antiparallel context. Finally, two null control models were created. The first assigns random weights to all possible residue pairs at core and edge positions (NULL), and the other uses similar residue weights but scores all homotypic interactions as -100.0 ($NULL_{HOMO}$).

All tested ESMs follow a common model-building framework, as described previously[7]. First, the test sequence pair was repacked using RosettaDesign onto either a set of 81 antiparallel or 121 parallel backbones. Next, structure relaxation using CHARMM energy minimization with Crick backbone restraints was used to relax the rotamer approximation. Finally, all repacked and minimized structures were evaluated with an appropriate energy function, with the lowest energy structure accepted as the

predicted structure. All ESMs shared this model-building framework and differed only in the energy function used. We tested four previously developed ESMs using the DFIRE, FoldX, Rosetta and GK energy functions[7]. Energy terms for previous ESMs are described in Chapter 3. In addition, the HP/S energy function was implemented as described previously[8]. The HP/S model was broken down according to the same terms used for ISMs (Table 4-1) but with the POINT term denoting helix propensities and the Rest term comprising all pairwise interactions not included in other terms.

The HP/S model was used as the basis for a series of hybrid models that incorporated implicit structure information. The archetype of this hybridization is the HP/S/C model as described in [8], which replaced core interactions with weights derived from the previously described FKS model. We implemented this model (referred to as HP/S/C_{FKS}) along with two variations that used core position weights from the CE (HP/S/C_{CE}) and RISP (HP/S/C_{RISP}) models. Finally, we tested two additional models that replaced the helix propensity portion of the HP/S/C model with a coiled-coil-specific propensity term derived using the Paircoil algorithm[18]. For each helix in the test interaction, individual residue propensities were calculated according to the single position frequencies from the Paircoil2 training database[14] as

$$R(r_i, h_i) = \ln \frac{f(r_i, h_i)}{f_{bkg}(r_i)}$$

where $f(r_i,h_i)$ is the frequency of residue r_i occurring with heptad position h_i in the coiled-coil database, and $f_{bkg}(r_i)$ is the frequency of residue r_i observed in Genbank[18]. Next, overlapping 7-residue windows were superimposed on each sequence and all residue propensities under each window were summed to produce a set of window scores. Residue scores were defined as the maximum window score over all windows

overlapping a given residue. The final score for the interaction is given as the sum over all residue scores. The models derived from this method are referred to as $CC/S/C_{FKS}$ and $CC/S/C_{RISP}$.

In addition, we tested a hybrid model known as RosettaCC, which combined Rosetta (omitting the Eref term) together with the coiled-coil score described above.

4.2.4 Performance metrics

We used two metrics to quantify performance. The first is referred to as FBR (Fraction-Best-Ranked). For a given positive with associated negative decoys, this metric records a success only if the positive scores lower than all associated decoys (best possible rank). This metric is easy to interpret, as it unambiguously provides a fraction of correct predictions out of all possible predictions. However, it is highly dependent on the average number of decoys per positive, making comparisons among test sets difficult. In addition, this metric is relatively insensitive to smaller changes in score that do not result in more positives achieving the best rank. Therefore, we developed the normalized total performance metric, known as TPMn, which corrects for both of these problems. Un-normalized TPM is computed as the sum over all positives of the reciprocal of the rank of each positive among its associated negative decoys, where the best rank is rank 1. This allows for "partial credit" given to positives that are low-ranked but not best-ranked. Normalized TPM is computed as the ratio of (TPM – TPMnull) / (TPMmax – TPMnull), where TPMnull is the value of TPM computed on a random distribution of positive ranks, and TPMmax is the TPM of the best possible distribution, where all positives have rank

1. This results in a score from +1 (where the model assigns every positive as rank 1) to zero (the model is indistinguishable from null). TPMn scores lower than zero are possible, and represent performance worse than null.

4.2.5 Homodimer preference

To quantify the preference of a scoring model for parallel homodimers, we tested a method whereby each sequence pair in a set of heterodimeric sequences was compared to the two unobserved homodimers that result from partnering each strand with itself. This method assigns a score of 0.5 to a correct prediction (favoring the heterodimer) and -0.5 to an incorrect prediction. The score for a sequence pair was the sum of scores from both component helices, and the final homodimerization score assigned to a score model was the average over all sequence-pair scores resulting from that model. Final scores fall in the range between 1.0 (exclusively favors heterodimer) and -1.0 (exclusively favors homodimer).

4.2.6 Model optimization

We used two approaches to optimize the performance of our scoring models on our test sets. To optimize the CC/S/V_{RISP} model on structure-derived antiparallel sequences, the Rest term was excluded, while the HP/S-based VERT term was replaced by the equivalent term from $RISP_{CCall}$. Next, a grid-based search over the CORE and

VERT scaling weights was used, with the grid ranging from -2.0 to 2.0 using a step size of 0.1.

For the parallel sequences, a Monte Carlo-based approach was used to simultaneously vary the scaling weights of all scoring terms for a given model. For each iteration of the optimization, one term was randomly selected and its weight was incremented by a random value between -0.5 and 0.5. Performance using the new weights was computed according to the TPMn metric, and the move was accepted either if performance increased or with a probability of $1-(0.97^{s})$ where s is the current iteration step. This procedure, iterating over 300 steps, was repeated 5 times for each test set and scoring model pair, and the final performance was taken as the maximum observed performance over all optimization runs.

4.3 Results

There are two major classes of dimeric coiled-coil alignment problems. The first involves searching for the alignment of two sequences where the coiled-coil boundaries of one or both sequences are not well defined. This problem arises most often when using coiled-coil detection methods, since these methods are not always accurate at predicting coiled-coil boundaries[19]. With this alignment problem, the predictor must discriminate mostly between the true coiled-coil-forming region and regions that have some coiled-coil-like features but do not form coiled coils. The second class of alignment problems seeks the alignment of two sequences that are both known to potentially form a coiled coil along their entire length. This can be found, for example, when aligning a

heterodimeric pair of designed leucine zippers with differing lengths. With this problem, the two sequences will have similar coiled-coil propensity across their entire lengths, causing prediction methods to rely more on interhelical interactions.

Our prediction framework addresses both of these problems, primarily through the choice of test set. The crystal-structure-based test sets relate directly to the first class, as they are constructed using the Paircoil2 method[14] to construct reasonable extensions of the known (by SOCKET) coiled-coil-forming regions into sequence that is often observed to not form a coiled coil. Figure 4-1 illustrates this process. In order to automate a typically manual process of observing Paircoil2 score graphs and selecting the valleys of well-scoring regions, we defined the alignment search boundaries at the points nearest to the known extent where the score increases at a per-residue rate faster than a defined cutoff. In contrast, all sequences in the bZIP test set have search boundaries that contain only coiled-coil-forming sequence. This test set simulates the aforementioned designed leucine-zipper alignment problem, albeit with native leucine zippers.

Given the problem as defined above, there are two possible approaches to evaluating any given candidate alignment. One approach considers the interaction as "blunt-ended" – in other words, sequence that is not part of the defined coiled-coil interaction is not modeled or considered. This approach is more appropriate for the first class of alignment problems, where any given model of the non-overlapping sequence is not likely to be correct. In addition, this approach is more compatible with ISMs, which consider mainly local interactions. The second approach models the entire complex, including non-overlapping sequence, and only compares the changes in interaction structure for each candidate alignment. This approach, similar to the orientation

prediction approach discussed in Chapter 3, does not require scoring models to contain reference state terms. However, the major disadvantage to this approach is that it relies only on interhelical interactions to determine alignment, and does not consider the specific preferences of residues to form amphipathic helices. Since proper consideration of the reference state has been previously discussed to be important in the prediction of coiled-coil partnering[8], a super-problem of coiled-coil alignment, we chose the blunt-ended approach.

4.3.1 Performance of ISMs

The implicit structure models are a class of simple, fast sequence scoring models that rely on previous characterization of model structure. Figure 4-2 summarizes the predictive performance of the seven tested ISMs using the TPMn metric. All non-null methods have positive predictive value, and many models show performance above 0.80 (corresponding to FBR 83-90%). The FKS model shows excellent performance on the bZIP sequences, as described previously[17]. However, this is not observed for other test sets, which are best predicted using the RISP models. The CE model, which was originally developed using a parallel bZIP model system[20], shows good performance on the crystal parallel and bZIP sets, but shows poor performance on the crystal antiparallel set. Indeed, both the CE and FKS models, which relied on terms from parallel models being applied to the analogous antiparallel interactions, perform poorly on the crystal antiparallel set. Intriguingly, the RISP models show different performance trends as the number of terms changes. The crystal antiparallel sequences are predicted better

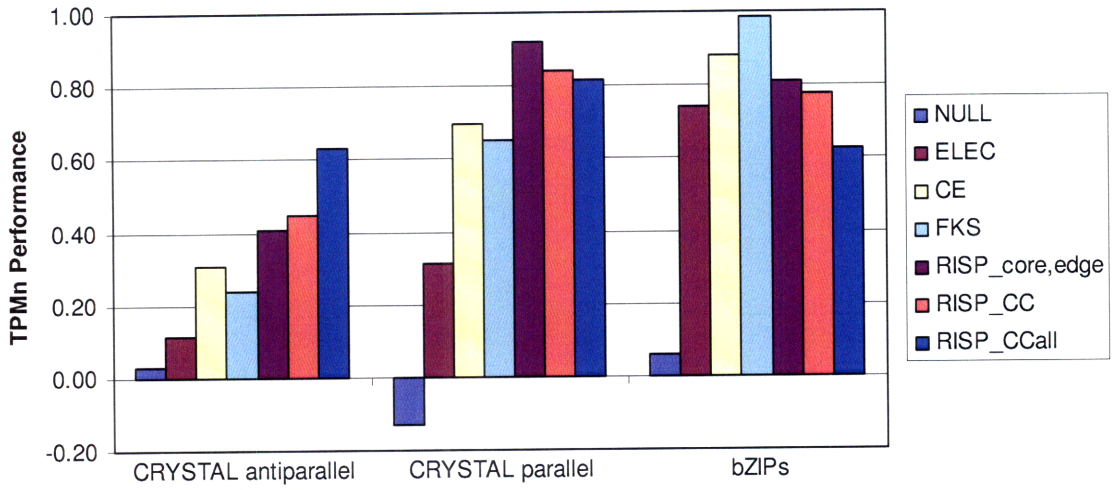


Figure 4-2. Prediction performance of ISMs. Performance is reported in units of TPMn and is evaluated separately for three test sets: antiparallel coiled coils with known structures, parallel coiled coils with known structures, and bZIPs, which are known to function as parallel dimers.

when more terms are added, while prediction performance for the crystal parallel and bZIP sets actually degrades with addition of these terms.

To understand these models in more detail, we used component analysis to isolate the contributions of each term in the model to the final performance. Component analysis looks at the change in performance as a model undergoes two different perturbations: one where a given term is excluded during scoring (*all but*) and the other where the same term is the only term used for scoring (*only*). If performance increases during either of these perturbations, this is an indication that the term either contributes negatively (in the case of 'all but' testing) or is a dominant contributor (in the case of 'only' testing). If performance decreases, this represents a more complicated relationship among the terms of the model.

We used component analysis to determine the causes of performance trends in the ISMs. Figure 4-3 highlights key findings. First, the poor performance of the $RISP_{CCall}$

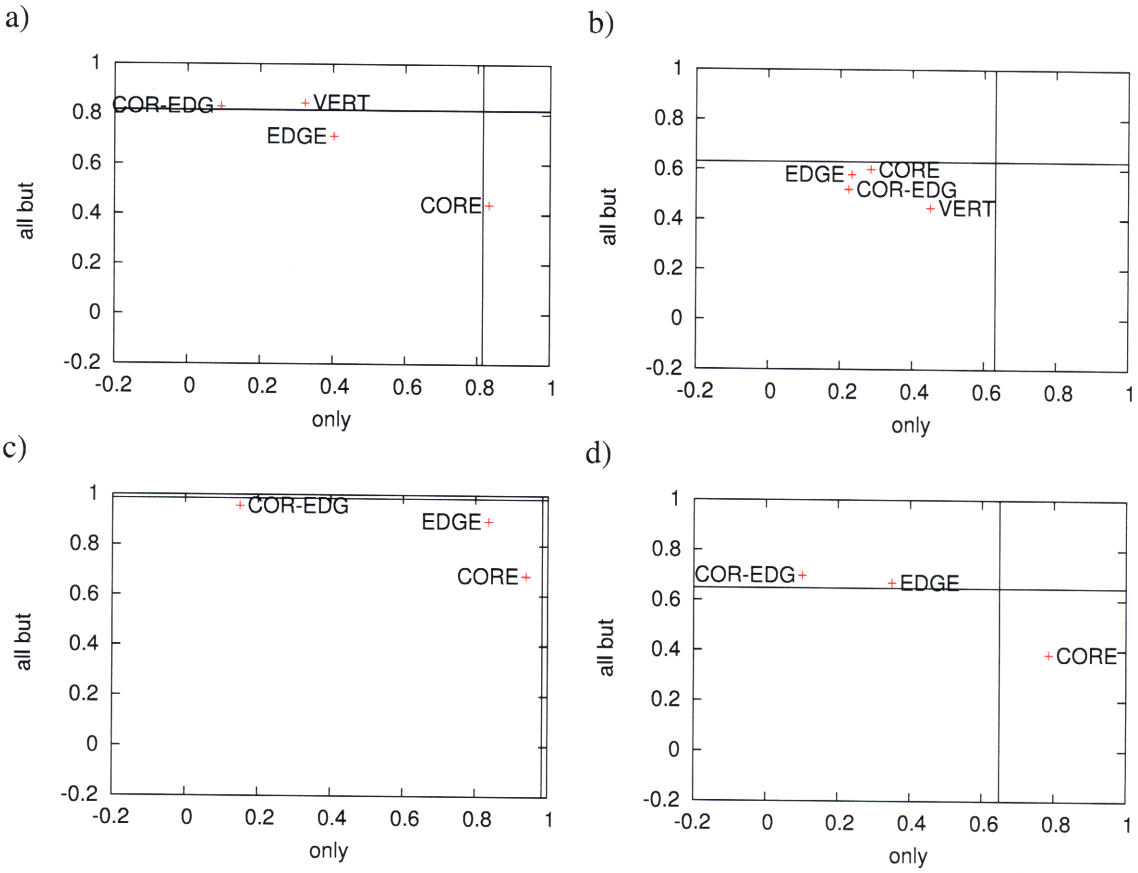


Figure 4-3. Component analysis of selected ISMs. $RISP_{CCall}$ on (a) crystal parallel and (b) crystal antiparallel sequences, as well as FKS on (c) bZIP and (d) crystal parallel sequences. Vertical and horizontal axes represent model performance in TPMn when a term is excluded (all but) or is used by itself (only), respectively. Vertical and horizontal lines represent performance of the unmodified model. Points below or to the left of these lines depict decreases in performance upon perturbation; points above or to the right depict increases in performance.

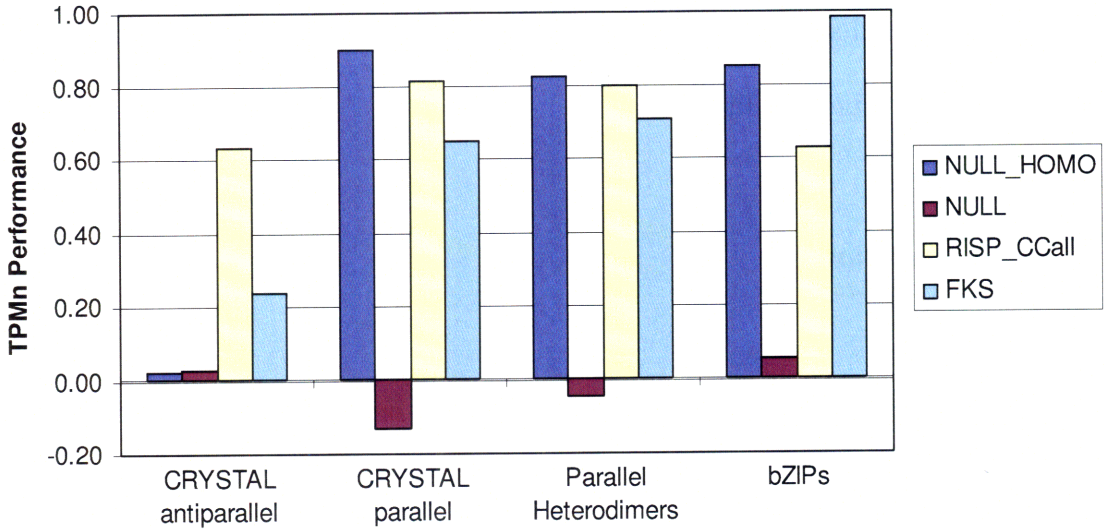
model on the crystal parallel sequences is a result of both the COR-EDG and VERT terms, since omitting either of these improves performance (Figure 4-3a). In contrast, the $RISP_{CCall}$ model applied to the crystal antiparallel sequences shows no such reduction in predictive power, suggesting that all terms contribute favorably to the prediction (Figure 4-3b). However, the VERT term (towards the lower right of the figure) appears particularly significant for the final prediction. Analysis of the FKS model applied to the bZIPs shows that all terms contribute, although the bulk of the predictive performance of

this model comes from the CORE term (Figure 4-3c). In contrast, the FKS model applied to the crystal parallel set shows negative contributions of the COR-EDG and EDGE term to the total model (Figure 4-3d).

4.3.2 Homotypic bias

A striking feature of the data in Figure 4-2 is the disparity in performance between the parallel and antiparallel sequences. One reason for this may be the large proportion of homodimer sequence in the crystal parallel set. Homodimers are expected to be much easier to align compared to heterodimers, given that their core interactions (**a-a'** and **d-d'**) are exclusively homotypic. However, models that excessively favor homotypic interactions may not be effective at predicting the interactions of parallel heterodimers. Therefore, to characterize these effects, we tested the performance of a composite set of nine parallel heterodimeric sequence pairs. Four of the nine sequence pairs (two bZIPs and two keratins) are paralogous in nature, being likely evolutionarily related[16,21]. This set is quite small relative to the size of the parallel homodimer set (70 sequences), and is not representative of all parallel heterodimeric coiled coils. Nevertheless, some interesting trends are obvious even with these few sequences, as seen in Figure 4-4a. Alignment prediction performance of the parallel heterodimer sequences as measured by TPMn is similar to that of both the crystal parallel and bZIP sequences. In addition to the models previously tested, we constructed a control null model (known as $NULL_{HOMO}$) which contained random weights for the core and edge interactions, with the exception of strongly favorable weights at core homotypic interactions. While this

a)



b)

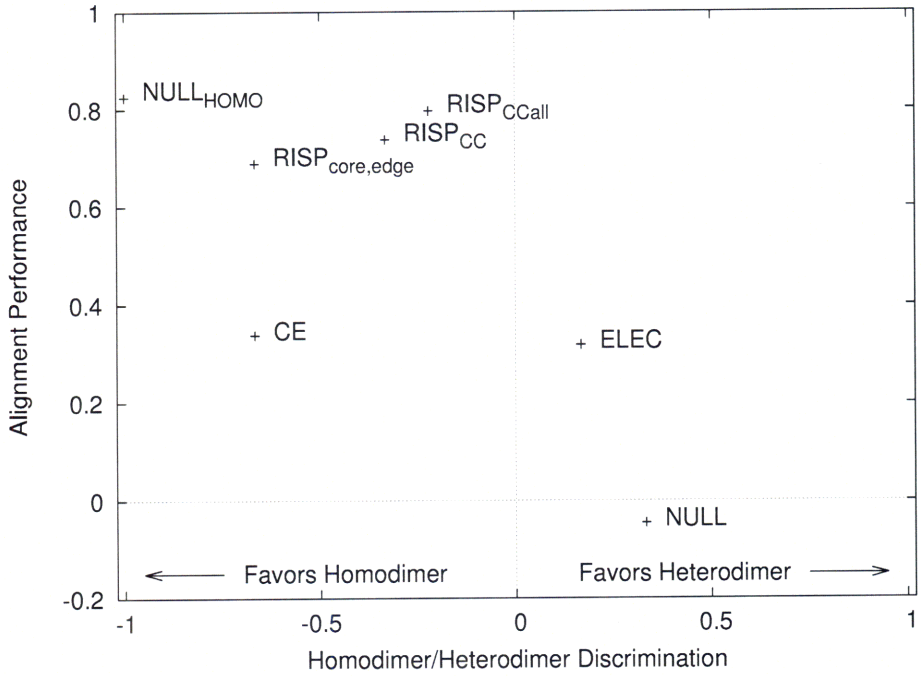


Figure 4-4. Homodimer bias analysis. (a) Performance of model $NULL_{HOMO}$ compared to high-performing ISMs on all test sets. Performance is in units of TPMn. (b) Correlation of ISM performance with homodimer preference. Alignment performance in units of TPMn.

model shows null performance on the antiparallel sequence set, we observe excellent performance on all parallel sequences, including the parallel heterodimer set. This suggests that favoring homotypic interactions in parallel sequences may be part of an effective strategy for predicting parallel coiled-coil alignment even in heterodimeric sequences. However, favoring homotypic interactions is not sufficient for some parallel sequences, given that certain models that incorporate more detail are able to make better predictions than the simple control model $NULL_{HOMO}$.

The standard alignment test is an indirect method of studying the propensity of a model to favor homotypic interactions. A more direct test compares the ability of a model to assign a favorable score to an observed heterodimer relative to the two un-observed homodimers that result from duplicating the individual strands of the heterodimer. Previous results using the FKS model on a set of keratin sequences showed reasonable performance on this test[12]. We compared the results of this test to the alignment prediction performance on the parallel heterodimer test set in order to understand whether favoring homotypic interactions correlates with alignment performance, either favorably or unfavorably. These data are plotted in Figure 4-4b. We observe no correlation between the ability of a model to predict the correct alignment of a pair of heterodimeric sequences and the ability of the same model to predict heterodimerization in the known heterodimer set over homodimerization. For the three RISP models, alignment prediction performance appears to increase slightly as hetero-preference increases, although all three prefer homodimers to heterodimers. The FKS model also shows high alignment prediction performance and high hetero-preference. However, we observe the highest

alignment prediction performance with the $NULL_{HOMO}$ model, which shows no hetero-preference at all.

4.3.3 Performance of ESMs

While implicit structure models are fast, simple and can provide good performance, the assumptions built into such simple models may limit their effectiveness. With all ISMs, performance is significantly different between the two parallel test sets. One possible reason for this disparity is the lack of consideration by the ISMs of the possible structural diversity between these sets. In order to investigate this hypothesis, we tested a series of explicit structure models (ESMs) that are much more detailed than their implicit counterparts. All ESM scores were calculated on the same set of structures; differences in predictions were solely due to the final evaluation step. We tested five evaluation functions: Rosetta, DFIRE, FoldX, GK [7] and HP/S [8].

Figure 4-5 shows the performance of the ESMs on our test sets, compared to the best-performing ISM for each set (crystal antiparallel: $RISP_{CCall}$, crystal parallel: $RISP_{core,edge}$, bZIPs: FKS). The GK model, which previously showed good performance on a dimeric coiled-coil orientation prediction test[7], is not able to distinguish alignment in this test. In contrast, the HP/S model, which was refined on a test set of bZIP interactions[8], shows excellent performance on our bZIP alignment set. We also observed lower than anticipated performance for the Rosetta model on all test sets. DFIRE performs very well on both parallel test sets, which is likely due to a higher

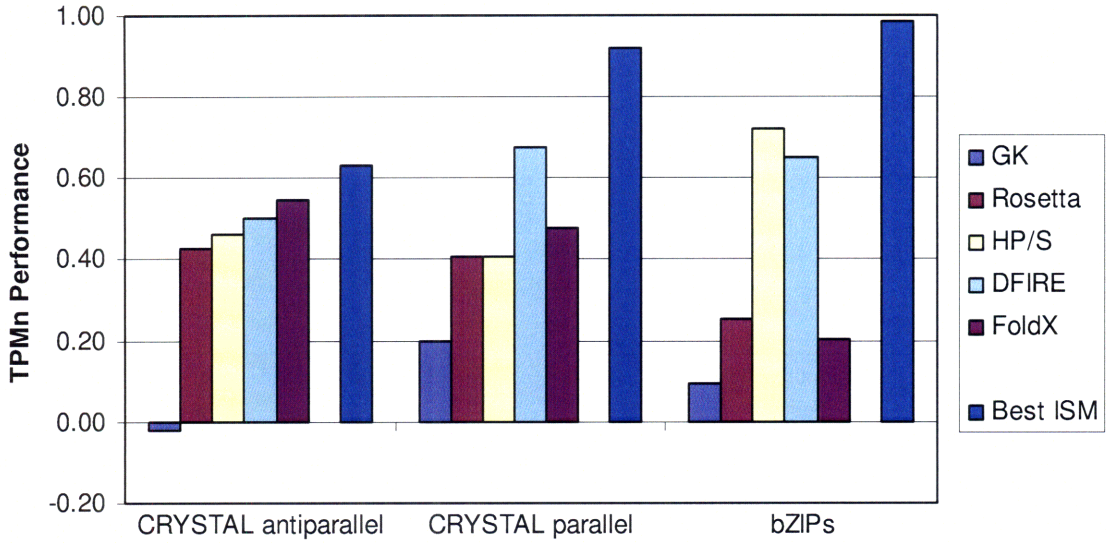


Figure 4-5. Performance of ESMs. Performance is in units of TPMn.

amount of homodimer bias in the potential[1]. Overall, we observe lower performance of the ESMs relative to the best ISM tested. Because this structural modeling framework was used successfully in the prediction of coiled-coil orientation, we focused on understanding and optimizing the energy functions used in these models.

Component analysis, shown in Figure 4-6, highlights some of the causes of poor performance among the explicit structure models. The choice of reference state has been previously discussed as being critical to the prediction of coiled-coil partnering[8]. Most of the ESMs contain reference states optimized using generic globular proteins, while the HP/S model contains a reference state derived from helix propensities that was shown to significantly improve coiled-coil partnering detection[8]. However, Figure 4-6 shows that for both the HP/S and Rosetta models, the reference state term strongly disfavors alignment prediction performance. In addition, the CORE term of the HP/S model shows

[1] Heterodimer preference (described in section 4.3.2) of DFIRE on the parallel heterodimer set is -0.333 (favoring homodimer), compared to the median over all other standard ESMs of 0.277 (favoring heterodimer).

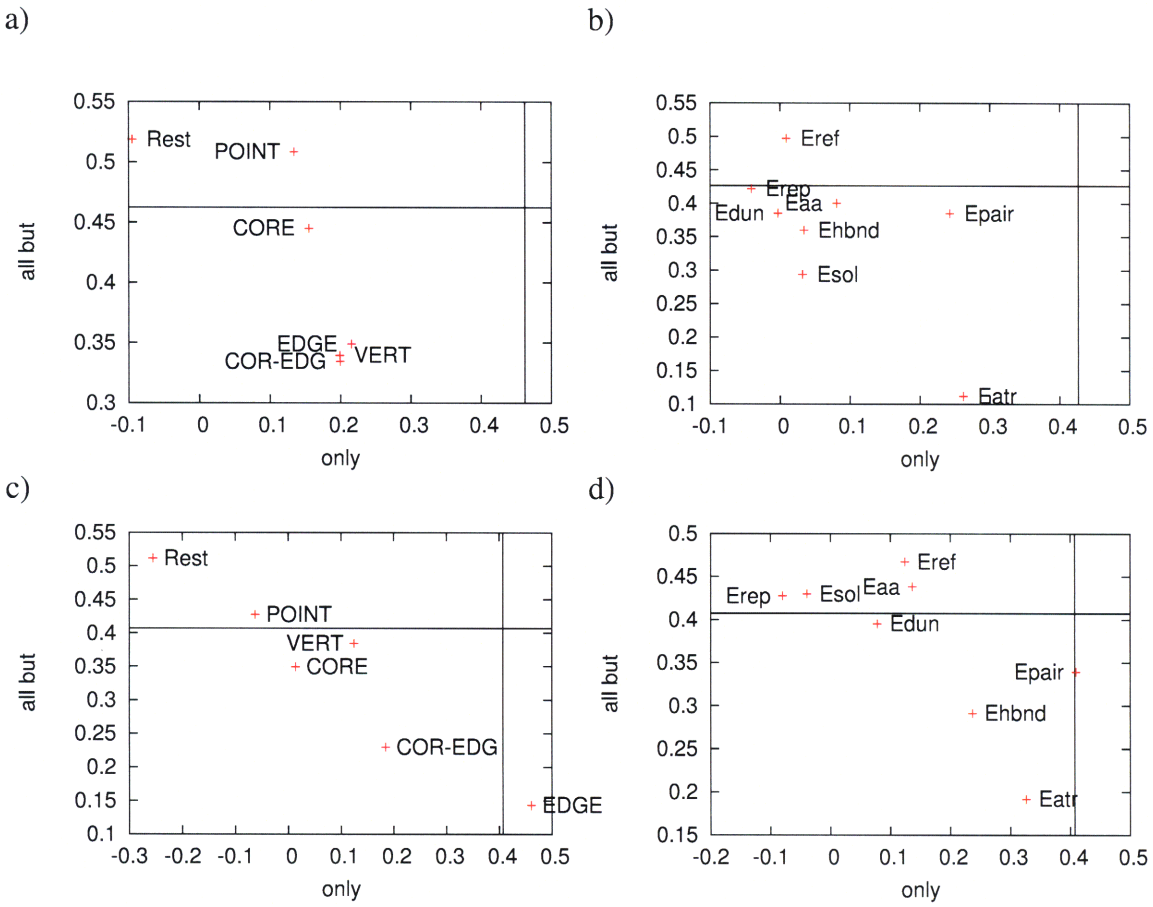


Figure 4-6. Component analysis of selected ESMs. Crystal antiparallel sequences analyzed using (a) HP/S and (b) Rosetta. Crystal parallel sequences tested with (c) HP/S and (d) Rosetta models. Axes and points as described in Figure 4-3.

poor predictive ability relative to the remaining terms.

Based on this analysis, we built several updated models that replace terms from poor-performing ESMs with analogous terms from ISMs. Figure 4-7 shows the performance of these new models. Model RosettaCC replaces the Eref reference-state energy term with a term derived from the coiled-coil propensity of the sequence. This term significantly improves performance on all test sets. The HP/S/C series of models replaces the core residue interactions from the HP/S model with the core contact potentials from three different ISMs: CE, FKS and RISP. This type of replacement was

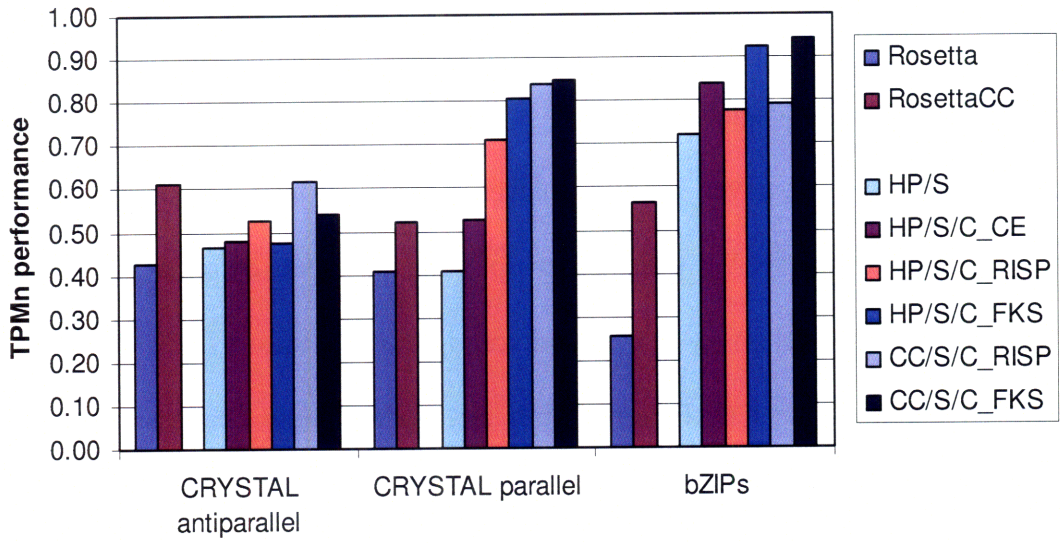


Figure 4-7. Performance of modified, unoptimized ESMs. Performance is in units of TPMn.

previously shown to improve performance for bZIP partner prediction[8], and is observed in this work to improve alignment prediction performance above the base HP/S model in all cases. Finally, the CC/S/C series of models was generated by replacing the HP reference term in the HP/S/C model with the same coiled-coil propensity term used in RosettaCC. This change improves performance slightly for all models, particularly that of the RISP-based models on the crystal sequence test set.

4.3.4 Model optimization

In the hybrid models described above, terms were combined from different sources without any adjustment in relative weights. Since these different models were not designed to produce scores on similar energy scales, we attempted to adjust the relative weights of component energy terms in order to improve performance. However, since

these results were not cross-validated, these tests only provide a theoretical maximum of performance of the models on our sequence sets.

Model $CC/S/C_{RISP}$ shows the best performance of all ESMs on the crystal antiparallel sequences. However, it only achieves a TPMn performance of 0.61, still far below the parallel sequences. We were able to optimize this model by replacing its VERT term with that from the $RISP_{CCall}$ model, and scaling the new term by 1.5. Furthermore, we removed the Rest and CORE terms from the model. These changes, resulting in model $PC/S/V_{RISP}$, improved performance to 0.71, which is the highest performance observed on the antiparallel sequence set. This performance corresponds to over 77% of the sequences predicted correctly. Interestingly, of the remaining 23% of sequences not predicted correctly by this optimized model, at least one other tested model predicts each sequence correctly.

The best un-optimized models for the parallel sequences have TPMn performance of 0.92 and 0.98 for the $RISP_{core,edge}$ model on the crystal parallel test set and the FKS model on the bZIPs, respectively. We used a simple Monte Carlo-based search routine to optimize term weights for these sets. On the crystal parallel pairs, TPMn performance of 0.96 was achieved using the $CC/S/C_{RISP}$ model by down-weighting the Rest term while emphasizing the CORE term. On the bZIP sequences, perfect performance of 1.00 was achieved through a variety of models including the FKS model, which down-weighted the EDGE and COR-EDG terms while again emphasizing the CORE term.

4.4 Discussion

We have described the performance of a variety of models developed for the prediction of coiled-coil alignment. Many models have the ability to favor the correct alignment relative to a reasonable set of decoy alignments, and some are able to predict the observed alignment of >90% of a set of parallel dimeric coiled coils. However, no single model is optimal for all test sets; significant differences in performance are observed among the crystal parallel and parallel bZIP sets. This indicates that much work remains in developing a universal model for predicting coiled-coil alignment.

Previous work in predicting the alignment of the spectrin antiparallel trimer only considered interhelical charge interactions, assuming that any variation in hydrophobic core interactions would not significantly affect alignment specificity[10]. This assumption was supported by experimental evidence from the Fos/Jun heterodimer, which suggested that partnering specificity was determined primarily by charged edge-position residues[22]. However, in our study, many parallel sequences can be aligned correctly using models that rely heavily on core terms, in contrast to these previous assumptions. Indeed, the crystal antiparallel sequences performed best on models such as $RISP_{CCall}$ and $CC/S/C_{RISP}$ that considered more detailed core vertical interactions, although performance of parallel sequences decreased using the same terms.

One hypothesis regarding the evolution of paralogous heterotypic interactions is that they commonly proceed through a homotypic intermediate[23,24,25]. This hypothesis implies that favorable homotypic interactions are easily formed at random, when selection pressure is low, and that when selection increases, heterotypic

compensatory mutations continue to stabilize the complex. The coiled coil is an interesting test system for this hypothesis, as it allows for comparison of homodimers and heterodimers in the same structural context. Our results support this hypothesis, as we have observed that a simple model favoring homotypic interactions is able to predict a small set of preferentially heterodimerizing coiled coils. This effect may be a physical preference for identical residues to be in contact with each other, given that both the CE model (based on experimental data) and the RISP models (trained on heterotypic protein-protein interfaces, section 3.3.6) show a clear preference for homotypic versus heterotypic interactions (Figure 4-4b). Upon closer inspection of these potentials, we observed a preference for homotypic interactions primarily among the commonly-found coiled-coil core hydrophobic residues (leucine, isoleucine and valine) as well as certain core polar interactions (asparagine in CE, asparagine and glutamine in RISP). Therefore, while physical homotypic preference is not expected for all interactions, it is observed in a key subset of residues commonly found in the cores of coiled coils.

One important application of coiled-coil alignment prediction is to accurately predict the alignment of a pair of designed leucine zippers. Since we do not have a test set of validated designed leucine zipper alignments, we used the bZIPs as a reasonable substitute. Our results indicate high predictive performance on the bZIPs for many models. One feature common among leucine zippers is the use of core polar interactions to constrain orientation[26], partnering specificity[27] and oligomerization state[28]. Such interactions are also assumed to favor proper helix alignment. While the high performance of the ELEC model (which lacks core terms) on the bZIPs suggests that

such interactions are not required, the best performing models on bZIP sequences (including CE and FKS) have strong favorable weights for core polar interactions.

The alignment performance of all methods on our crystal antiparallel dimer test set is significantly lower than that observed on the parallel sets. There are several possible reasons for this. First, the CE and FKS models attempted to use analogous weights from parallel studies in the antiparallel context. This approach, while simplistically true for edge charge patterning[10,29], proves to be inappropriate for detailed models. Second, our implicit structure models assume the same generic structural environment for all sequences, which may limit performance given our understanding of the importance of modeling structural diversity[7,30]. However, our explicit structure models, which incorporate backbone flexibility and full structure modeling, did not perform significantly better than the implicit structure models. Finally, the parallel test set contains a significant number of homodimers, which are expected to be trivial to predict. This may result in a general overestimate of performance for parallel sequences. However, a control test set of parallel heterodimers had similar performance to the other parallel test sets, and no correlation was observed between homodimer bias and alignment prediction performance.

Given the good performance of the ESMs in predicting coiled-coil orientation, it was reasonable to expect similar performance on the related question of coiled-coil alignment. However, the poor performance observed can be explained for a number of reasons. First, consideration of the reference state is a major factor in modeling protein-protein interactions, and many explicit structure energy functions do not model this state effectively[8]. Coiled-coil alignment, unlike orientation prediction, requires an effective

reference state to compare structures containing different sequence. For example, the GK model, which performed well at orientation prediction despite lacking a reference state term[7], performs near the null level for alignment prediction on all sequences. We observe significant improvement of the Rosetta and HP/S ESMs when generic reference states were replaced with a coiled-coil-specific term. Interestingly, the helix-propensity reference state of the HP/S model, previously shown to be effective at predicting bZIP association[8], was counter-productive on both crystal parallel and crystal antiparallel sets (although it improved alignment performance on the bZIP set). This may be due to the prevalence of helical sequence being included within alignment test boundaries. Instead, the coiled-coil-specific term may assist the energy function to locate the most likely coiled-coil region even within regions already assigned to be coiled coil by Paircoil2. Finally, core interaction modeling was shown to be insufficient, as replacing either core or vertical terms with well-performing terms from ISMs improved performance on crystal parallel and crystal antiparallel sequences, respectively. This effect has been previously discussed as a necessary advancement of explicit structure models for coiled coils[8].

Several simple extensions of this work may prove fruitful. First, the RISP ISMs do not explicitly model the known physicochemical differences among coiled-coil heptad positions, e.g. differences in solvation of core versus edge interactions. Rather than applying the single RISP contact matrix to all pairwise interactions, it may be possible to construct individual potentials for each heptad pair interaction that are better representations of the stability of an interaction in a particular structural context. Such models may prove to be more effective without sacrificing the speed and simplicity of

traditional RISP. Second, the ESMs showed poorer than expected performance, even given the modifications described above. Detailed analysis of the predicted structures, along with comparison to observed crystal structures, may show trends among poorly performing sequences. This type of analysis was previously instrumental in highlighting key features promoting dimeric coiled-coil orientation[7].

In order to refine our antiparallel models, new techniques to understand the relationship between model and reality are required. While the component analysis method provides general indications of the strengths and weaknesses of a given model, it does not provide sufficient resolution to understand which scores or residues may be modeled incorrectly. Future experimental data, possibly obtained using model coiled coils such as those developed by Hadley et al.[31], may prove useful for refining predictive models of antiparallel coiled coils. In addition, it may be possible to discern features that divide the set of antiparallel coiled coils into distinct classes and predict each class with a slightly different model. This approach is supported by the observation that all antiparallel sequences are predicted correctly by at least one model. However, this approach requires a framework for predicting such classification while avoiding overtraining to the currently small set of test sequences.

The development of improved models for predicting coiled-coil alignment is an important step on the road towards coiled-coil interaction prediction. Existing models for predicting parallel alignment may be sufficient, although antiparallel models are not yet complete. While future developments in explicit structure modeling may provide highly accurate predictions, further refinement of simple implicit structure models may improve performance without the need for detailed structure.

4.5 References

1. Fessler J (1974) Self-assembly of collagen. Journal of Supramolecular Structure 2: 99-102.
2. Engel J, Prockop DJ (1991) The Zipper-Like Folding of Collagen Triple Helices and the Effects of Mutations that Disrupt the Zipper. Annual Review of Biophysics and Biophysical Chemistry 20: 137-152.
3. Mo JM, Holtzer ME, Holtzer A (1991) Kinetics of self-assembly of alpha alpha-tropomyosin coiled coils from unfolded chains. Proceedings of the National Academy of Sciences of the United States of America 88: 916-920.
4. Araya E, Berthier C, Kim E, Yeung T, Wang X, et al. (2002) Regulation of coiled-coil assembly in tropomyosins. Journal of structural biology 137: 176-183.
5. Kammerer R, Schulthess T, Landwehr R, Lustig A, Engel J, et al. (1998) An autonomous folding unit mediates the assembly of two-stranded coiled coils. Proceedings of the National Academy of Sciences of the United States of America 95: 13419-13424.
6. Wolf E, Kim PS, Berger B (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein science : a publication of the Protein Society 6: 1179-1189.
7. Apgar JR, Gutwin KN, Keating AE (2008) Predicting helix orientation for coiled-coil dimers. Proteins 72: 1048-1065.
8. Grigoryan G, Keating AE (2006) Structure-based prediction of bZIP partnering specificity. J Mol Biol 355: 1125-1142.
9. Grigoryan G, Keating A (2008) Structural specificity in coiled-coil interactions. Current Opinion in Structural Biology 18: 477-483.
10. Parry DA, Dixon TW, Cohen C (1992) Analysis of the three-alpha-helix motif in the spectrin superfamily of proteins. Biophysical journal 61: 858-867.
11. Yan Y, Winograd E, Viel A, Cronin T, Harrison SC, et al. (1993) Crystal structure of the repetitive segments of spectrin. Science (New York, NY) 262: 2027-2030.
12. Singh M, Kim P. Towards predicting coiled-coil protein interactions; 2001. ACM. pp. 279-286.
13. Mason JM, Schmitz MA, Müller KM, Arndt KM (2006) 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 103: 8989-8994.
14. McDonnell AV, Jiang T, Keating AE, Berger B (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 22: 356-358.
15. Newman JR, Keating AE (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science 300: 2097-2101.
16. Coulombe P, Omary B (2002) [`]Hard' and [`]soft' principles defining the structure, function and regulation of keratin intermediate filaments. Current Opinion in Cell Biology 14: 110-122.
17. Fong JH, Keating AE, Singh M (2004) Predicting specificity in bZIP coiled-coil protein interactions. Genome Biol 5.

18. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, et al. (1995) Predicting coiled coils by use of pairwise residue correlations. Proceedings of the National Academy of Sciences of the United States of America 92: 8259-8263.
19. Gruber M, Söding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. J Struct Biol 155: 140-145.
20. Krylov D, Mikhailenko I, Vinson C (1994) A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. The EMBO journal 13: 2849-2861.
21. Deppmann CD, Alvania RS, Taparowsky EJ (2006) Cross-species annotation of basic leucine zipper factor interactions: Insight into the evolution of closed interaction networks. Mol Biol Evol 23: 1480-1492.
22. O'Shea EK, Rutkowski R, Kim PS (1992) Mechanism of specificity in the Fos-Jun oncoprotein heterodimer. Cell 68: 699-708.
23. Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res 33: 3629-3635.
24. Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI (2006) Structural Similarity Enhances Interaction Propensity of Proteins. J Mol Biol.
25. Lukatsky DB, Zeldovich KB, Shakhnovich EI (2006) Statistically Enhanced Self-Attraction of Random Patterns. Physical Review Letters 97.
26. Oakley MG, Kim PS (1998) A buried polar interaction can direct the relative orientation of helices in a coiled coil. Biochemistry 37: 12603-12610.
27. Zeng X, Herndon A, Hu J (1997) Buried Asparagines Determine the Dimerization Specificities of Leucine Zipper Mutants. Proceedings of the National Academy of Sciences of the United States of America 94: 3673-3678.
28. Junius FK, Mackay JP, Bubb WA, Jensen SA, Weiss AS, et al. (1995) Nuclear magnetic resonance characterization of the Jun leucine zipper domain: unusual properties of coiled-coil interfacial polar residues. Biochemistry 34: 6164-6174.
29. McClain DL, Woods HL, Oakley MG (2001) Design and characterization of a heterodimeric coiled coil that forms exclusively with an antiparallel relative helix orientation. J Am Chem Soc 123: 3151-3152.
30. Apgar JR, Hahn S, Grigoryan G, Keating AE (2009) Cluster expansion models for flexible-backbone protein energetics. Journal of computational chemistry.
31. Hadley E, Gellman S (2006) An Antiparallel [alpha]-Helical Coiled-Coil Model System for Rapid Assessment of Side-Chain Recognition at the Hydrophobic Interface. Journal of the American Chemical Society 128: 16444-16445.

Chapter 5

Conclusions and Future Directions

5.1 Prediction of coiled-coil structure

The coiled coil, despite being studied for over a half century, is still not fully understood. We have yet to describe the full range of proteins and processes in which coiled-coil structures play an important role. To this end, we seek to develop methods of predicting coiled-coil structure from sequence alone. Such methods, when fully realized, will enable organism-wide studies of coiled-coil structures and interactions from existing genomic sequence data. However, current methods only address one structural parameter at a time, while predicting uncharacterized sequence will require simultaneous prediction of all aspects of coiled-coil structure: partnering, oligomerization number, orientation and alignment. As discussed in the previous chapters, we have made improvements in three key subproblems of coiled-coil structure prediction, paving the way towards the goal of a unified coiled-coil structure prediction framework.

5.2 Coiled-coil databases and statistics-based prediction

As the number of known protein sequences increases, so does the amount of known coiled-coil sequence. The earliest database of coiled-coil sequence consisted of just a few thousand residues from parallel dimeric fibrous proteins[1]. This database, while small, proved instrumental in the development of early statistical coiled-coil prediction methods[2]. As described in Chapter 2, we have now compiled over 124,000 residues from 158 families that are known to adopt coiled-coil structures, covering many combinations of helix number and helix orientation. Because our database is annotated by structure as well as heptad, it is useful for training prediction methods that predict coiled-coil structure.

We demonstrated that as database size increases, the prediction of previously uncharacterized sequence improves. However, many families remain poorly predicted under cross-validation, indicating that performance is highly sensitive to individual family composition, and that additional sequences are required to improve performance. Because leave-family-out cross-validation performance is dictated by the composition of the remaining families, we expect that further performance improvement of the Multicoil method will most likely be realized by the discovery of new families that share common residue pairs with currently underperforming families. In addition, performance could improve simply by enlarging the size of existing families, assuming that the same underrepresented residue pairs occur in existing families at low frequency. However, using sequence homology searching methods to increase family size can be problematic,

since increasingly diverged sequence (which is the most valuable from a statistical perspective) may also have diverged structure[3]. Indeed, preliminary results from training Multicoil on a database containing large amounts of homology-search-derived sequence showed a slight decrease in overall performance, suggesting that such sequence may be less reliable. However, there are other possibilities for this poor performance, such as greater bias from over-represented families in the training set. Therefore, more work must be done to understand this result.

One possible method of targeting performance improvements would be to use the statistical data already collected to search for novel sequences that specifically improve under-represented observations. A similar approach was used in the Learncoil method, where an iterative process was used to detect sequences using a frequency table, then to update that frequency table according to the detected sequences[4]. Learncoil has been shown to improve detection of certain difficult-to-detect coiled-coil families[5,6], although it has not yet been used to enrich general coiled-coil prediction.

In addition to demonstrating the importance of high quality training databases, Chapter 2 showed how the choice of validation protocol has a significant impact on perceived performance. Currently, there is no strong consensus in the field of coiled-coil prediction as to the proper way to train and validate prediction methods. Most authors recognize the need to test using databases distinct from those used in training; however, there is currently no agreement on appropriate similarity or evolutionary cutoffs between these two sets. This lack of consensus makes fair comparisons between prediction methods very difficult. Our results indicate that previous methods of cross-validation such as leave-sequence-out overestimate performance when compared to stringent yet

realistic cross-validation approaches like leave-family-out. The contribution of our training database, broken down by distinct families, along with clear cross-validation routines, is an important contribution to this field.

Finally, further advancements to statistical frameworks will likely improve predictions, by using training data more efficiently. Sequence-profile methods have been shown to significantly improve prediction of secondary structure[7], and have recently been implemented for the prediction of coiled-coil location[8]. These methods may also prove useful in predicting further coiled-coil structural features, particularly for families with significant sequence diversity. However, similar to the concerns with using homology-search methods to expand training databases, profile-based methods of structure prediction may not improve performance if the profile inadvertently contains sequences of heterogeneous structure.

5.3 Current prediction of coiled-coil structural features

The earliest hypotheses about the coiled coil did not place any restrictions on the orientation or number of helices involved in the complex[9]. This turned out to be prescient, for although the first studied coiled coils were determined to be parallel dimers, it soon became clear that coiled coils could be found in a wide array of structural topologies[10]. However, relatively few methods have been developed to distinguish between these possible topologies based on protein sequence alone. The Multicoil method, retrained in Chapter 2 as Multicoil2, is currently the only widely-used method of predicting dimer versus trimer propensity. Chapter 3 describes a novel structure-based

method for predicting helix orientation in dimers, and illustrates possible molecular mechanisms influencing this preference. Finally, Chapter 4 describes the performance of structure-based methods for the prediction of helix alignment in parallel and antiparallel dimers.

Both of the methods in Chapters 3 and 4 utilize two classes of structural models: implicit structure models (ISMs) that score pre-defined residue pair interactions, and explicit structure models (ESMs) that evaluate a full 3D all-atom model of the interaction. For the implicit structure models, we showed that predictions made using simple charge-patterning rules, or weights from parallel systems used in antiparallel contexts, were insufficient to predict many sequences correctly. However, the interfacial contact potential RISP was able to make useful predictions about both orientation and alignment, especially for antiparallel sequences. Finally, the observation that homo-specific models can accurately predict heterodimeric sequence is interesting in light of hypotheses about the conservation of homotypic interactions during the evolution of protein complexes[11,12].

In contrast, the ESMs have shown varied performance in our tests. For the orientation prediction problem, ESMs that have poor models of the unfolded reference state still perform well, due to this state canceling. However, the same ESMs perform much worse in the alignment prediction problem, which does require a reasonable reference state. We tested the use of a coiled-coil-specific reference state, which improved performance particularly on antiparallel sequences. Also, the HP/S ESM appears to have difficulty modeling certain core interactions. Replacement of core or

vertical interaction scores with ISM-derived terms, as suggested previously[13], improves performance for the parallel and antiparallel sequences respectively.

5.4 The future of coiled-coil prediction methods

The ultimate goal of coiled-coil prediction is the concurrent prediction of all structural variations (number of helices, helix orientation, helix alignment and partnering preference). Currently, no method or combination of methods is able to achieve this goal. A major challenge is that the determinants of coiled-coil structure specificity are often subtle and involve a balance of factors. Unlike other domains, high sequence identity does not always connote similar structure, as coiled coils have been observed to change structure significantly upon a single point mutation. Instead, smaller differences in factors such as core packing, hydrogen bonding or rotamer selection can determine coiled-coil structural specificity. Existing models can capture some of these trends, however, more work is necessary in order to improve accuracy enough to predict all features simultaneously.

5.4.1 Improvements to existing methods

Several possible advancements may improve the performance of existing prediction methods. First, the RISP ISM has much potential for improvement. Currently, it does not consider the differences in residue environments between interactions (e.g. the difference in solvation between core and edge) or in residue geometry between coiled-

coil orientations (e.g. differences between parallel and antiparallel). Updating this model to consider such differences may improve its ability to model certain residue interactions. This can be done, for example, using methods from a variety of orientation-dependent potentials[14,15,16]. Such potentials have shown improved performance in decoy discrimination tests[17].

In addition, further developments in modeling the unfolded state of coiled coils, as well as improved evaluation of core interactions, will likely improve the performance of many ESMs. This has been discussed in work by Grigoryan and Keating[13].

5.4.2 Folding-based models

Current models used for coiled-coil structure prediction only consider the stability of the final folded complex. However, protein folding is a complex kinetic process, and prediction models that consider aspects of this process may show enhanced prediction performance. Many coiled coils have been shown to fold through a two-state mechanism that is hypothesized to require certain sequence patterns to nucleate folding[18,19]. In particular, heterodimeric tropomyosin was shown to require a specific region within the long sequence for proper folding and heterodimerization, and mutation of this region was shown to modify the association preferences of the entire coiled coil[20]. A model which is able to recognize such regions as nucleation sites may reduce the amount of structural sampling necessary for current alignment prediction methods. So-called "trigger sequence" motifs, while demonstrated to be important for the folding of certain coiled coils, are likely not a general solution to this problem[21]. However, further

understanding of the folding determinants of coiled coils, particularly whether they are localized or distributed, may focus predictions on more important sequence regions, decreasing prediction noise. Currently, the major challenge to folding-based prediction methods lies in understanding the critical "activated state" structure that is the nucleus of folding. This activated state is extremely short-lived, making structural characterization very difficult[18]. In addition, by focusing on the small activated-state region, models of the interactions within this region may need to be more detailed in order to make accurate predictions of the entire interaction.

5.5 Applications of coiled-coil structure prediction

Once methods of predicting coiled-coil structure improve, there are many potential applications with broad impacts for biology. Much experimental work has been done to characterize the structures of various coiled-coil proteins, with implications for their function. For example, the crystal structure of the SNARE core domain demonstrated a possible mechanism for membrane fusion: the folding of the SNARE core domain into a parallel four-stranded coiled coil places the C-terminal transmembrane domains of the membrane-embedded SNARE proteins in close proximity, overcoming the natural repulsion of the donor and acceptor membranes[22]. Interestingly, some SNARE proteins do not appear to be specific for the parallel configuration, suggesting a lack of competency for membrane fusion and possible regulatory roles for these proteins[23,24]. However, the importance of this structural specificity could have been predicted without extensive experimental results, if such prediction methods were

available. In this final section, I highlight several more applications of coiled-coil structure prediction, and discuss how advancements in this field will improve our ability to predict, understand and design biological systems.

One important application for coiled-coil structure prediction is to aid whole-sequence structure prediction. An example of a coiled-coil motif that frequently occurs within larger protein folds is the intramolecular antiparallel hairpin. This structure, typified by the seryl-tRNA synthetase protein[25], consists of two coiled-coil-forming helices connected by a loop or other intervening sequence. The ability to detect such structures on a genomic scale would not only show how common these motifs are in various protein families, but also would aid in predicting the structure of such proteins. This is also an intermediate application that is simpler than the general coiled-coil structure prediction problem, because partnering is restricted to a single sequence, and targets can potentially be identified by a coiled-coil—break—coiled-coil pattern predicted by detection methods. Such detection on a genomic scale will require accurate prediction of coiled-coil alignment and helix boundaries, as well as an estimate of the propensity of a sequence pair to form a hairpin.

In addition, the combination of experimental data with coiled-coil structure prediction methods can potentially elucidate some of the important interaction networks and structural complexes that are known to contain coiled coils. For example, the spindle pole body is hypothesized to contain many coiled coils, and while some have recently been studied experimentally[26], accurate prediction methods would help greatly in understanding how coiled coils play a role in the assembly of this complex. In particular, predictions of the possible connections between the individual components of the

complex can be combined with existing low-resolution structure data to build detailed models of the complex architecture[27]. Such approaches have previously been used to develop a model of the nuclear pore complex[28].

Finally, protein engineering is a fast-growing subfield of biological engineering. Its main goal is to utilize existing and novel protein structures to test and modify cellular function. One major part of this goal is to develop protein-protein interaction motifs for use in constructing protein complexes. For example, protein-protein interaction specificity has been implicated in the specificity of signaling networks[29,30], and re-engineering certain interactions using designed coiled coils has recently been used to alter the fundamental properties of such networks[31]. While this is a useful proof-of-concept, further modifications will require more interaction-mediating motifs with well-characterized stabilities, association preferences and structures. The leucine zipper is particularly suitable for this, being highly specific and relatively well understood[32,33]. However, current leucine-zipper design methods do not model all possible structural states. Better methods for predicting coiled-coil structure could address this and could, e.g., be used in conjunction with the recently developed CLASSY specificity design framework to expand the structural space available for design[32].

5.6 References

1. Parry DA (1982) Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. Bioscience reports 2: 1017-1024.
2. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science (New York, NY) 252: 1162-1164.
3. Galkin V, Yu X, Bielnicki J, Heuser J, Ewing C, et al. (2008) Divergence of Quaternary Structures Among Bacterial Flagellar Filaments. Science 320: 382-385.
4. Berger B, Singh M (1997) An iterative method for improved protein structural motif recognition. Journal of computational biology : a journal of computational molecular cell biology 4: 261-273.
5. Singh M, Berger B, Kim P, Berger J, Cochran A (1998) Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. Proceedings of the National Academy of Sciences of the United States of America 95: 2738-2743.
6. Singh M, Berger B, Kim PS (1999) LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins. Journal of molecular biology 290: 1031-1041.
7. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202.
8. Gruber M, Söding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. J Struct Biol 155: 140-145.
9. Crick FHC (1953) The packing of alpha-helices: simple coiled-coils. Acta Crystallographica 6: 689-697.
10. Lupas AN, Gruber M (2005) The structure of alpha-helical coiled coils. Adv Protein Chem 70: 37-78.
11. Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res 33: 3629-3635.
12. Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI (2006) Structural Similarity Enhances Interaction Propensity of Proteins. J Mol Biol.
13. Grigoryan G, Keating AE (2006) Structure-based prediction of bZIP partnering specificity. J Mol Biol 355: 1125-1142.
14. Buchete NV, Straub JE, Thirumalai D (2004) Orientational potentials extracted from protein structures improve native fold recognition. Protein Sci 13: 862-874.
15. Mukherjee A, Bhimalapuram P, Bagchi B (2005) Orientation-dependent potential of mean force for protein folding. The Journal of Chemical Physics 123.
16. Wu Y, Lu M, Chen M, Li J, Ma J (2007) OPUS-Ca: A knowledge-based potential function requiring only C{alpha} positions. Protein Sci 16: 1449-1463.
17. Miyazawa S, Jernigan R (2005) How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? The Journal of Chemical Physics 122.

18. Mo JM, Holtzer ME, Holtzer A (1991) Kinetics of self-assembly of alpha alpha-tropomyosin coiled coils from unfolded chains. Proceedings of the National Academy of Sciences of the United States of America 88: 916-920.
19. Steinmetz MO, Stock A, Schulthess T, Landwehr R, Lustig A, et al. (1998) A distinct 14 residue site triggers coiled-coil formation in cortexillin I. The EMBO journal 17: 1883-1891.
20. Araya E, Berthier C, Kim E, Yeung T, Wang X, et al. (2002) Regulation of coiled-coil assembly in tropomyosins. Journal of structural biology 137: 176-183.
21. Lee DL, Lavigne P, Hodges RS (2001) Are trigger sequences essential in the folding of two-stranded alpha-helical coiled-coils? Journal of molecular biology 306: 539-553.
22. Sutton B, Fasshauer D, Jahn R, Brunger A (1998) Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. Nature 395: 347-353.
23. Weninger K, Bowen ME, Chu S, Brunger AT (2003) Single-molecule studies of SNARE complex assembly reveal parallel and antiparallel configurations. Proceedings of the National Academy of Sciences of the United States of America 100: 14800-14805.
24. Jahn R, Scheller R (2006) SNAREs — engines for membrane fusion. Nature Reviews Molecular Cell Biology 7: 631-643.
25. Cusack S, Berthet-Colominas C, Härtlein M, Nassar N, Leberman R (1990) A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 A. Nature 347: 249-255.
26. Zizlsperger N, Malashkevich VN, Pillay S, Keating AE (2008) Analysis of coiled-coil interactions between core proteins of the spindle pole body. Biochemistry 47: 11858-11868.
27. Alber F, Förster F, Korkin D, Topf M, Sali A (2008) Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. Annu Rev Biochem.
28. Alber F, Dokudovskaya S, Veenhoff L, Zhang W, Kipper J, et al. (2007) The molecular architecture of the nuclear pore complex. Nature 450: 695-701.
29. Ranganathan R, Ross E (1997) PDZ domain proteins: Scaffolds for signaling complexes. Current Biology 7: R770-R773.
30. Park SH, Zarrinpar A, Lim WA (2003) Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. Science 299: 1061-1064.
31. Bashor C, Helman N, Yan S, Lim W (2008) Using Engineered Scaffold Interactions to Reshape MAP Kinase Pathway Signaling Dynamics. Science 319: 1539-1543.
32. Grigoryan G, Reinke A, Keating A (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature 458: 859-864.
33. Newman JR, Keating AE (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science 300: 2097-2101.

Appendix A

Residue frequencies from NPS database

Database construction as described in section 2.3.1. Frequencies reported as "*raw / normalized* (*counts*)", where *raw* represents the fraction of the specified residue over all residues of that heptad position, *normalized* is the raw frequency divided by the average frequency of occurrence of that residue in Genbank as reported in [1], and *counts* is the number of occurrences of the specified residue in the specified heptad position. Frequencies are reported over the entire database ("All"), and over each structural subclass: parallel dimers, parallel trimers, parallel tetramers, antiparallel dimers, antiparallel trimers and antiparallel tetramers.

All : 2105 strands 124034 residues

	a	b	c	d	e	f	g
L	0.267 / 2.859 (4770)	0.036 / 0.388 (643)	0.033 / 0.356 (588)	0.422 / 4.518 (7565)	0.054 / 0.578 (945)	0.040 / 0.432 (710)	0.055 / 0.591 (970)
I	0.141 / 2.628 (2515)	0.014 / 0.260 (247)	0.018 / 0.336 (318)	0.051 / 0.949 (911)	0.028 / 0.522 (490)	0.020 / 0.383 (361)	0.025 / 0.458 (431)
V	0.130 / 2.027 (2328)	0.025 / 0.395 (450)	0.025 / 0.386 (439)	0.059 / 0.918 (1058)	0.027 / 0.415 (467)	0.025 / 0.397 (449)	0.028 / 0.444 (501)
M	0.034 / 1.467 (614)	0.013 / 0.542 (225)	0.014 / 0.582 (241)	0.039 / 1.681 (706)	0.014 / 0.619 (254)	0.013 / 0.560 (231)	0.020 / 0.848 (349)
F	0.021 / 0.552 (383)	0.005 / 0.139 (96)	0.013 / 0.335 (230)	0.025 / 0.653 (455)	0.006 / 0.165 (112)	0.005 / 0.132 (90)	0.004 / 0.100 (68)
Y	0.035 / 1.115 (630)	0.004 / 0.121 (68)	0.007 / 0.211 (118)	0.035 / 1.109 (629)	0.005 / 0.173 (96)	0.005 / 0.171 (95)	0.008 / 0.239 (133)
G	0.007 / 0.095 (120)	0.028 / 0.397 (501)	0.031 / 0.443 (557)	0.006 / 0.089 (113)	0.024 / 0.336 (418)	0.035 / 0.487 (610)	0.014 / 0.192 (240)
A	0.075 / 0.983 (1335)	0.108 / 1.418 (1911)	0.083 / 1.100 (1477)	0.130 / 1.713 (2333)	0.043 / 0.561 (746)	0.113 / 1.486 (1988)	0.077 / 1.020 (1362)
K	0.069 / 1.208 (1236)	0.099 / 1.732 (1759)	0.091 / 1.594 (1614)	0.023 / 0.394 (404)	0.112 / 1.951 (1957)	0.114 / 1.993 (2009)	0.114 / 1.987 (1999)
R	0.049 / 0.918 (885)	0.074 / 1.377 (1318)	0.077 / 1.421 (1356)	0.011 / 0.206 (199)	0.095 / 1.765 (1668)	0.097 / 1.798 (1708)	0.092 / 1.703 (1614)
H	0.013 / 0.559 (225)	0.017 / 0.741 (296)	0.018 / 0.791 (315)	0.011 / 0.490 (198)	0.014 / 0.634 (250)	0.017 / 0.734 (291)	0.013 / 0.579 (229)
E	0.017 / 0.273 (298)	0.181 / 2.967 (3214)	0.191 / 3.131 (3380)	0.050 / 0.814 (891)	0.226 / 3.701 (3959)	0.151 / 2.467 (2653)	0.214 / 3.515 (3771)
D	0.002 / 0.040 (36)	0.105 / 2.094 (1870)	0.097 / 1.924 (1713)	0.006 / 0.120 (108)	0.040 / 0.796 (702)	0.075 / 1.491 (1322)	0.063 / 1.254 (1109)
Q	0.016 / 0.385 (294)	0.099 / 2.328 (1765)	0.090 / 2.097 (1585)	0.027 / 0.633 (485)	0.138 / 3.223 (2413)	0.083 / 1.952 (1469)	0.139 / 3.262 (2450)
N	0.055 / 1.300 (988)	0.063 / 1.476 (1114)	0.068 / 1.597 (1201)	0.018 / 0.425 (324)	0.056 / 1.323 (986)	0.062 / 1.452 (1088)	0.036 / 0.848 (634)
S	0.028 / 0.382 (497)	0.079 / 1.090 (1409)	0.088 / 1.207 (1555)	0.033 / 0.449 (586)	0.056 / 0.765 (977)	0.084 / 1.151 (1477)	0.052 / 0.708 (906)
T	0.025 / 0.413 (441)	0.045 / 0.756 (801)	0.050 / 0.842 (890)	0.041 / 0.683 (732)	0.059 / 0.988 (1034)	0.054 / 0.901 (948)	0.043 / 0.712 (748)
C	0.014 / 0.758 (252)	0.002 / 0.124 (41)	0.004 / 0.210 (69)	0.006 / 0.321 (107)	0.001 / 0.067 (22)	0.004 / 0.204 (67)	0.003 / 0.153 (50)
W	0.002 / 0.139 (35)	0.001 / 0.060 (15)	0.001 / 0.100 (25)	0.008 / 0.561 (142)	0.002 / 0.142 (35)	0.002 / 0.133 (33)	0.001 / 0.089 (22)
P	0.000 / 0.002 (2)	0.000 / 0.004 (4)	0.001 / 0.027 (25)	0.000 / 0.000 (0)	0.000 / 0.002 (2)	0.001 / 0.020 (19)	0.000 / 0.000 (0)

Parallel Dimer : 1332 strands 94516 residues

	a	b	c	d	e	f	g
L	0.279 / 2.987 (3780)	0.031 / 0.335 (421)	0.030 / 0.317 (398)	0.451 / 4.831 (6101)	0.053 / 0.565 (709)	0.040 / 0.432 (546)	0.049 / 0.523 (658)
I	0.129 / 2.416 (1753)	0.013 / 0.237 (171)	0.017 / 0.321 (231)	0.033 / 0.625 (453)	0.025 / 0.467 (336)	0.022 / 0.407 (295)	0.019 / 0.363 (262)
V	0.123 / 1.912 (1665)	0.026 / 0.399 (345)	0.025 / 0.386 (333)	0.049 / 0.757 (658)	0.021 / 0.325 (281)	0.025 / 0.384 (334)	0.024 / 0.376 (325)
M	0.029 / 1.257 (399)	0.012 / 0.495 (156)	0.013 / 0.569 (179)	0.034 / 1.440 (456)	0.015 / 0.645 (203)	0.014 / 0.599 (190)	0.015 / 0.637 (201)
F	0.019 / 0.492 (259)	0.005 / 0.134 (70)	0.014 / 0.364 (190)	0.024 / 0.617 (324)	0.006 / 0.144 (75)	0.005 / 0.129 (68)	0.002 / 0.048 (25)
Y	0.037 / 1.171 (502)	0.003 / 0.080 (34)	0.006 / 0.200 (85)	0.040 / 1.269 (543)	0.004 / 0.132 (56)	0.004 / 0.140 (60)	0.005 / 0.153 (65)
G	0.006 / 0.078 (75)	0.024 / 0.345 (330)	0.028 / 0.389 (372)	0.006 / 0.078 (75)	0.015 / 0.214 (204)	0.034 / 0.473 (455)	0.010 / 0.138 (132)
A	0.064 / 0.848 (873)	0.118 / 1.557 (1593)	0.080 / 1.051 (1073)	0.131 / 1.731 (1779)	0.037 / 0.482 (492)	0.114 / 1.499 (1542)	0.069 / 0.913 (934)
K	0.085 / 1.479 (1147)	0.099 / 1.739 (1341)	0.090 / 1.580 (1216)	0.024 / 0.415 (321)	0.115 / 2.017 (1552)	0.119 / 2.088 (1619)	0.119 / 2.082 (1605)
R	0.057 / 1.057 (773)	0.075 / 1.400 (1017)	0.076 / 1.406 (1020)	0.008 / 0.152 (111)	0.101 / 1.873 (1358)	0.104 / 1.924 (1406)	0.092 / 1.714 (1245)
H	0.014 / 0.632 (193)	0.017 / 0.772 (234)	0.019 / 0.832 (252)	0.010 / 0.450 (137)	0.014 / 0.611 (185)	0.016 / 0.718 (219)	0.012 / 0.538 (163)
E	0.017 / 0.277 (229)	0.195 / 3.200 (2631)	0.206 / 3.381 (2775)	0.061 / 0.999 (825)	0.249 / 4.077 (3345)	0.152 / 2.485 (2055)	0.240 / 3.928 (3229)
D	0.002 / 0.037 (25)	0.103 / 2.046 (1387)	0.100 / 1.992 (1348)	0.007 / 0.129 (88)	0.034 / 0.667 (451)	0.071 / 1.406 (959)	0.065 / 1.289 (874)
Q	0.011 / 0.266 (154)	0.103 / 2.420 (1393)	0.089 / 2.075 (1192)	0.025 / 0.595 (344)	0.143 / 3.359 (1929)	0.079 / 1.861 (1077)	0.151 / 3.545 (2040)
N	0.067 / 1.574 (907)	0.054 / 1.276 (731)	0.061 / 1.441 (824)	0.013 / 0.301 (173)	0.053 / 1.237 (707)	0.057 / 1.343 (774)	0.034 / 0.810 (464)
S	0.024 / 0.330 (326)	0.076 / 1.047 (1027)	0.089 / 1.226 (1201)	0.034 / 0.464 (457)	0.052 / 0.716 (701)	0.086 / 1.182 (1166)	0.048 / 0.659 (647)
T	0.019 / 0.315 (255)	0.041 / 0.689 (554)	0.049 / 0.820 (659)	0.035 / 0.591 (478)	0.062 / 1.044 (838)	0.052 / 0.879 (711)	0.042 / 0.707 (569)
C	0.017 / 0.892 (225)	0.003 / 0.152 (38)	0.005 / 0.260 (65)	0.007 / 0.353 (89)	0.001 / 0.064 (16)	0.004 / 0.234 (59)	0.002 / 0.120 (30)
W	0.002 / 0.110 (21)	0.000 / 0.016 (3)	0.001 / 0.105 (20)	0.009 / 0.655 (125)	0.001 / 0.063 (12)	0.001 / 0.047 (9)	0.001 / 0.047 (9)
P	0.000 / 0.001 (1)	0.000 / 0.003 (2)	0.002 / 0.031 (22)	0.000 / 0.000 (0)	0.000 / 0.001 (1)	0.001 / 0.017 (12)	0.000 / 0.000 (0)

Parallel Trimer : 354 strands 14544 residues

	a	b	c	d	e	f	g
L	0.255 / 2.731 (532)	0.049 / 0.520 (103)	0.054 / 0.579 (113)	0.340 / 3.647 (718)	0.069 / 0.743 (141)	0.043 / 0.460 (87)	0.061 / 0.653 (126)
I	0.187 / 3.500 (391)	0.013 / 0.247 (28)	0.020 / 0.375 (42)	0.105 / 1.967 (222)	0.035 / 0.652 (71)	0.016 / 0.304 (33)	0.020 / 0.370 (41)
V	0.170 / 2.648 (355)	0.023 / 0.360 (49)	0.026 / 0.402 (54)	0.089 / 1.388 (188)	0.041 / 0.643 (84)	0.030 / 0.461 (60)	0.036 / 0.557 (74)
M	0.029 / 1.248 (61)	0.016 / 0.685 (34)	0.012 / 0.531 (26)	0.059 / 2.511 (124)	0.013 / 0.567 (27)	0.008 / 0.358 (17)	0.021 / 0.909 (44)
F	0.024 / 0.630 (51)	0.007 / 0.170 (14)	0.010 / 0.246 (20)	0.021 / 0.550 (45)	0.013 / 0.329 (26)	0.002 / 0.064 (5)	0.010 / 0.249 (20)
Y	0.021 / 0.652 (43)	0.008 / 0.268 (18)	0.005 / 0.151 (10)	0.011 / 0.360 (24)	0.004 / 0.140 (9)	0.004 / 0.125 (8)	0.018 / 0.566 (37)
G	0.013 / 0.182 (27)	0.038 / 0.531 (80)	0.058 / 0.815 (121)	0.007 / 0.100 (15)	0.030 / 0.429 (62)	0.044 / 0.625 (90)	0.029 / 0.415 (61)
A	0.123 / 1.622 (257)	0.078 / 1.031 (166)	0.088 / 1.159 (184)	0.125 / 1.642 (263)	0.059 / 0.784 (121)	0.086 / 1.130 (174)	0.101 / 1.337 (210)
K	0.015 / 0.260 (31)	0.097 / 1.697 (206)	0.083 / 1.454 (174)	0.024 / 0.423 (51)	0.108 / 1.891 (220)	0.105 / 1.835 (213)	0.090 / 1.572 (186)
R	0.007 / 0.124 (14)	0.072 / 1.329 (152)	0.053 / 0.984 (111)	0.017 / 0.308 (35)	0.087 / 1.605 (176)	0.085 / 1.582 (173)	0.099 / 1.838 (205)
H	0.007 / 0.319 (15)	0.011 / 0.482 (23)	0.014 / 0.616 (29)	0.015 / 0.674 (32)	0.012 / 0.524 (24)	0.021 / 0.942 (43)	0.013 / 0.559 (26)
E	0.005 / 0.086 (11)	0.147 / 2.410 (312)	0.131 / 2.155 (275)	0.011 / 0.186 (24)	0.171 / 2.797 (347)	0.148 / 2.432 (301)	0.154 / 2.528 (319)
D	0.002 / 0.038 (4)	0.098 / 1.939 (207)	0.092 / 1.834 (193)	0.006 / 0.122 (13)	0.042 / 0.831 (85)	0.076 / 1.509 (154)	0.076 / 1.518 (158)
Q	0.039 / 0.920 (82)	0.102 / 2.395 (217)	0.104 / 2.429 (217)	0.019 / 0.455 (41)	0.126 / 2.948 (256)	0.108 / 2.528 (219)	0.105 / 2.468 (218)
N	0.027 / 0.631 (56)	0.077 / 1.818 (164)	0.086 / 2.025 (180)	0.041 / 0.970 (87)	0.066 / 1.562 (135)	0.084 / 1.971 (170)	0.044 / 1.035 (91)
S	0.027 / 0.368 (56)	0.094 / 1.288 (199)	0.101 / 1.385 (211)	0.034 / 0.469 (72)	0.066 / 0.912 (135)	0.066 / 0.907 (134)	0.070 / 0.963 (145)
T	0.038 / 0.642 (80)	0.066 / 1.113 (141)	0.062 / 1.041 (130)	0.071 / 1.191 (150)	0.047 / 0.782 (95)	0.058 / 0.974 (118)	0.042 / 0.704 (87)
C	0.008 / 0.438 (17)	0.000 / 0.000 (0)	0.000 / 0.026 (1)	0.001 / 0.076 (3)	0.000 / 0.026 (1)	0.002 / 0.132 (5)	0.007 / 0.390 (15)
W	0.002 / 0.170 (5)	0.002 / 0.167 (5)	0.000 / 0.034 (1)	0.001 / 0.101 (3)	0.009 / 0.662 (19)	0.010 / 0.699 (20)	0.003 / 0.206 (6)
P	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.009 (1)	0.000 / 0.000 (0)

Parallel Tetramer : 122 strands 5794 residues

	a	b	c	d	e	f	g
L	0.238 / 2.554 (208)	0.086 / 0.917 (74)	0.015 / 0.161 (13)	0.270 / 2.891 (236)	0.043 / 0.458 (33)	0.016 / 0.167 (12)	0.066 / 0.708 (51)
I	0.148 / 2.762 (129)	0.027 / 0.497 (23)	0.008 / 0.151 (7)	0.131 / 2.457 (115)	0.043 / 0.799 (33)	0.012 / 0.218 (9)	0.082 / 1.525 (63)
V	0.222 / 3.461 (194)	0.032 / 0.504 (28)	0.020 / 0.306 (17)	0.115 / 1.798 (101)	0.083 / 1.291 (64)	0.032 / 0.505 (25)	0.039 / 0.605 (30)
M	0.079 / 3.378 (69)	0.027 / 1.136 (23)	0.012 / 0.493 (10)	0.045 / 1.905 (39)	0.008 / 0.332 (6)	0.010 / 0.443 (8)	0.067 / 2.879 (52)
F	0.049 / 1.269 (43)	0.008 / 0.209 (7)	0.002 / 0.060 (2)	0.051 / 1.325 (45)	0.004 / 0.100 (3)	0.005 / 0.134 (4)	0.010 / 0.267 (8)
Y	0.001 / 0.036 (1)	0.008 / 0.256 (7)	0.005 / 0.146 (4)	0.011 / 0.362 (10)	0.013 / 0.410 (10)	0.018 / 0.575 (14)	0.014 / 0.451 (11)
G	0.011 / 0.161 (10)	0.039 / 0.554 (34)	0.020 / 0.276 (17)	0.021 / 0.290 (18)	0.110 / 1.551 (85)	0.048 / 0.676 (37)	0.027 / 0.383 (21)
A	0.119 / 1.570 (104)	0.043 / 0.564 (37)	0.051 / 0.669 (44)	0.094 / 1.235 (82)	0.065 / 0.853 (50)	0.091 / 1.196 (70)	0.027 / 0.358 (21)
K	0.002 / 0.040 (2)	0.082 / 1.435 (71)	0.109 / 1.898 (94)	0.006 / 0.100 (5)	0.069 / 1.200 (53)	0.092 / 1.610 (71)	0.148 / 2.582 (114)
R	0.013 / 0.234 (11)	0.079 / 1.458 (68)	0.104 / 1.928 (90)	0.048 / 0.891 (42)	0.065 / 1.202 (50)	0.053 / 0.987 (41)	0.080 / 1.490 (62)
H	0.001 / 0.051 (1)	0.027 / 1.182 (23)	0.021 / 0.924 (18)	0.000 / 0.000 (0)	0.035 / 1.554 (27)	0.013 / 0.576 (10)	0.021 / 0.921 (16)
E	0.002 / 0.038 (2)	0.111 / 1.819 (96)	0.203 / 3.332 (176)	0.002 / 0.037 (2)	0.122 / 1.996 (94)	0.162 / 2.658 (125)	0.084 / 1.380 (65)
D	0.000 / 0.000 (0)	0.160 / 3.172 (138)	0.140 / 2.778 (121)	0.000 / 0.000 (0)	0.073 / 1.442 (56)	0.139 / 2.759 (107)	0.047 / 0.927 (36)
Q	0.001 / 0.027 (1)	0.087 / 2.031 (75)	0.119 / 2.785 (103)	0.067 / 1.579 (59)	0.089 / 2.093 (69)	0.093 / 2.187 (72)	0.115 / 2.700 (89)
N	0.003 / 0.081 (3)	0.054 / 1.278 (47)	0.074 / 1.739 (64)	0.053 / 1.237 (46)	0.073 / 1.707 (56)	0.062 / 1.465 (48)	0.070 / 1.646 (54)
S	0.050 / 0.692 (44)	0.081 / 1.112 (70)	0.064 / 0.872 (55)	0.021 / 0.283 (18)	0.067 / 0.925 (52)	0.083 / 1.140 (64)	0.073 / 0.996 (56)
T	0.050 / 0.844 (44)	0.043 / 0.716 (37)	0.033 / 0.561 (29)	0.056 / 0.938 (49)	0.034 / 0.564 (26)	0.062 / 1.043 (48)	0.027 / 0.456 (21)
C	0.001 / 0.062 (1)	0.001 / 0.062 (1)	0.000 / 0.000 (0)	0.005 / 0.246 (4)	0.005 / 0.279 (4)	0.001 / 0.070 (1)	0.001 / 0.070 (1)
W	0.006 / 0.406 (5)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.003 / 0.243 (3)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)
P	0.000 / 0.000 (0)	0.001 / 0.022 (1)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)

Antiparallel Dimer : 276 strands 8525 residues

	a	b	c	d	e	f	g
L	0.175 / 1.878 (222)	0.033 / 0.357 (40)	0.048 / 0.512 (57)	0.364 / 3.906 (484)	0.044 / 0.472 (52)	0.048 / 0.519 (57)	0.105 / 1.129 (124)
I	0.180 / 3.364 (228)	0.017 / 0.311 (20)	0.031 / 0.580 (37)	0.084 / 1.562 (111)	0.041 / 0.775 (49)	0.019 / 0.349 (22)	0.051 / 0.953 (60)
V	0.077 / 1.205 (98)	0.021 / 0.324 (25)	0.027 / 0.418 (32)	0.076 / 1.185 (101)	0.029 / 0.448 (34)	0.023 / 0.357 (27)	0.058 / 0.900 (68)
M	0.063 / 2.698 (80)	0.008 / 0.356 (10)	0.020 / 0.860 (24)	0.066 / 2.800 (87)	0.012 / 0.506 (14)	0.014 / 0.581 (16)	0.042 / 1.815 (50)
F	0.021 / 0.549 (27)	0.003 / 0.086 (4)	0.013 / 0.346 (16)	0.028 / 0.718 (37)	0.005 / 0.131 (6)	0.009 / 0.241 (11)	0.010 / 0.263 (12)
Y	0.066 / 2.073 (83)	0.007 / 0.211 (8)	0.014 / 0.451 (17)	0.038 / 1.191 (50)	0.012 / 0.375 (14)	0.011 / 0.350 (13)	0.016 / 0.511 (19)
G	0.006 / 0.089 (8)	0.047 / 0.657 (56)	0.038 / 0.531 (45)	0.004 / 0.053 (5)	0.057 / 0.798 (67)	0.021 / 0.299 (25)	0.021 / 0.299 (25)
A	0.074 / 0.977 (94)	0.090 / 1.185 (108)	0.137 / 1.811 (164)	0.151 / 1.984 (200)	0.063 / 0.836 (75)	0.165 / 2.172 (194)	0.163 / 2.149 (192)
K	0.043 / 0.759 (55)	0.110 / 1.921 (132)	0.106 / 1.861 (127)	0.020 / 0.342 (26)	0.104 / 1.819 (123)	0.085 / 1.485 (100)	0.071 / 1.248 (84)
R	0.067 / 1.245 (85)	0.061 / 1.128 (73)	0.111 / 2.053 (132)	0.003 / 0.056 (4)	0.061 / 1.130 (72)	0.065 / 1.214 (77)	0.083 / 1.545 (98)
H	0.013 / 0.561 (16)	0.011 / 0.481 (13)	0.012 / 0.522 (14)	0.022 / 0.971 (29)	0.010 / 0.451 (12)	0.013 / 0.566 (15)	0.018 / 0.793 (21)
E	0.043 / 0.699 (54)	0.133 / 2.184 (160)	0.117 / 1.924 (140)	0.026 / 0.432 (35)	0.136 / 2.233 (161)	0.137 / 2.242 (161)	0.124 / 2.034 (146)
D	0.006 / 0.110 (7)	0.110 / 2.185 (132)	0.038 / 0.750 (45)	0.002 / 0.045 (3)	0.082 / 1.631 (97)	0.077 / 1.537 (91)	0.025 / 0.507 (30)
Q	0.043 / 0.998 (54)	0.062 / 1.443 (74)	0.051 / 1.197 (61)	0.027 / 0.635 (36)	0.130 / 3.051 (154)	0.079 / 1.850 (93)	0.083 / 1.950 (98)
N	0.017 / 0.390 (21)	0.139 / 3.272 (167)	0.107 / 2.525 (128)	0.012 / 0.283 (16)	0.074 / 1.752 (88)	0.075 / 1.759 (88)	0.020 / 0.460 (23)
S	0.054 / 0.737 (68)	0.087 / 1.189 (104)	0.066 / 0.910 (79)	0.025 / 0.341 (33)	0.073 / 0.999 (86)	0.089 / 1.225 (105)	0.046 / 0.630 (54)
T	0.046 / 0.767 (58)	0.056 / 0.934 (67)	0.057 / 0.955 (68)	0.038 / 0.631 (50)	0.063 / 1.063 (75)	0.059 / 0.996 (70)	0.057 / 0.954 (67)
C	0.003 / 0.170 (4)	0.001 / 0.045 (1)	0.001 / 0.045 (1)	0.008 / 0.445 (11)	0.001 / 0.045 (1)	0.002 / 0.091 (2)	0.003 / 0.137 (3)
W	0.003 / 0.224 (4)	0.005 / 0.354 (6)	0.003 / 0.238 (4)	0.008 / 0.534 (10)	0.001 / 0.060 (1)	0.003 / 0.241 (4)	0.003 / 0.181 (3)
P	0.001 / 0.015 (1)	0.001 / 0.016 (1)	0.002 / 0.032 (2)	0.000 / 0.000 (0)	0.001 / 0.016 (1)	0.005 / 0.097 (6)	0.000 / 0.000 (0)

Antiparallel Trimer : 15 strands 419 residues

	a	b	c	d	e	f	g
L	0.283 / 3.037 (17)	0.051 / 0.545 (3)	0.115 / 1.230 (7)	0.194 / 2.074 (12)	0.117 / 1.250 (7)	0.100 / 1.072 (6)	0.088 / 0.940 (5)
I	0.167 / 3.115 (10)	0.085 / 1.584 (5)	0.016 / 0.306 (1)	0.097 / 1.809 (6)	0.017 / 0.312 (1)	0.017 / 0.312 (1)	0.035 / 0.656 (2)
V	0.200 / 3.115 (12)	0.051 / 0.792 (3)	0.033 / 0.511 (2)	0.113 / 1.759 (7)	0.050 / 0.779 (3)	0.050 / 0.779 (3)	0.035 / 0.547 (2)
M	0.067 / 2.849 (4)	0.017 / 0.724 (1)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.033 / 1.425 (2)	0.000 / 0.000 (0)	0.000 / 0.000 (0)
F	0.033 / 0.859 (2)	0.017 / 0.437 (1)	0.000 / 0.000 (0)	0.065 / 1.663 (4)	0.033 / 0.859 (2)	0.017 / 0.430 (1)	0.053 / 1.356 (3)
Y	0.017 / 0.527 (1)	0.000 / 0.000 (0)	0.033 / 1.038 (2)	0.032 / 1.021 (2)	0.083 / 2.637 (5)	0.000 / 0.000 (0)	0.018 / 0.555 (1)
G	0.000 / 0.000 (0)	0.017 / 0.239 (1)	0.033 / 0.462 (2)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.050 / 0.704 (3)	0.000 / 0.000 (0)
A	0.017 / 0.220 (1)	0.068 / 0.893 (4)	0.148 / 1.944 (9)	0.097 / 1.275 (6)	0.017 / 0.220 (1)	0.100 / 1.318 (6)	0.035 / 0.462 (2)
K	0.017 / 0.291 (1)	0.119 / 2.074 (7)	0.033 / 0.573 (2)	0.000 / 0.000 (0)	0.083 / 1.457 (5)	0.067 / 1.166 (4)	0.123 / 2.147 (7)
R	0.033 / 0.618 (2)	0.102 / 1.887 (6)	0.016 / 0.304 (1)	0.113 / 2.095 (7)	0.100 / 1.855 (6)	0.117 / 2.165 (7)	0.053 / 0.976 (3)
H	0.000 / 0.000 (0)	0.034 / 1.507 (2)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.017 / 0.741 (1)	0.033 / 1.481 (2)	0.035 / 1.559 (2)
E	0.017 / 0.273 (1)	0.136 / 2.223 (8)	0.115 / 1.881 (7)	0.065 / 1.058 (4)	0.167 / 2.732 (10)	0.067 / 1.093 (4)	0.175 / 2.876 (10)
D	0.000 / 0.000 (0)	0.051 / 1.011 (3)	0.066 / 1.304 (4)	0.048 / 0.962 (3)	0.167 / 3.313 (10)	0.117 / 2.319 (7)	0.140 / 2.790 (8)
Q	0.017 / 0.390 (1)	0.068 / 1.588 (4)	0.131 / 3.071 (8)	0.065 / 1.511 (4)	0.033 / 0.781 (2)	0.050 / 1.171 (3)	0.035 / 0.822 (2)
N	0.000 / 0.000 (0)	0.051 / 1.196 (3)	0.082 / 1.929 (5)	0.016 / 0.380 (1)	0.000 / 0.000 (0)	0.117 / 2.745 (7)	0.035 / 0.826 (2)
S	0.033 / 0.458 (2)	0.102 / 1.397 (6)	0.098 / 1.351 (6)	0.032 / 0.443 (2)	0.033 / 0.458 (2)	0.083 / 1.145 (5)	0.018 / 0.241 (1)
T	0.050 / 0.838 (3)	0.017 / 0.284 (1)	0.049 / 0.824 (3)	0.048 / 0.811 (3)	0.000 / 0.000 (0)	0.017 / 0.279 (1)	0.053 / 0.882 (3)
C	0.050 / 2.688 (3)	0.000 / 0.000 (0)	0.016 / 0.881 (1)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)
W	0.000 / 0.000 (0)	0.017 / 1.202 (1)	0.000 / 0.000 (0)	0.016 / 1.144 (1)	0.050 / 3.546 (3)	0.000 / 0.000 (0)	0.070 / 4.977 (4)
P	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.016 / 0.310 (1)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)

Antiparallel Tetramer : 6 strands 236 residues

	a	b	c	d	e	f	g
L	0.314 / 3.369 (11)	0.065 / 0.691 (2)	0.000 / 0.000 (0)	0.400 / 4.287 (14)	0.086 / 0.919 (3)	0.059 / 0.630 (2)	0.171 / 1.837 (6)
I	0.114 / 2.136 (4)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.114 / 2.136 (4)	0.000 / 0.000 (0)	0.029 / 0.550 (1)	0.086 / 1.602 (3)
V	0.114 / 1.780 (4)	0.000 / 0.000 (0)	0.032 / 0.502 (1)	0.086 / 1.335 (3)	0.029 / 0.445 (1)	0.000 / 0.000 (0)	0.057 / 0.890 (2)
M	0.029 / 1.221 (1)	0.032 / 1.379 (1)	0.065 / 2.757 (2)	0.000 / 0.000 (0)	0.057 / 2.442 (2)	0.000 / 0.000 (0)	0.057 / 2.442 (2)
F	0.029 / 0.736 (1)	0.000 / 0.000 (0)	0.065 / 1.663 (2)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.029 / 0.758 (1)	0.000 / 0.000 (0)
Y	0.000 / 0.000 (0)	0.032 / 1.021 (1)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.057 / 1.808 (2)	0.000 / 0.000 (0)	0.000 / 0.000 (0)
G	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.029 / 0.402 (1)
A	0.171 / 2.259 (6)	0.097 / 1.275 (3)	0.097 / 1.275 (3)	0.086 / 1.129 (3)	0.200 / 2.635 (7)	0.059 / 0.775 (2)	0.086 / 1.129 (3)
K	0.000 / 0.000 (0)	0.065 / 1.128 (2)	0.032 / 0.564 (1)	0.029 / 0.500 (1)	0.114 / 1.998 (4)	0.059 / 1.028 (2)	0.086 / 1.499 (3)
R	0.000 / 0.000 (0)	0.065 / 1.197 (2)	0.065 / 1.197 (2)	0.000 / 0.000 (0)	0.171 / 3.180 (6)	0.118 / 2.183 (4)	0.029 / 0.530 (1)
H	0.000 / 0.000 (0)	0.032 / 1.434 (1)	0.065 / 2.867 (2)	0.000 / 0.000 (0)	0.029 / 1.270 (1)	0.059 / 2.614 (2)	0.029 / 1.270 (1)
E	0.029 / 0.468 (1)	0.226 / 3.702 (7)	0.226 / 3.702 (7)	0.029 / 0.468 (1)	0.057 / 0.937 (2)	0.206 / 3.375 (7)	0.057 / 0.937 (2)
D	0.000 / 0.000 (0)	0.097 / 1.924 (3)	0.065 / 1.283 (2)	0.029 / 0.568 (1)	0.086 / 1.704 (3)	0.118 / 2.339 (4)	0.086 / 1.704 (3)
Q	0.057 / 1.338 (2)	0.065 / 1.511 (2)	0.129 / 3.022 (4)	0.029 / 0.669 (1)	0.086 / 2.007 (3)	0.147 / 3.444 (5)	0.086 / 2.007 (3)
N	0.029 / 0.672 (1)	0.065 / 1.518 (2)	0.000 / 0.000 (0)	0.029 / 0.672 (1)	0.000 / 0.000 (0)	0.029 / 0.692 (1)	0.000 / 0.000 (0)
S	0.029 / 0.392 (1)	0.097 / 1.329 (3)	0.097 / 1.329 (3)	0.114 / 1.570 (4)	0.029 / 0.392 (1)	0.088 / 1.212 (3)	0.086 / 1.177 (3)
T	0.029 / 0.479 (1)	0.032 / 0.540 (1)	0.032 / 0.540 (1)	0.057 / 0.957 (2)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.029 / 0.479 (1)
C	0.057 / 3.072 (2)	0.032 / 1.734 (1)	0.032 / 1.734 (1)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.029 / 1.536 (1)
W	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)
P	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)	0.000 / 0.000 (0)

References

1. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science (New York, NY) 252: 1162-1164.

Appendix B

Supplementary material for Chapter 3 :

Predicting helix orientation for coiled-coil dimers

Author Contributions

This work was prepared in collaboration with James R. Apgar and Amy E. Keating. J.R.A. prepared explicit structure models, implemented the antiparallel Crick parameterization and computed structure-based statistics.

Attribution

Table B-1: List of PQS structures in the test set.

PQS ID	Strand 1	hep	Strand 2	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.[1] "Strand 1" and "Strand 2" indicate the residue numbers and chain for the first and second helix of the coiled coil. "hep" columns indicate the first heptad position of the first and second helices, respectively.

Table B-2: 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**).

Table B-3. 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.[2]
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 [3] and [4]

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 [5]

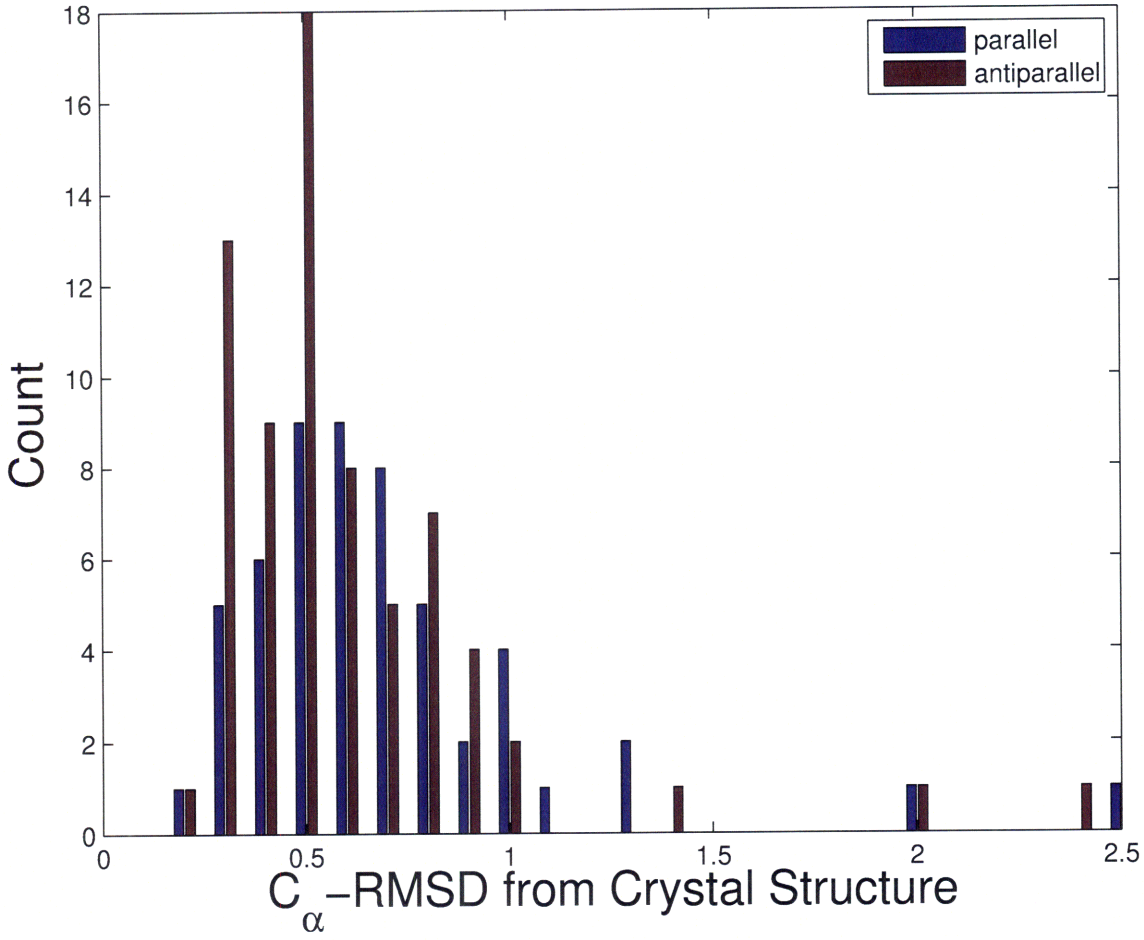
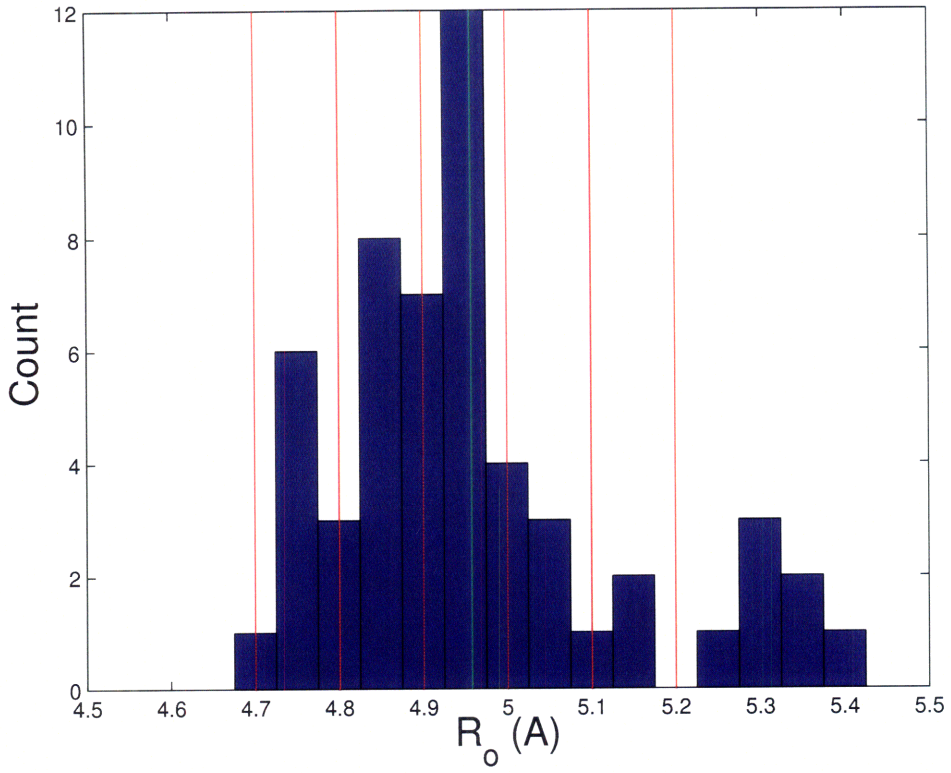


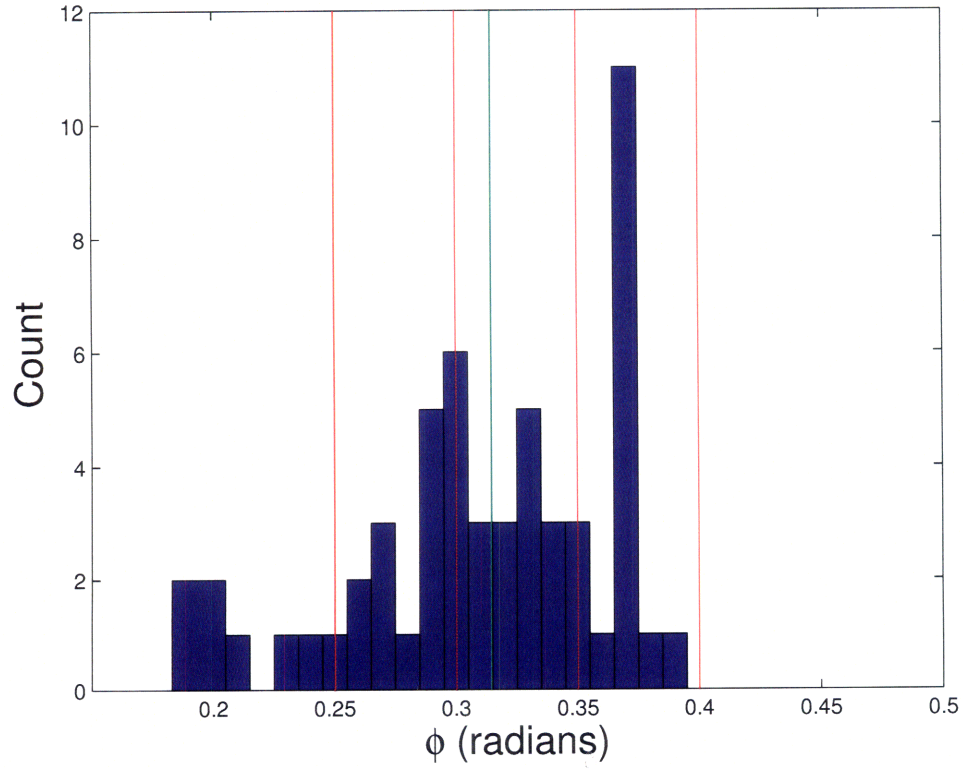
Figure B-1: Native coiled-coil variation described using Crick parameterization.

Histogram of the backbone RMSD (Cα-only) between each test-set structure and its closest Crick backbone.

a)


Count
12
10
8
6
4
2
0
4.5
4.6
4.7
4.8
4.9
5
5.1
5.2
5.3
5.4
5.5
R_o (A)

b)


Count
12
10
8
6
4
2
0
0.2
0.25
0.3
0.35
0.4
0.45
0.5
ϕ (radians)

c)

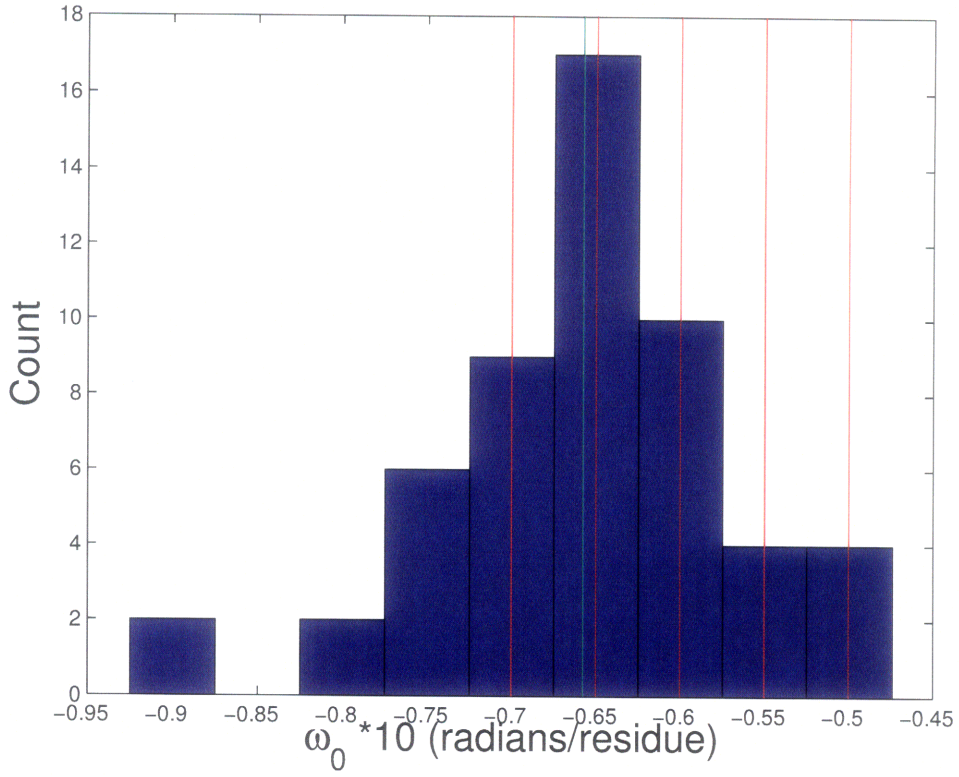
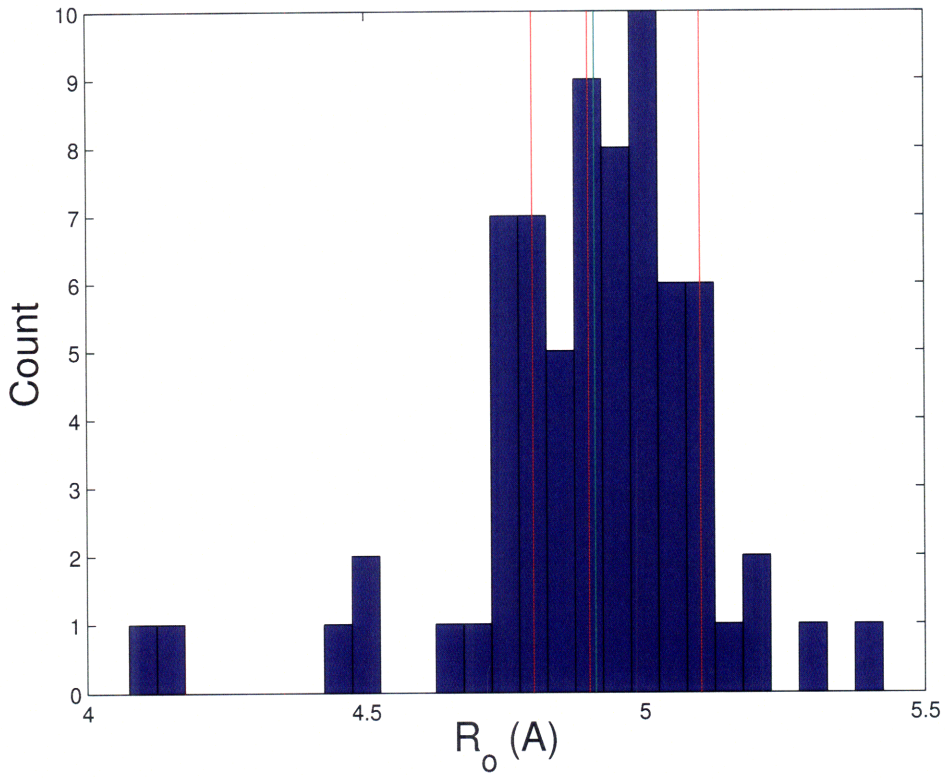
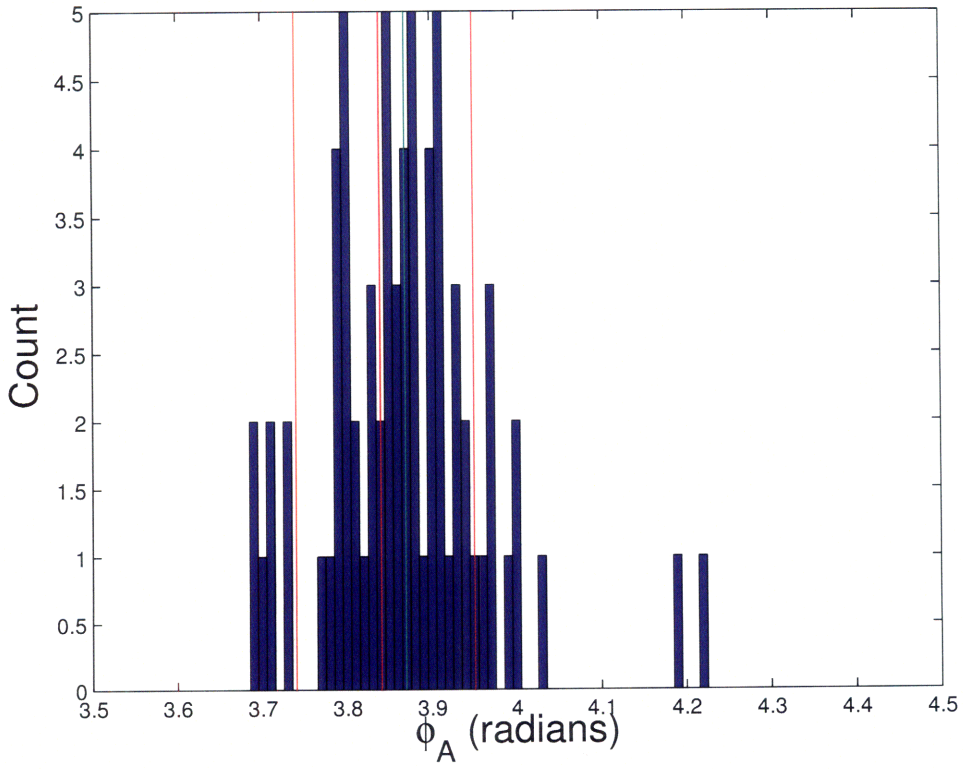


Figure B-2: Histogram of the parallel Crick parameters generated by fitting parallel test-set structures to the best possible Crick backbone. Parameters a) R_0, b) ϕ and c) ω_0 are shown. The mean value of the parameter is indicated in green, and in red are the values used to generate the parallel-structure 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.

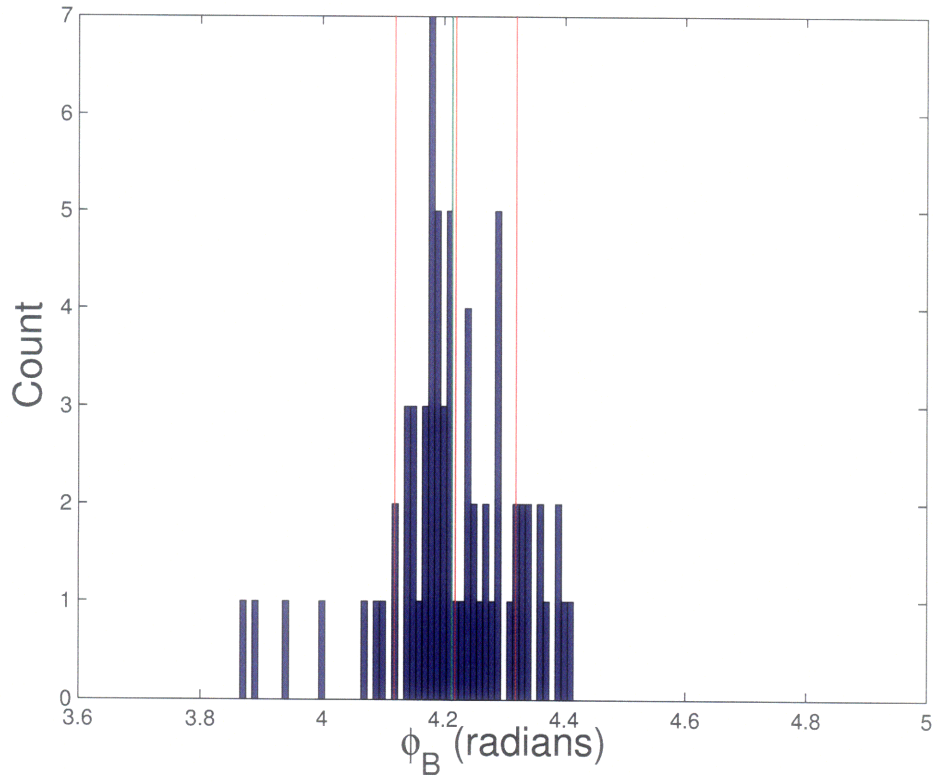
a)



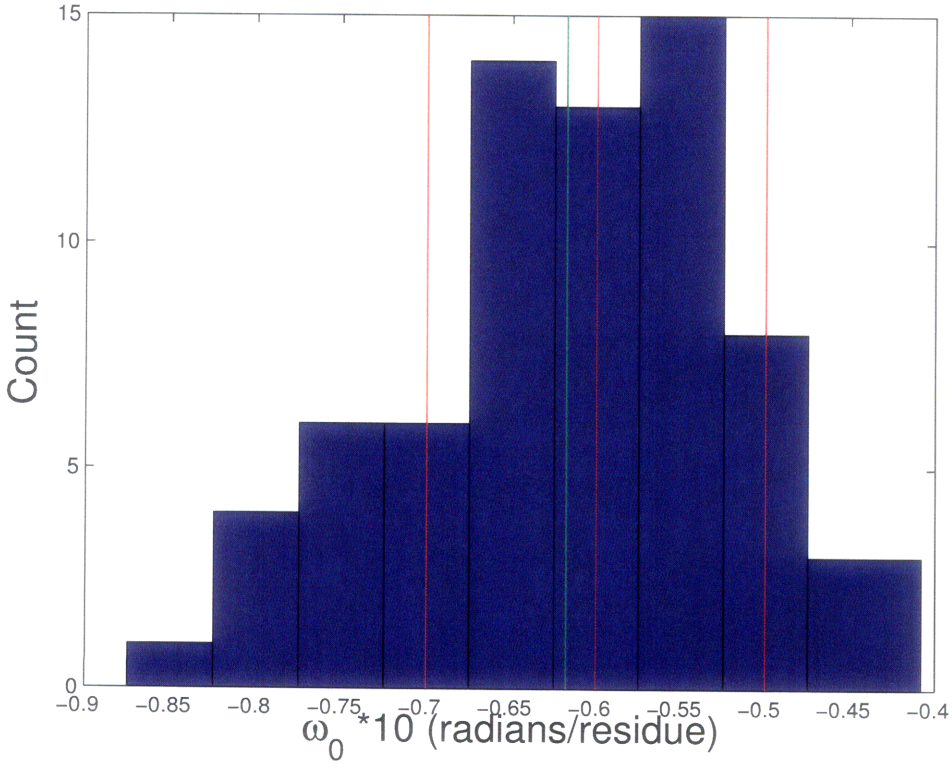
b)



c)



d)



e)

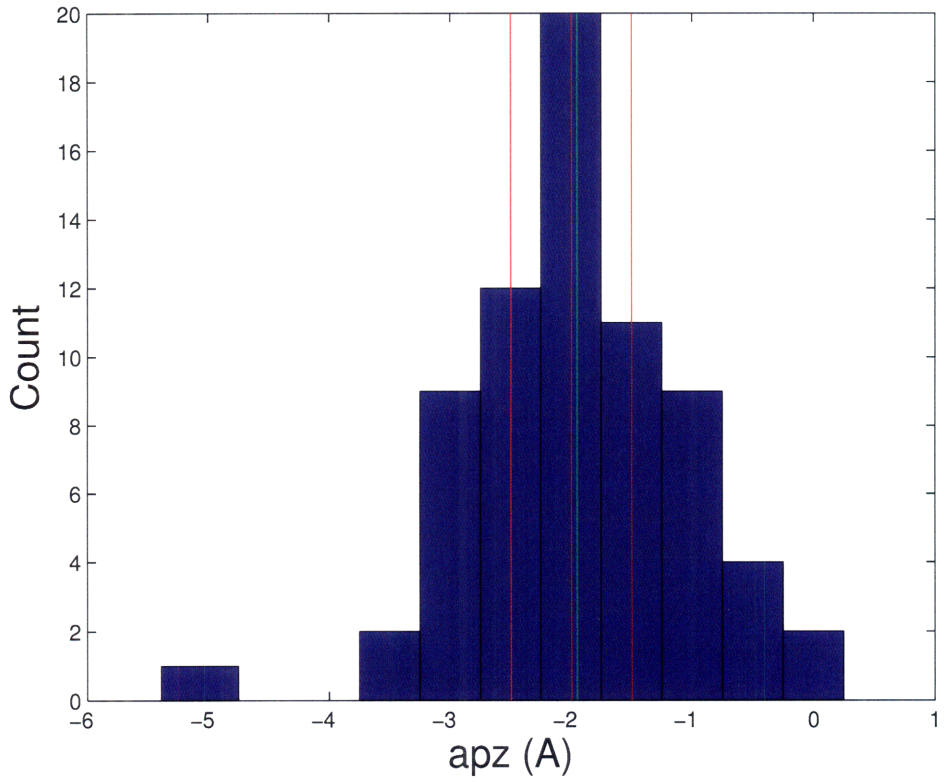


Figure B-3: Histogram of the antiparallel Crick parameters generated by fitting antiparallel test-set structures to the best possible Crick backbone. Parameters a) R_0 b) ϕ_A, c) ϕ_B d) ω_0 and e) apz are shown. The mean value of the parameter is indicated in green, and in red are the sampled values, as in Figure B-2.

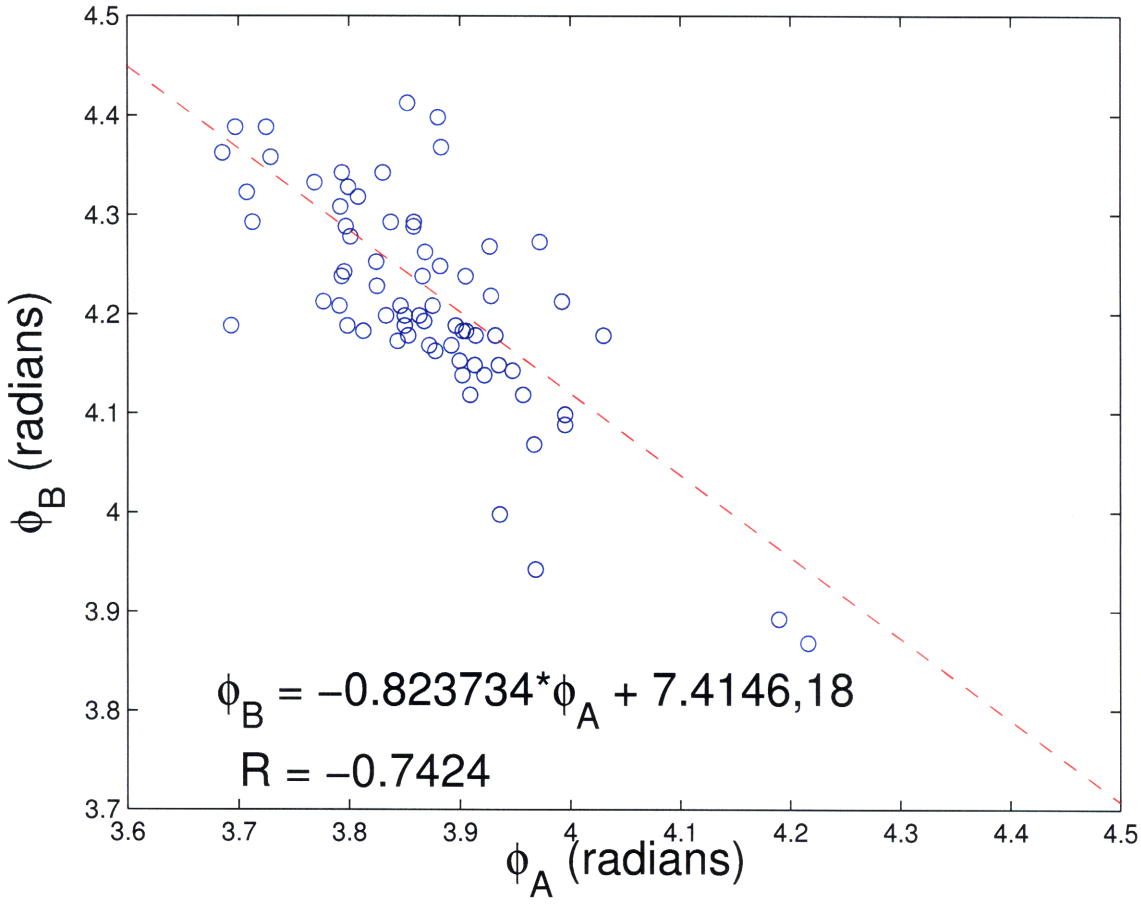
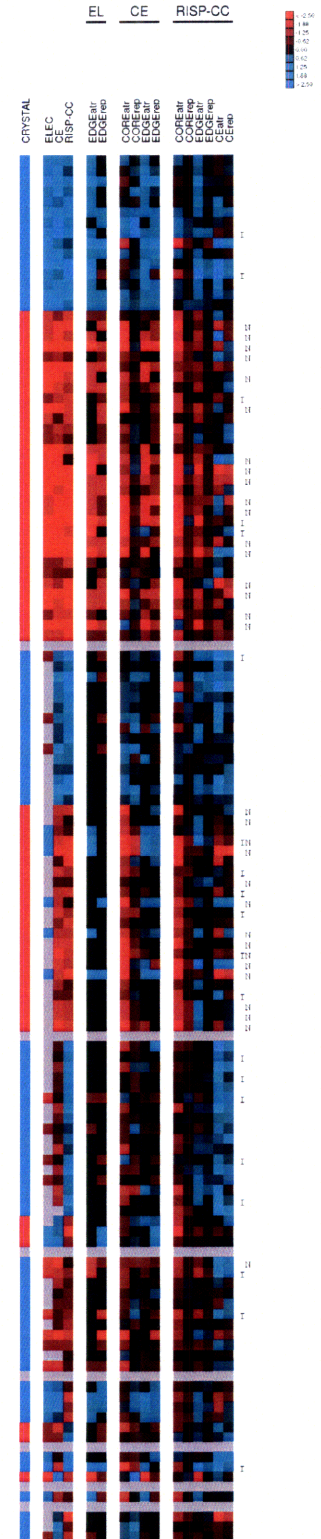


Figure B-4: Antiparallel ϕ_A and ϕ_B correlation for all structures in the test set.

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.

a)


EL
CE
RISP-CC
CRYSTAL
ELEC
CE
RISP-CC
EDGEatr
EDGErep
COREatr
CORErep
EDGEatr
EDGErep
COREatr
CORErep
EDGEatr
EDGErep
CEatr
CErep

b)

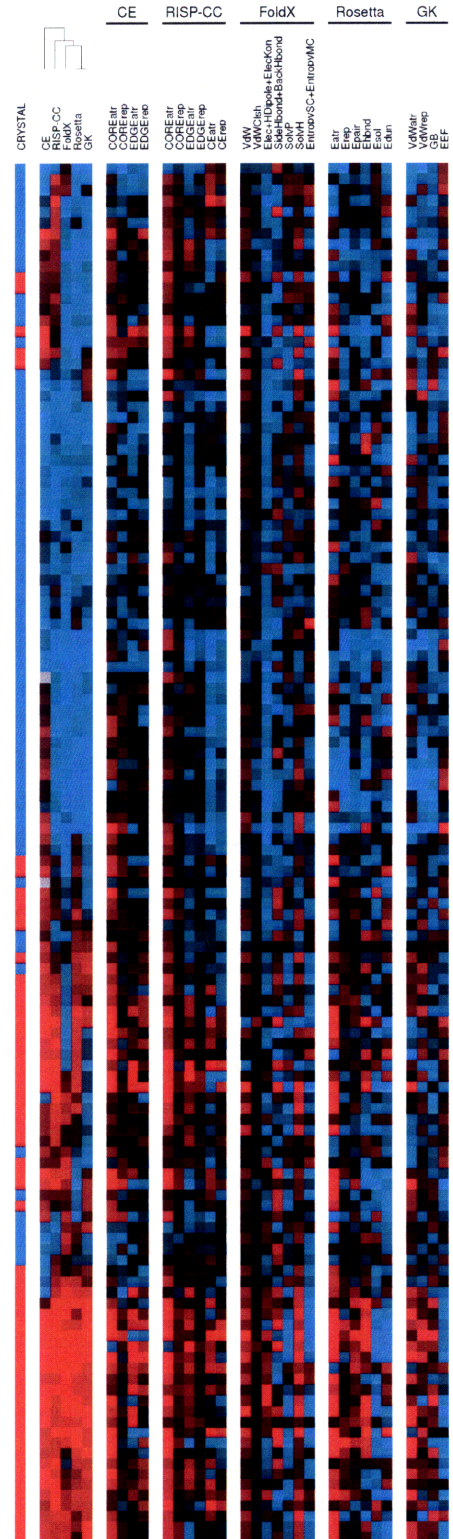

CE
RISP-CC
FoldX
Rosetta
GK
CRYSTAL
CE
RISP-CC
FoldX
Rosetta
GK
COREatr
CORErep
EDGEatr
EDGErep
COREatr
CORErep
EDGEatr
EDGErep
CEatr
CErep
VdW
VdWClash
Elec+HDipole+ElecKon
Side+Bond+BackHbond
SolvAC
SolvH
EntropySC+EntropyMC
Eatr
Erep
Epair
Ehbond
Exal
Edun
VdWatr
VdWrep
GB
EEF

Figure B-5. Component analysis of ESMs and ISMs. Methods and components are as in Figure 3-3. (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 3-3.

References:

1. Newman JR, Keating AE (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science 300: 2097-2101.
2. Selzer T, Albeck S, Schreiber G (2000) Rational design of faster associating and tighter binding protein complexes. Nat Struct Biol 7: 537-541.
3. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320: 369-387.
4. Pey AL, Stricher F, Serrano L, Martinez A (2007) Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. Am J Hum Genet 81: 1006-1024.
5. Grigoryan G, Keating AE (2006) Structure-based prediction of bZIP partnering specificity. J Mol Biol 355: 1125-1142.

Appendix C

Alignment prediction test sets

Test set construction described in section 4.2.2. Accession code and number of total alignments per sequence are provided. Alignment test boundaries are all sequence from N- to -C on both strands. Bold regions are known extent with heptad assigned. The > character denotes continuation of the same sequence on the following line.

Crystal Antiparallel

```
:: 1a36.0 -- 17 possible alignments
                  abcdefgabcdefgabcdefgabcd                                      >
               N-MMNLQTKIDAKKEQLADARRDLKSAKADAKVMKDAKTKKVVESKKKAVQRLEEQLMKLEVQATDREENK >
  C-NLKSTGLAIQKNEERDTAQVELKMLQEELRQVAKKKSEVVKKTKADKMVKADAKASKLDRRADALQEKKADIKTQLNMMSK-N >
                  dcbagfedcbagfedcbagfedcba                                      >

>
> QIALG-C
>
>


:: 1a38.0 -- 9 possible alignments
                                                 defgabcdefgabcdefga
  N-LVQKAKLAEQAERYDDMAACMKSVTEQGAELSNEERNLLSVAYKNVVGARRSSWRVVSSIE-C
                                     C-FKELLSLVDNCIDRLETEIKERYERA-N
                                              agfedcbagfedcbagfed
```

```
:: 1alu.0 -- 5 possible alignments
             abcdefgabcdefgabcdefga
  N-LTSSERIDKQIRYILDGISALRKETCNKS-C
         C-LARLSSQLFEKFSRLILHTTMDQLWQNQAQLKTLLS-N
             dcbagfedcbagfedcbagfed


:: 1bjt.0 -- 9 possible alignments
                   abcdefgabcdefgabcd
  N-YVRLEYYQKRKDHMSERLQWEVEKYSFQVKFIKMIIEKE-C
  C-KLDTNWIDKASLKLLNELETEKEQKQKLLKQYREKTLSWIRMGLLY-N
                   dcbagfedcbagfedcba


:: 1bmf.0 -- 5 possible alignments
             abcdefgabcdefgabcdefgabcdefga
           N-LKDITRRLKSIKNIQKITKSMKMVAAAKYARAEREL-C
  C-KTIVAQRTRNFTLTLKDIMESANKSANDMATMRASQESTTSEK-N
             dcbagfedcbagfedcbagfedcbagfed


:: 1cii.0 -- 35 possible alignments
                                                                                >
                                                                                >
  C-NKQQTLQQEARTLANRATNLSETVSSLKNRAESIARDNRTQQGLLRAQQADLVALAAGGNKELESLRTDLIGKRREYEALMDA >
                                                                                >

>                                          defgabcdefgabcdefgabcdefgabcd        >
>                       N-IEKRDKEITAYKNTLSAQQKENENKRTEAGKRLSAAIAAREKDENTLKTLRAGNADAADITRQ >
> AKKDANQADKQRKEAQEILSRAERPSIRLSDADAFLMRSETHLRLADYGAIETRFGYERLEAQLLRFEQRTIDAADANGARLTKL >
>                                          agfedcbagfedcbagfedcbagfedcba        >

>
> EFRLLQAE-C
> TNEDKERAAIAASLRKGAETRKNENEKQQASLTNKYATIEKDRKEIDSRT-N
>


:: 1cii.2 -- 26 possible alignments
                                                                                >
                                                  N-RSLIEQAEKRQKDAQNADKKAADMLAEYERRKGILDTRLSE >
  C-SSIDSLKVSELAAAIAARDKANIKKNIDARYKEYTKLAEEVNRIKKGKAQGAMERALQEAKAGYKESVSKLFETTFNIADKTA >
                                                                                >

>          abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcd                 >
> LEKNGGAALAVLDAQQARLLGQQTRNDRAISEARNKLSSVTESLNTARNALTRAEQQLTQQ-C               >
> KLEDQKRKEEAIKQNIGSLQNRINEKEKLLANLADNAHKQENTRASLNNRASNVASEASTIKNRADLLRQRLKDWEAVEANYIKK >
>          dcbagfedcbagfedcbagfedcbagfedcbagfedcbagfedcbagfedcba                 >

>
>
> DGEV-N
>


:: 1cnt_2.1 -- 5 possible alignments
             abcdefgabcdefgabcd
  N-RDLCSRSIWLARKIRSDLTALTESY-C
  C-QHSSIFRLDHISRVTWQSLEQLVKLGWLKKEF-N
             dcbagfedcbagfedcba


:: 1cxz.0 -- 14 possible alignments
             abcdefgabcdefgabcdefga
  N-QQQLELERERLRREIRKELKLKEGAENLRRATTDLGRSLGPVELLLRGSSRRLDLLHQQLQELH-C
         C-LEQLQQHLLDLRRSSGRLLLEVPGLSRGLDTTARRLNEAGEKLKLEKRIERRLRERELELQQQV-N
             dcbagfedcbagfedcbagfed
```

```
:: 1ecm.0 -- 1 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcd
  N-LLALREKISALDEKLLALLAERRELAVEVGKA-C
  C-AKGVEVALERREALLALLKEDLASIKERLALL-N
    dcbagfedcbagfedcbagfedcbagfedcba


:: 1ecr.0 -- 4 possible alignments
          abcdefgabcdefgabcd
        N-LNTTFRQMEQELAIFAAHLEQHKLL-C
  C-LESEVTVIHEFTTKLKNIHQIHSVL-N
          dcbagfedcbagfedcba


:: 1egw_1.0 -- 3 possible alignments
                abcdefgabcdefgabcd
  N-QITRIMDERNRQVTFTKRKFGLMKKAYELSVL-C
              C-LVSLEYAKKMLGFKRKTFTVQRNREDMIRTIQ-N
                dcbagfedcbagfedcba


:: 1exj.0 -- 2 possible alignments
          defgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-KKAQDLEMEELFAFYTEQERQIREKLDFLSALEQTISLVKKRMKRQM-C
        C-MQRKMRKKVLSITQELASLFDLKERIQREQETYFAFLEEMELDQAKK-N
          agfedcbagfedcbagfedcbagfedcbagfedcbagfed


:: 1few.0 -- 37 possible alignments
                                                                                abcdef >
                                                                  N-LSSEALMRRAVSLVTDSTSTFLSQTTY >
  C-RLYAEQESEAREEGEEQTKQKLEEIQAEALKTEAKRSLQHVEEVQLKVLQIHNRATISAQDAGTQYAAEAAMESLGVATMWTT >
                                                                                dcbagf >

> gabcdefgabcdefgabcdefgabcd                                                        >
> ALIEAITEYTKAVYTLTSLYRQYTSLLGKMNSEEEDEVWQVIIGARAEMTSKHQEYLKLETTWMTAVGLSEMAAEAAYQTGADQA >
> ELKLYEQHKSTMEARAGIIVQWVEDEEESNMKGLLSTYQRYLSTLTYVAKTYET-N                         >
> edcbagfedcbagfedcbagfedcba                                                        >

>
> SITARNHIQLVKLQVEEVHQLSRKAETKLAEAQIEELKQKTQEEGEERAESEQEAYLRED-C
>
>


:: 1few.2 -- 18 possible alignments
                                                                      abcdefgabcdefgabcd   >
          N-ITEYTKAVYTLTSLYRQYTSLLGKMNSEEEDEVWQVIIGARAEMTSKHQEYLKLETTWMTAVGLSEMAAEAAYQTG >
  C-LYAEQESEAREEGEEQTKQKLEEIQAEALKTEAKRSLQHVEEVQLKVLQIHNRATISAQDAGTQYAAEAAMESLGVATMWTTE >
                                                                      dcbagfedcbagfedcba   >

>
> ADQASITARNHIQLVKLQVEEVHQLSRKAETKLAEAQIEELKQKTQEEGEERAESEQEAYL-C
> LKLYEQHKSTMEARAGIIVQWVEDEEESNMKGLLSTYQRYLSTLTYVAKTYETI-N
>


:: 1fs0.0 -- 3 possible alignments
          abcdefgabcdefgabcdefgabcd
        N-MRKSQDRMAASRPYAETMRKVIGHL-C
  C-MAVMRAAQESALNEVVGQYVQSEVYRRLLTDL-N
          dcbagfedcbagfedcbagfedcba
```

```
:: 1fxk.0 -- 24 possible alignments
                    abcdefgabcdefgabcdefgabcd                                                  >
  N-VQHQLAQFQQLQQQAQAISVQKQTVEMQINETQKALEELSRAADDAEVYKSSGNILIRVAKDELTEELQEKLETLQLREKTIE >
  C-AEQINVQMEQLKKMVREEQREITKERLQLTELKEQLEETLEDKAVRILINGSSKYVEADDAARSLEELAKQTENIQMEVTQKQ >
                    dcbagfedcbagfedcbagfedcba                                                  >

>
> RQEERVMKKLQEMQVNIQE-C
> VSIAQAQQQLQQFQALQHQ-N
>


:: 1fxk.1 -- 5 possible alignments
     abcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-LAEIVAQLNIYQSQVELIQQQMEAVRATISELEILEKTLSDIQGKD-C
  C-LEEAQPSLKMMIDTIARLNEGMKQLTSELENKQSKISEMADEFNKKIAVGAGV-N
     dcbagfedcbagfedcbagfedcbagfedcbagfedcba


:: 1gmj_2.0 -- 12 possible alignments
                                        abcdefgabcdefgabcd
  N-REQAEEERYFRARAKEQLAALKKHKENEISHHAKEIERLQKEIERHKQSIKKL-C
                               C-LKKISQKHREIEKQLREIEKAHHSIENEKHKKLAALQEKARARFYREEEAQER-N
                                        dcbagfedcbagfedcba


:: 1h8e.0 -- 3 possible alignments
           abcdefgabcdefgabcdefgabcdefgabcdefga
         N-LKDITRRLKSIKNIQKITKSMKMVAAAKYARAEREL-C
  C-KTIVAQRTRNFTLTLKDIMESANKSANDMATMRASQESTTSEK-N
           dcbagfedcbagfedcbagfedcbagfedcbagfed


:: 1ili.0 -- 13 possible alignments
                                  abcdefgabcdefgabcd
         N-EKSIKMGKRNGLHLSEHIRNEIKSMKKRMSELCIDFNKNLNEDDTSLVFSKAE-C
  C-DDLFAAVRSTSKATNLELVFDAHTNYGLLKAVQARLPLLQQLIATNEQKCRTHFAMEMKR-N
                                  dcbagfedcbagfedcba


:: 1ilr.0 -- 5 possible alignments
     abcdefgabcdefgabcd
  N-IQRLNWMLWVIDECFRDLCYRTGIC-C
  C-LVRVAQGAFKEMASLVYFSAFHRVWYKLGQLR-N
     dcbagfedcbagfedcba


:: 1i4d.1 -- 6 possible alignments
                    abcdefgabcdefgabcdefga
  N-VSSINTLVTKTMEDTLMTVKQYEAARLEYDAYRTDL-C
  C-KHMVKIKNEELFKLKIAVDGRLKEYKDRHAQFTAQA-N
                    dcbagfedcbagfedcbagfed


:: 1ii8.0 -- 12 possible alignments
                                                      defgabcdefgabcdefgabcdefga
                                  N-AILESDEAREKVVREVLNLDKFETAYKKLSELKKTINNRIKEYRDIL-C
  C-EKGEWVVFLRVKNEEARVVVESYKGETFEAFIESALEGIKSLAAERALAKYKKVKEILEETFDKAKELDKIEKKV-N
                                                      agfedcbagfedcbagfedcbagfed


:: 1io1.0 -- 17 possible alignments
           abcdefgabcdefgabcdefgabcdefgabcd                                                >
  N-IKGLTQASRNANDGISIAQTTEGALNEINNNLQRVRELAVQSANSTNSQSDLDSIQAEITQRLNEIDRVSGQTQFNGVKVLAQ >
  C-LNNVTNGLNTIASNFRNQVAALDSRLTDVQALAADIKQL-N                                             >
           dcbagfedcbagfedcbagfedcbagfedcba                                                >

>
> DNTLTIQVGANDGETIDIDLKQINSQTLGLDTLNVQQKYKVSDTAAT-C
>
>


:: 1ivs_2.0 -- 4 possible alignments
           defgabcdefgabcdefga
  N-VEEWRRRQEKRLKELLALAERSQRKL-C
         C-LAERIREAQELNEKLRAEEAEVVEKP-N
           agfedcbagfedcbagfed
```

```
:: 1jbg.0 -- 2 possible alignments
            abcdefgabcdefgabcd
  N-KAALQSQKEILMKKKQRMDEMIQTI-C
         C-ITQIMEDMRQKKKMLIEKQSQLAAK-N
            dcbagfedcbagfedcba


:: 1k1f_2.0 -- 6 possible alignments
                    abcdefgabcdefgabcdefgabcd
  N-VGDIEQELERAKASIRRLEQEVNQERFRMIYLQTLLAKE-C
  C-EKALLTQLYIMRFREQNVEQELRRISAKARELEQEIDGV-N
                    dcbagfedcbagfedcbagfedcba


:: 1kvk.0 -- 7 possible alignments
                          defgabcdefgabcdefga
  N-TKALVAGVRSRLIKFPEIMAPLLTSIDAISLECERVLGEM-C
         C-TVQCLQDLSAHGVGLANLHHQNMDMLEELVLYQ-N
                          agfedcbagfedcbagfed


:: 1l8d.0 -- 3 possible alignments
     defgabcdefgabcdefgabcdefgabcdefgabcd
  N-LETKKTTIEEERNEITQRIGELKNKIGDLKTAIEELKKAKGKC-C
  C-IEMDIRRLERELESKRDILKALTNKSNNLDLHYKSL-N
     agfedcbagfedcbagfedcbagfedcbagfedcba


:: 1m5i.1 -- 11 possible alignments
            abcdefgabcdefgabcd
          N-LEELEKERSLLLADLDKEEKEKDWYYAQLQNLTKRIDSL-C
  C-QSQLLQRIRLIDKEIQQIRAIRRQARKEMDQCTGLQEEMAVRIQRAEYELQRRTMDTQLS-N
            dcbagfedcbagfedcba


:: 1nt2.0 -- 19 possible alignments
                    defgabcdefgabcdefga                                                    >
  N-LAVSEKMVEKELRREDRYVVALVKALEEIDESINMLNEKLEDIRAVKESEITEKFEKKIRELRELRRDVEREIEE-C        >
                 C-IKEMVEEIEREVDRRLERLERIKKEFKETIESEKVARIDELKENLMNISEDIEELAKVLAVVYRDERRL >
                    agfedcbagfedcbagfed                                                    >

>
>
> EKEVMKESVALAT-N
>


:: 1o9c.0 -- 26 possible alignments
                                                      abcdefgabcdefgabcd                 >
  N-ENVYMAKLAEQAERYEEMVEFMEKVSNSLGSEELTVEERNLLSVAYKNVIGARRASWRIISSIEQKEESRGNEEHVNSIREYR >
                                                 C-LKLIGDCIKSLENEIKSRYERISNVHEENGRSEEKQEISSI >
                                                      dcbagfedcbagfedcba                 >

>
> SKIENELSKICDGILKLLD-C
> IRWSARRAGIVNKYAVSLLNREEVTLEESGLSNSVKEMFEVMEEYREAQEALKAMYVNEER-N
>
```

```
:: 1orj.1 -- 17 possible alignments
    defgabcdefgabcdefga                                                                                  >
  N-LEQIILLYDKAIECLERAIEIYDQVNELEKRKEFVENIDRVYDIISALKSFLDHEKGKEIAKNLDTIYTIILNTLVKVDKTKE >
  C-KVEEWAERLDKLIELIKQLEEKTKDVKVLTNLIITYITDL-N                                                         >
    agfedcbagfedcbagfed                                                                                  >

>
> ELQKILEILKDLREAWEEVK-C
>
>


:: 1ov9.0 -- 3 possible alignments
          abcdefgabcdefgabcd
  N-AYARELTIEQLEEALDKLTTVVQERKEAEAEE-C
  C-EEAEAEKREQVVTTLKDLAEELQEITLERAYA-N
          dcbagfedcbagfedcba


:: 1q05.0 -- 3 possible alignments
                  abcdefgabcdefgabcd
  N-QRHSADVKRRTLEKVAEIERHIEELQSMRDQLLALANAC-C
        C-CANALALLQDRMSQLEEIHREIEAVKELTRRK-N
                  dcbagfedcbagfedcba


:: 1q06.0 -- 2 possible alignments
          defgabcdefgabcdefgabcdefga
  N-SADVKRRTLEKVAEIERHIEELQSMRDQLLALA-C
        C-ALALLQDRMSQLEEIHREIEAVKELTRRKVDAS-N
          agfedcbagfedcbagfedcbagfed


:: 1q08.0 -- 3 possible alignments
          defgabcdefgabcdefgabcdefga
  N-CQESKGIVQERLQEVEARIAELQSMQRSLQRLNDACCGTA-C
  C-ATGCCADNLRQLSRQMSQLEAIRAEVEQLREQVIGKSEQC-N
          agfedcbagfedcbagfedcbagfed


:: 1qsd.1 -- 8 possible alignments
                              abcdefgabcdefgabcdefgabcd
                      N-LDIKVKALKRLTKEEGYYQQELKDQEAHVAKLKEDKSVD-C
  C-QASTIASRADSVDETGQYTKLFQELDEKFERIKEYLTPLLRKTDDLVEEQKKL-N
                              dcbagfedcbagfedcbagfedcba


:: 1qvr_1.0 -- 38 possible alignments
                                                                                    defgabcdefgab >
                                                                N-IDEAAARLRMALESAPEEIDALERKKL >
  C-ETVELRVFRAGRLKESLAEVEAELKPLEGYRLEAARNLDYQREALEIERRVEDLRHQAERLKRLIERERWEARLKAIEETLK >
                                                                                    agfedcbagfedc >

> cdefga                                                                                             >
> QLEIEREALKKEKDPDSQERLKAIEAEIAKLTEEIAKLRAEWERELRKLREAQHRLDEVRREIELAERQYDLNRAAELRYGE >
> AIEAEIAKLREQSDPDKEKKLAEREIELQLKKRELADIEEPASELAMRLRAAAEDILDIAKD-N                                  >
> bagfed                                                                                             >

>
> LPKLEAEVEALSEKLRGARFVRLEVTEEDIAEI-C
>
>


:: 1qz2.2 -- 3 possible alignments
                  defgabcdefgabcdefga
  N-KTQLAVCQQRIRRQLAREKKLYANMFERLAEEE-C
              C-EEEALREFMNAYLKKERALQRRIRQQCVALQTK-N
                agfedcbagfedcbagfed


:: 1r6f.0 -- 9 possible alignments
                        defgabcdefgabcdefga
  N-DDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSSSGTIN-C
        C-MVSDYKQIFRNLAEIASNFRSTIDSLQTTKQSV-N
                        agfedcbagfedcbagfed


:: 1r6t.0 -- 9 possible alignments
          abcdefgabcdefgabcd
```

```
        N-LLELFNSIATQGELVRSLKAGNASKDEIDSAVKMLVSLKMSYKAAA-C
  C-AKYDEGAAAKYSMKLSVLMKVASDIEDKSANGAKLSRVL-N
          dcbagfedcbagfedcba


:: 1r7j.0 -- 4 possible alignments
                       abcdefgabcdefgabcdefgabcd
  N-MDLEIIRQEGKQYMLTKKGEELLEDIRKFNEMRKNMDQLKEKINSV-C
                     C-VSNIKEKLQDMNKRMENFKRIDELLEEGKKTLMYQKGEQRIIELDM-N
                       dcbagfedcbagfedcbagfedcba


:: 1rq0_2.0 -- 14 possible alignments
                     abcdefgabcdefgabcd
  N-QMKNYGMEYAKIEEIENITNRIKETQEFIELLREEGENELEIEKYEKELDQLYQELLFLL-C
                C-LFLLEQYLQDLEKEYKEIELENEGEERLLEIFEQTEKIRNTINEIEEIKAYEMGYNKMQE-N
                  dcbagfedcbagfedcba


:: 1s4b.0 -- 12 possible alignments
                                                                  abcdefgabcdefgabcd          >
                                                               N-IKRIKKKLSLLCIDFNKN-C         >
  C-GRRECEARKLELIVAREQEGLPKLKQALEDLFTAVTQSTKAMNMELVYDAHTHFGLLRSKQARLTVLEKLIASNEEKCRSNFA >
                                                                  dcbagfedcbagfedcba          >

>
>
> EEVKR-N
>


:: 1ses.1 -- 8 possible alignments
                     abcdefgabcdefgabcd
  N-RAIREKGVALDLEALLALDREVQELKKRLQEVQTERNQVAKRVPKA-C
                C-LAELRAEKERLAEELRKAEEGLAKGRAILAEK-N
                  dcbagfedcbagfedcba


:: 1t3j.0 -- 4 possible alignments
                     abcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-ATTFARLCQQVDMTQKHLEEEIARLSKEIDQLEKMQNNSKLLRNKAVQLESELENFSKQF-C
          C-FQKSFNELESELQVAKNRLLKSNNQMKELQDIEKSLRAIEEELHKQTMDVQQCLRAFTTA-N
                     dcbagfedcbagfedcbagfedcbagfedcbagfedcba


:: 1tjl_1.0 -- 4 possible alignments
               abcdefgabcdefgabcdefgabcd
  N-EAQLAHFRRILEAWRNQLRDEVDRTVTHMQDE-C
          C-VKKLTKEIKKILKRERDRNRLELSFEEEQAAR-N
            dcbagfedcbagfedcbagfedcba


:: 1twf.1 -- 3 possible alignments
               abcdefgabcdefgabcd
          N-VRGIDTLQKKVASILLAL-C
  C-QLNWETEFNTKLAGLKNIISNCANK-N
               dcbagfedcbagfedcba
```

```
:: 1vp7_1.0 -- 6 possible alignments
                                        defgabcdefgabcdefga                                  >
  N-FETALAELESLVSAMENGTLPLEQSLSAYRRGVELARVCQDRLAQAEQQVKVLE-C                                 >
                                      C-ELVKVQQEAQALRDQCVRALEVGRRYASLSQELPLTGNEMASVLSELE >
                                        agfedcbagfedcbagfed                                  >

>
>
> ALATEF-N
>


:: 1wle.0 -- 13 possible alignments
    defgabcdefgabcdefga
  N-TWQELRQLREQIRSLEEEKEAVTEAVRALVVNQDNSQVQQ-C
  C-FQEELQAEKPYLLTLQKRIERGRARLSQYQPDQQVQSNDQNVVLARVAETVAEKEEELSRIQERLQRLEQWTSII-N
    agfedcbagfedcbagfed


:: 1x03.0 -- 5 possible alignments
                                    defgabcdefgabcdefgabcdefgabcdefga                        >
  N-RQALEKFDESKEIAESSMFNLLEMDIEQVSQLSALVQAQLEYHKQAVQILQQVTVRLEERI-C                         >
                                C-IREELRVTVQQLIQVAQKHYELQAQVLASLQSVQEIDMELLNFMSSEAIEKSEDF >
                                  agfedcbagfedcbagfedcbagfedcbagfed                          >

>
>
> KELAQR-N
>


:: 1x75.1 -- 7 possible alignments
    defgabcdefgabcdefga
  N-TIFELRKARDRAHILEALAVALANIDPIIELIR-C
  C-LIRLLEAIQDLLEKYEDLLKEHELGTLKQLRLDLIAQAQQ-N
    agfedcbagfedcbagfed


:: 1xd4_2.0 -- 3 possible alignments
    defgabcdefgabcdefga
  N-VKAFMAEIRQYIRELNLII-C
  C-VTDEIHGLLKVSLEHIDVIRSFINEV-N
    agfedcbagfedcbagfed


:: 1xnp.0 -- 10 possible alignments
                defgabcdefgabcdefgabcdefga
         N-INVKMRELAEFLHELNERIREIIEEKRELEEARILIETYIENTMRRLAEENRQI-C
  C-IQRNEEALRRMTNEIYTEILIRAEELERKEEIIERIRENLEHLFEALERMKVNI-N
                agfedcbagfedcbagfedcbagfed


:: 1ybz.0 -- 1 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcd
  N-LKLLRKEIDKIDNQIISLLKKRLEIAQAIGKI-C
  C-IKGIAQAIELRKKLLSIIQNDIKDIEKRLLKL-N
    dcbagfedcbagfedcbagfedcbagfedcba


:: 1ydx.0 -- 7 possible alignments
                defgabcdefgabcdefgabcd
  N-NKNEQHAIANTLSVFDERLENLASLIEINRKLRDEYAHKLFSL-C
         C-LKKLLTDRIVTLSSLEKKYQDLKQDLLFVIKGAKRQ-N
                agfedcbagfedcbagfedcba


:: 1yf2.0 -- 6 possible alignments
          abcdefgabcdefgabcdefgabcd
  N-QKQIAKILTKIDEGIEIIEKSINKLERIKKGLMHKLLTK-C
  C-GTLLLEMIKKKMRQLKEKKQKKLEISKDVSSLIKAIQKQ-N
          dcbagfedcbagfedcbagfedcba


:: 1yf2.1 -- 8 possible alignments
          abcdefgabcdefgabcd
  N-QKQIAKILTKIDEGIEIIEKSINKLERIKKGLMHKLLTK-C
  C-GTLLLEMIKKKMRQLKEKKQKKLEISKDVSSLIKAIQKQ-N
          dcbagfedcbagfedcba
```

```
:: 1z0j.0 -- 8 possible alignments
    abcdefgabcdefgabcd
  N-EELLLQQIDNIKAYIFDAKQCGRLDEVEVLTENLRELKHTLAKQKG-C
  C-QKALTHKLERLNETLVEVEDLRGCQKADFIYA-N
    dcbagfedcbagfedcba


:: 1z0p.0 -- 4 possible alignments
          defgabcdefgabcdefgabcd
  N-MSYEKEFLKDFEDWVKTQIQVNQLAMATS-C
         C-KYNDFKGLLFEYADLKSEYRIFADKARED-N
          agfedcbagfedcbagfedcba


:: 1zke_1.0 -- 12 possible alignments
          abcdefgabcdefgabcd
        N-IRKILADIEDSQNEIEMLLKLANLSLGDFIEIKRGSMDMPKGVNEAFFTQLSEEVERLKELINALNKIKKGLLV-C
  C-FVLLGKKIKNLANILEKLREVEESLQTFFAEN-N
          dcbagfedcbagfedcba


:: 1zpy.0 -- 13 possible alignments
          defgabcdefgabcdefga
  N-ETRDMHRAIISLREELEAVDLYNQRVNACKDKELKAILAHNRDEEKEHAAMLLE-C
         C-IWELLMAAHEKEEDRNHALIAKLEKDKCANVRQNYLDVAELEERLSIIARHMDRTEDSLEQ-N
          agfedcbagfedcbagfed


:: 2avr.0 -- 26 possible alignments
                                  defgabcdefgabcdefgabcd                          >
  N-AASLVGELQALDAEYQNLANQEEARFNEERAQADAARQALAQNEQVYNELSQRAQRLQAEANTRFYKSQYQELASKYEDALKK >
                      C-LNGARLAQIKEFDSIVAKQQEMEAELKKLADEYKSALEQYQSKYFRTNAEAQLRQARQSLEN >
                                  agfedcbagfedcbagfedcba                          >

>
> LEAEMEQQKAVISDFEKIQALRA-C
> YVQENQALAQRAADAQAREENFRAEEQNALNQYEADLAQLEGVL-N
>


:: 2b5u.0 -- 7 possible alignments
                 defgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-AERNYERARAELNQANEDVARNQERQAKAVQVYNSRKSELDAANKTLADAIAEIKQFNRFA-C
  C-LNNEASRKKDEKKKRSEMASSLAADADSKEKAAADFAAQKNNVDTQARQAKLGA-N
                 agfedcbagfedcbagfedcbagfedcbagfedcbagfed


:: 2bde.0 -- 13 possible alignments
    defgabcdefgabcdefga
  N-IEKKIGEAMAIKKELEQKYVDLCTRSIDESSQQYDQEIHDLQLQISTVDLQISRLLQEQNS-C
  C-QEQLLRSIQLDVTSIQLQLDHIEQDYQQSSEDISRTCLDVYKQELEKKIAMAEG-N
    agfedcbagfedcbagfed
```

```
:: 2br9.0 -- 15 possible alignments
                                                    abcdefgabcdefgabcd                  >
  N-REDLVYQAKLAEQAERYDEMVESMKKVAGMDVELTVEERNLLSVAYKNVIGARRASWRIISSIEQKEENKGGEDKLKMIREYR >
                                              C-VDLIDCCILKLETEVMQRYERIMKLKDEGGKNEEKQEIS-N >
                                                    dcbagfedcbagfedcba                  >

>
> QMVET-C
>
>


:: 2btp.0 -- 11 possible alignments
                                                              defgabcdefgabcdefga
  N-GTENLYFQSMEKTELIQKAKLAEQAERYDDMATCMKAVTEQGAELSNEERNLLSVAYKNVVGGRRSAWRVISSIE-C
                                                          C-YKDLLELVTTCISRLESEVKERYDKI-N
                                                              agfedcbagfedcbagfed


:: 2d4c_1.1 -- 12 possible alignments
                                  defgabcdefgabcdefgabcdefga                        >
  N-RQALEKFDESKEIAESSMFNLLEMDIEQVSQLSALVQAQLEYHKQAVQILQQVTVRLEERI-C                >
                                C-IREELRVTVQQLIQVAQKHYELQAQVLASLQSVQEIDMELLNFMSSEAIEKSEDF >
                                  agfedcbagfedcbagfedcbagfed                        >

>
>
> KELAQR-N
>


:: 2d4x.0 -- 12 possible alignments
                    abcdefgabcdefgabcdefgabcd
  N-QAQNSQYALARTFATQKVSLEESVLSQVTTAIQTAQEKIVYAGNGTLSDDDRASLATDLQGIRDQLMNLANSTDGNGRYIF-C
                 C-LESLQTGLEARVTLVNNLSNKLGRNTKDIAAAAKEKEVN-N
                    dcbagfedcbagfedcbagfedcba


:: 2d4y_1.0 -- 10 possible alignments
    abcdefgabcdefgabcdefga
  N-LRGAQNQSSGLTTRYEQMSKIDNLLADKSSSLSGSLQSFFTSLQTLVSNA-C
  C-LQKVVNAQTTSSTKLTSTKNGVDSVLTAYADNFTKNGGVVNSNQLDLLAQ-N
    dcbagfedcbagfedcbagfed


:: 2d8e.0 -- 1 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcd
  N-IQALRKEVDRVNREILRLLSERGRLVQEIGRL-C
  C-LRGIEQVLRGRESLLRLIERNVRDVEKRLAQI-N
    dcbagfedcbagfedcbagfedcbagfedcba


:: 2esh.0 -- 1 possible alignments
    defgabcdefgabcdefgabcdefga
  N-LREILRSLEDMKRRIETLEERIKRVL-C
  C-LVRKIREELTEIRRKMDELSRLIERL-N
    agfedcbagfedcbagfedcbagfed


:: 2etn_3.0 -- 18 possible alignments
                    abcdefgabcdefgabcd
  N-VKLTKAGYERLMKQLEQERERLQEATKILQELMESSDDYDDSGLEAAKQEKARIEARIDSLEDVLSRAVILEEG-C
  C-TGEELIVARSLVDELSDIRAEIRAKEQKAAELGSDDYDDSSEMLEQLIKTAEQLREREQELQKMLREYGAKTLK-N
                    dcbagfedcbagfedcba
```

```
:: 2fup.0 -- 15 possible alignments
                                                          abcdefgabcdefgabcdefga              >
                                                       N-LLDLFAEDIGHANQLLQLVDEEFQALERR-C      >
  C-RGNRLNAQQCRELLEGLEDGRALLEAGDARERAYRALGERDLSVGAERLIEARARGNRELQQMLPQKAGLLQQLVPLERRELA >
                                                          dcbagfedcbagfedcbagfed              >

>
>
> QFEEDVLQLLQNAHGIDEAFLDL-N
>


:: 2h7v_2.0 -- 18 possible alignments
                                          defgabcdefgabcdefga                                 >
                          N-KHLDQTHSFSDIGSLVRAHKHLETLLEVLVTLSQQGQPVSSETYGFLNRLAEAKITLSQQLN >
  C-FRQLSQRAVDAWSGSRNILISLQAKASEQQQQLTNLQQSLTIKAEALRNLFGYT-N                               >
                                          agfedcbagfedcbagfed                                 >

>
> TLQQQQESAKAQLSILINRSGSWADVARQSLQRF-C
>
>


:: 2h94.0 -- 7 possible alignments
                          abcdefgabcdefgabcdefga
  N-EKHVKDEQIEHWKKIVKTQEELKELLNKMVNLKEKIKELHQQY-C
                  C-LEQLKEELKGQTEALEDYEKCLATLDRHKSKVLFEA-N
                          dcbagfedcbagfedcbagfed


:: 2h1d_1.0 -- 8 possible alignments
              abcdefgabcdefgabcdefgabcdefgabcd
          N-LKEVEMRLKSIKNIEKITKTMKIVASTRLSKAEKAKISAKKMDEAEQLFYKNA-C
  C-NTIVAQRTRNYLISYRNIMDGANKSANDMANRRASIEAAYGQAMATLMQNALT-N
              dcbagfedcbagfedcbagfedcbagfedcba


:: 2ic6_1.0 -- 4 possible alignments
    abcdefgabcdefgabcdefga
  N-LKEVQDNITLHEQRLVTTRQKLKDAERAV-C
  C-LDALERKLEGLKTELASVAARRSQLTSKN-N
    dcbagfedcbagfedcbagfed


:: 2jdi.0 -- 8 possible alignments
              abcdefgabcdefgabcd
          N-LKDITRRLKSIKNIQKITKSMKMVAAAKYARA-C
  C-KTIVAQRTRNFTLTLKDIMESANKSANDMATMRASQESTTSEKLSY-N
              dcbagfedcbagfedcba


:: 2nov_1.0 -- 13 possible alignments
                          defgabcdefgabcdefga
          N-SSYIAHRREVILARSRFDKEKAEKRLHIVEGLIRVISILDEVIALIRASENKADAKENLKV-C
  C-FKKKVERLEKKMLNYMTREDGIIAALMAIKERLEAEEEQLVVVDTNTLRYLQLT-N
                          agfedcbagfedcbagfed


:: 2o98.1 -- 5 possible alignments
                                  abcdefgabcdefgabcd
                              N-FNELNQLAEEAKRRAEIARQRELHTLKGHVESVVKLKGLDIETIQQ-C
  C-QQITEIDLGKLKVVSEVHGKLTHLERQRAIEARRKAEEALQNLENF-N
                                  dcbagfedcbagfedcba


:: 2pah.1 -- 2 possible alignments
              abcdefgabcdefgabcd
  N-VLDNTQQLKILADSINSEIGILCSA-C
          C-ASCLIGIESNISDALIKLQQTNDLV-N
              dcbagfedcbagfedcba
```

Crystal Parallel

```
:: 1am9_1.2 -- 6 possible alignments
                                      defgabcdefgabcdefgabcdefga
  N-IEKRYRSSINDKIIELKDLVVGTEAKLNKSAVLRKAIDYIRFLQHSNQKLKQENLSLRTAVHKSKSLK-C
  N-IEKRYRSSINDKIIELKDLVVGTEAKLNKSAVLRKAIDYIRFLQHSNQKLKQENLSLRTAV-C
```

```
                                        defgabcdefgabcdefgabcdefga

:: 1ber.0 -- 8 possible alignments
                                        abcdefgabcdefgabcd
  N-RSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQMARRLQVTSEKVGNLAFLDVTGRIAQTLL-C
  N-RSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQMARRLQVTSEKVGNLAFLDVTGRIAQTLL-C
                                        abcdefgabcdefgabcd


:: 1c1g_2.0 -- 30 possible alignments
                                defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgab >
  N-IKKKMQMLKLDKENALDRADEAEADKKAAEDRSKQLEDELVSLQKKLKATEDELDKYSEALKDAQEKLELAEKKATDAEADVA >
  N-IKKKMQMLKLDKENALDRADEAEADKKAAEDRSKQLEDELVSLQKKLKATEDELDKYSEALKDAQEKLELAEKKATDAEADVA >
                                defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgab >

> cdefgabcdefgabcdefga                                                              >
> SLNRRIQLFEEELDRAQERLATALQKLEEAEKAADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKYEEVARKLVI >
> SLNRRIQLFEEELDRAQERLATALQKLEEAEKAADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKYEEVARKLVI >
> cdefgabcdefgabcdefga                                                              >

>                                                                                   >
> IESDLERAEERAELSEGKCAELEEEIKTVTNNLKSLEAQAEKYSQKEDKYEEEIKVLSDKLKEAETRAEFAERSVTKLEKSIDDL >
> IESDLERAEERAELSEGKCAELEEEIKTVTNNLKSLEAQAEKYSQKEDKYEEEIKVLSDKLKEAETRAEFAERSVTKLEKSIDDL >
>                                                                                   >

>
> EDELYAQKLKYKAISEELDHALNDM-C
> EDELYAQKLKYKAISEELDHALNDM-C
>


:: 1c1g_2.2 -- 32 possible alignments
                                                                                    >
  N-IKKKMQMLKLDKENALDRADEAEADKKAAEDRSKQLEDELVSLQKKLKATEDELDKYSEALKDAQEKLELAEKKATDAEADVA >
  N-IKKKMQMLKLDKENALDRADEAEADKKAAEDRSKQLEDELVSLQKKLKATEDELDKYSEALKDAQEKLELAEKKATDAEADVA >
                                                                                    >

>                                                                                   >
> SLNRRIQLFEEELDRAQERLATALQKLEEAEKAADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKYEEVARKLVI >
> SLNRRIQLFEEELDRAQERLATALQKLEEAEKAADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKYEEVARKLVI >
>                                                                                   >

> defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga                      >
> IESDLERAEERAELSEGKCAELEEEIKTVTNNLKSLEAQAEKYSQKEDKYEEEIKVLSDKLKEAETRAEFAERSVTKLEKSIDDL >
> IESDLERAEERAELSEGKCAELEEEIKTVTNNLKSLEAQAEKYSQKEDKYEEEIKVLSDKLKEAETRAEFAERSVTKLEKSIDDL >
> defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga                      >

>
> EDELYAQKLKYKAISEELDHALNDM-C
> EDELYAQKLKYKAISEELDHALNDM-C
>
```

```
:: 1c1g_2.3 -- 36 possible alignments
                                                                                                                >
  N-IKKKMQMLKLDKENALDRADEAEADKKAAEDRSKQLEDELVSLQKKLKATEDELDKYSEALKDAQEKLELAEKKATDAEADVA >
  N-IKKKMQMLKLDKENALDRADEAEADKKAAEDRSKQLEDELVSLQKKLKATEDELDKYSEALKDAQEKLELAEKKATDAEADVA >
                                                                                                                >

>                                                                                                               >
> SLNRRIQLFEEELDRAQERLATALQKLEEAEKAADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKYEEVARKLVI >
> SLNRRIQLFEEELDRAQERLATALQKLEEAEKAADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKYEEVARKLVI >
>                                                                                                               >

>                                                                                          defgabcdefgabcd >
> IESDLERAEERAELSEGKCAELEEEIKTVTNNLKSLEAQAEKYSQKEDKYEEEIKVLSDKLKEAETRAEFAERSVTKLEKSIDDL >
> IESDLERAEERAELSEGKCAELEEEIKTVTNNLKSLEAQAEKYSQKEDKYEEEIKVLSDKLKEAETRAEFAERSVTKLEKSIDDL >
>                                                                                          defgabcdefgabcd >

> efgabcdefgabcdefgabcd
> EDELYAQKLKYKAISEELDHALNDMTSI-C
> EDELYAQKLKYKAISEELDHALNDMTSI-C
> efgabcdefgabcdefgabcd


:: 1cz7_2.0 -- 2 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-VVHLRQRTEELLRCNEQQAAELETCKEQLFQSNMERKELHNTVMDLRGNIRVF-C
  N-VVHLRQRTEELLRCNEQQAAELETCKEQLFQSNMERKELHNTVMDLRGNIRVF-C
    abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcd


:: 1deb.0 -- 3 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcdefga
  N-YDQLLKQVEALKMENSNLRQELEDNSNHLTKLETEASNMKEVLKQLQGSI-C
  N-YDQLLKQVEALKMENSNLRQELEDNSNHLTKLETEASNMKEVLKQLQGSI-C
    abcdefgabcdefgabcdefgabcdefgabcdefga


:: 1dgc.0 -- 5 possible alignments
                        abcdefgabcdefgabcdefgabcd
  N-ALKRARNTEAARRSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGER-C
  N-ALKRARNTEAARRSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGER-C
                        abcdefgabcdefgabcdefgabcd


:: 1fos_1.0 -- 8 possible alignments
                  defgabcdefgabcdefgabcdefgabcdefga
  N-RERNKMAAAKSRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFIL-C
  N-RMRNRIAASKSRKRKLERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKV-C
                  defgabcdefgabcdefgabcdefgabcdefga


:: 1fos_2.0 -- 6 possible alignments
                  defgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-RERNKMAAAKSRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFIL-C
  N-RMRNRIAASKSRKRKLERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKV-C
                  defgabcdefgabcdefgabcdefgabcdefgabcdefga


:: 1ft9.0 -- 6 possible alignments
                              defgabcdefgabcdefgabcd
  N-LVEATERTEVRFADIRTFEQKLQTCPSMAWGLIAILGRALTSCMRTIEDLMFHDIKQ-C
  N-LVEATERTEVRFADIRTFEQKLQTCPSMAWGLIAILGRALTSCMRTIEDLMFHDIKQ-C
                              defgabcdefgabcdefgabcd


:: 1gd2_2.0 -- 5 possible alignments
                        abcdefgabcdefgabcdefgabcdefgabcd
  N-SKRKAQNRAAQRAFRKRKEDHLKALETQVVTLKELHSSTTLENDQLRQKVRQLEEELRIL-C
  N-SKRKAQNRAAQRAFRKRKEDHLKALETQVVTLKELHSSTTLENDQLRQKVRQLEEELRIL-C
                        abcdefgabcdefgabcdefgabcdefgabcd


:: 1go4.0 -- 5 possible alignments
            defgabcdefgabcdefgabcdefgabcdefga
  N-AEQSFLFSREEADTLRLKVEELEGERSRLEEEKRMLEAQLERRALQGDYDQSRTKVLHMSL-C
  N-AEQSFLFSREEADTLRLKVEELEGERSRLEEEKRMLEAQLERRALQGDYDQSRTKVLHMSL-C
            defgabcdefgabcdefgabcdefgabcdefga


:: 1go4.1 -- 5 possible alignments
            defgabcdefgabcdefgabcdefga
```

```
  N-SREEADTLRLKVEELEGERSRLEEEKRMLEAQLERRALQGDYDQSRTKVLHMSL-C
  N-SREEADTLRLKVEELEGERSRLEEEKRMLEAQLERRALQGDYDQSRTKVLHMSL-C
           defgabcdefgabcdefgabcdefga


:: 1h88.0 -- 4 possible alignments
                          defgabcdefgabcdefgabcdefgabcdefga
  N-RERNNIAVRKSRDKAKMRNLETQHKVLELTAENERLQKKVEQLSRELSTLRNLF-C
  N-RERNNIAVRKSRDKAKMRNLETQHKVLELTAENERLQKKVEQLSRELSTLRNLF-C
                          defgabcdefgabcdefgabcdefgabcdefga


:: 1hlo.2 -- 4 possible alignments
                   defgabcdefgabcdefga
  N-LQGEKASRAQILDKATEYIQYMRRKNHTHQQDIDDLKRQN-C
  N-LQGEKASRAQILDKATEYIQYMRRKNHTHQQDIDDLKRQN-C
                   defgabcdefgabcdefga


:: 1hw5.0 -- 6 possible alignments
                                       abcdefgabcdefgabcd
  N-RSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQMARRLQVLAEKVGNL-C
  N-RSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQMARRLQVLAEKVGNL-C
                                       abcdefgabcdefgabcd


:: 1ik9.0 -- 2 possible alignments
    defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga                                                    >
  N-IRELIAYALDTIAENQAKNEHLQKENERLLRDWNDVQGRFEKAVSAKEALETDLYKRFILVLNEKKTKIRSLHNKLLNAAQER >
  N-IRELIAYALDTIAENQAKNEHLQKENERLLRDWNDVQGRFEKAVSAKEALETDL-C                                       >
    defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga                                                   >

>
> EKDIKQ-C
>
>


:: 1j1d_1.0 -- 6 possible alignments
                          abcdefgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-REKKKKILAERRKVLAIDHLNEDQLREKAKELWQTIYNLEAEKFDLQEKFKQQKYEINVLRNRI-C
                 N-LELAGLGFAELQDLARQLHARVDKVDEERYDIEAKVTKNITEIADLTQKI-C
                          abcdefgabcdefgabcdefgabcdefgabcdefgabcdefga


:: 1jnm.0 -- 3 possible alignments
                   defgabcdefgabcdefgabcdefgabcdefga
  N-RMRNRIAASKSRKRKLERIARLEEKVKTLKAQNSELASTANMLREQV-C
  N-RMRNRIAASKSRKRKLERIARLEEKVKTLKAQNSELASTANMLREQV-C
                   defgabcdefgabcdefgabcdefgabcdefga


:: 1joc.0 -- 4 possible alignments
    defgabcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-LLERCLKGEGEIEKLQTKVLELQRKLDNTTAAVQELGRENQSLQIKHTQALNRKWAEDNEVQNC-C
  N-LLERCLKGEGEIEKLQTKVLELQRKLDNTTAAVQELGRENQSLQIKHTQALNRKWAEDNEVQNC-C
    defgabcdefgabcdefgabcdefgabcdefgabcdefgabcd
```

```
:: 1kql.0 -- 2 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-VEELLSKNYHLENEVARLKKLVDDLEDELYAQKLKYKAISEELDHA-C
  N-VEELLSKNYHLENEVARLKKLVDDLEDELYAQKLKYKAISEELDHA-C
    abcdefgabcdefgabcdefgabcdefgabcdefgabcd


:: 1lih.0 -- 8 possible alignments
                  abcdefgabcdefgabcd
  N-NELRQQQSELTSTWDLMLQTRINLSRSAARMMMDASNQQSSAKTDLLQNAKTTLAQAAAHYANFKNM-C
  N-NELRQQQSELTSTWDLMLQTRINLSRSAARMMMDASNQQSSAKTDLLQNAKTTLAQAAAHYANFKNM-C
                  abcdefgabcdefgabcd


:: 1lj2.0 -- 4 possible alignments
           abcdefgabcdefgabcd
         N-IPQQQAHIAELQVYNNKLERDLQNKIGSLTSS-C
  N-MHSLQNVIPQQQAHIAELQVYNNKLERDLQNKIGSLTSS-C
           abcdefgabcdefgabcd


:: 1lj2.1 -- 2 possible alignments
    defgabcdefgabcdefgabcdefga
  N-PLHAFDDLESVIRNLISDYDKLFLMFKGLIQRS-C
  N-PLHAFDDLESVIRNLISDYDKLFLMFKGLIQRS-C
    defgabcdefgabcdefgabcdefga


:: 1nkp_1.0 -- 9 possible alignments
                        defgabcdefgabcdefga
         N-KVVILKKATAYILSVQAEEQKLISEEDLLRKRREQLKHKL-C
  N-LQGEKASRAQILDKATEYIQYMRRKNHTHQQDIDDLKRQNALLEQQV-C
                        defgabcdefgabcdefga


:: 1no4_2.0 -- 4 possible alignments
                  abcdefgabcdefgabcdefgabcdefgabcd
  N-QSERTEALQQLRVNYGSFVSEYNDLTKSHEKLAAEKDDLIVSNSKLFRQIGLT-C
  N-QSERTEALQQLRVNYGSFVSEYNDLTKSHEKLAAEKDDLIVSNSKLFRQIGLT-C
                  abcdefgabcdefgabcdefgabcdefgabcd


:: 1nwq.0 -- 5 possible alignments
                                defgabcdefgabcdefgabcdefga
  N-SNEYRVRRERNNIAVRKSRDKAKQRNVETQQKVLELTSDNDRLRKRVEQLSREL-C
  N-SNEYRVRRERNNIAVRKSRDKAKQRNVETQQKVLELTSDNDRLRKRVEQLSREL-C
                                defgabcdefgabcdefgabcdefga


:: 1nyh.0 -- 4 possible alignments
                  abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-LSFVDIVLSKAASALDEKEKQLAVANEIIRSLSDEVMRNEIRITSLQGDLTFTKKCLENARSQISEKDAKINKL-C
  N-LSFVDIVLSKAASALDEKEKQLAVANEIIRSLSDEVMRNEIRITSLQGDLTFTKKCLENARSQISEKDAKINKL-C
                  abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcd


:: 1o5l.0 -- 3 possible alignments
                  abcdefgabcdefgabcd
  N-LDLLMKDRELLLFFLKDVSEHFRVVSEKLFFL-C
  N-LDLLMKDRELLLFFLKDVSEHFRVVSEKLFFL-C
                  abcdefgabcdefgabcd


:: 1omi.0 -- 7 possible alignments
                                      abcdefgabcdefgabcd
  N-DTETSVGYYNLEVISEQATAYVIKINELKELLSKNLTHFFYVFQTLQKQVSYSLAKFNDF-C
  N-DTETSVGYYNLEVISEQATAYVIKINELKELLSKNLTHFFYVFQTLQKQVSYSLAKFNDF-C
                                      abcdefgabcdefgabcd
```

```
:: 1pl5.0 -- 12 possible alignments
                                                                      defgabcdefgabcdefgab >
  N-LDDITDSNTTEILTSVDVLGTHSQTGTQQSNMYTSTQKTELEIDNKDSVTECSKDMKEDGLSFVDIVLSKAASALDEKEKQLA >
  N-LDDITDSNTTEILTSVDVLGTHSQTGTQQSNMYTSTQKTELEIDNKDSVTECSKDMKEDGLSFVDIVLSKAASALDEKEKQLA >
                                                                      defgabcdefgabcdefgab >

> cdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga
> VANEIIRSLSDEVMRNEIRITSLQGDLTFTKKCLENARSQISEKDAKINKLMEKDFQVNKEI-C
> VANEIIRSLSDEVMRNEIRITSLQGDLTFTKKCLENARSQISEKDAKINKLMEKDFQVNKEI-C
> cdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga


:: 1qo0.0 -- 4 possible alignments
            defgabcdefgabcdefgabcd
  N-VLVSARRISEEMAKLKQKTEQLQDRIAGQARINQAKVLLMQRH-C
  N-VLVSARRISEEMAKLKQKTEQLQDRIAGQARINQAKVLLMQRH-C
            defgabcdefgabcdefgabcd


:: 1qp9_1.0 -- 2 possible alignments
            abcdefgabcdefgabcd
          N-ELLKDNELKKLRERVKSL-C
  N-WAEEAEKELLKDNELKKLRERVKSL-C
            abcdefgabcdefgabcd


:: 1s1c.1 -- 7 possible alignments
                                                 abcdefgabcdefgabcd
  N-KDIEILRRENEELTEKMKKAEEEYKLEKEEEISNLKAAFEKNINTERTLKTQAVNKLAEI-C
  N-KDIEILRRENEELTEKMKKAEEEYKLEKEEEISNLKAAFEKNINTERTLKTQAVNKLAEI-C
                                                 abcdefgabcdefgabcd


:: 1t6f.0 -- 5 possible alignments
                    abcdefgabcdefgabcdefgabcdefgabcd
  N-SQYWKEVAEKRRKALYEALKENEKLHKEIEQKDNEIARLKKENKELAEVAEHVQYMAELI-C
  N-SQYWKEVAEKRRKALYEALKENEKLHKEIEQKDNEIARLKKENKELAEVAEHVQYMAELI-C
                    abcdefgabcdefgabcdefgabcdefgabcd


:: 1tu3_3.0 -- 4 possible alignments
            abcdefgabcdefgabcdefgabcd
  N-AQRLQTELDVSEQVQRDFVKLSQTLQVQLERIRQADSLERIRAILN-C
  N-AQRLQTELDVSEQVQRDFVKLSQTLQVQLERIRQADSLERIRAILN-C
            abcdefgabcdefgabcdefgabcd


:: 1uii.0 -- 2 possible alignments
    defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-WKEVAEKRRKALYEALKENEKLHKEIEQKDNEIARLKKENKELAEVAEHVQYMA-C
  N-WKEVAEKRRKALYEALKENEKLHKEIEQKDNEIARLKKENKELAEVAEHVQYMA-C
    defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga


:: 1ukl_1.3 -- 6 possible alignments
                                defgabcdefgabcdefgabcd
  N-SINDKIIELKDLVMGTDAKMHKSGVLRKAIDYIKYLQQVNHKLRQENMVLKLANQKN-C
  N-SINDKIIELKDLVMGTDAKMHKSGVLRKAIDYIKYLQQVNHKLRQENMVLKLANQKN-C
                                defgabcdefgabcdefgabcd


:: 1uuj_2.0 -- 2 possible alignments
            abcdefgabcdefgabcd
  N-LEKKWTSVIRLQKKVMELESKLNEA-C
  N-LEKKWTSVIRLQKKVMELESKLNEA-C
            abcdefgabcdefgabcd
```

```
:: 1wlq_1.0 -- 4 possible alignments
    defgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-WKEVAEQRRKALYEALKENEKLHKEIEQKDSEIARLRKENKDLAEVAEHVQYMAEVIERLS-C
  N-WKEVAEQRRKALYEALKENEKLHKEIEQKDSEIARLRKENKDLAEVAEHVQYMAEVIERLS-C
    defgabcdefgabcdefgabcdefgabcdefgabcdefga


:: 1wu9.0 -- 2 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcd
  N-AAELMQQVNVLKLTVEDLEKERDFYFGKLRNIELICQEN-C
  N-AAELMQQVNVLKLTVEDLEKERDFYFGKLRNIELICQEN-C
    abcdefgabcdefgabcdefgabcdefgabcd


:: 1yhn.0 -- 2 possible alignments
    abcdefgabcdefgabcdefgabcd
  N-FEQILQERNELKAKVFLLKEELAYFQRELLTD-C
  N-FEQILQERNELKAKVFLLKEELAYFQRELLTD-C
    abcdefgabcdefgabcdefgabcd


:: 1yke_1.1 -- 8 possible alignments
            abcdefgabcdefgabcd
  N-HQSRESLIMLLEEQLEYKRGEIREIEQVCKQV-C
  N-AEEQLRKIDMLQKKLVEVEDEKIEAIKKKEKLLRHVDSLIEDFVDG-C
            abcdefgabcdefgabcd


:: 1ytz.0 -- 5 possible alignments
    defgabcdefgabcdefgabcdefgabcdefgabcd
  N-LRDKAKELWDWLYQLQTEKYDFAEQIKRKKYEIVTLRNRIDQA-C
  N-LQELSKKLHAKIDSVDEERYDTEVKLQKTNKELEDLSQKLFDLRGKFKRP-C
    defgabcdefgabcdefgabcdefgabcdefgabcd


:: 1zme.0 -- 3 possible alignments
            abcdefgabcdefgabcd
  N-IVVSTKYLQQLQKDLNDKTEENNRLKALLLER-C
  N-IVVSTKYLQQLQKDLNDKTEENNRLKALLLER-C
            abcdefgabcdefgabcd


:: 2aze.0 -- 5 possible alignments
    abcdefgabcdefgabcdefgabcdefga
  N-CQNLEVERQRRLERIKQKQSQLQELILQQIAFKNLVQRNRHAE-C
  N-LEGLTQDLRQLQESEQQLDHLMNICTTQLRLLSEDT-C
    abcdefgabcdefgabcdefgabcdefga


:: 2c91.0 -- 3 possible alignments
                defgabcdefgabcdefgabcdefgabcd
  N-RYKNRVAARKSRAKFKQLLQHYREVAAAKSSENDRLRLLLKQM-C
  N-RYKNRVAARKSRAKFKQLLQHYREVAAAKSSENDRLRLLLKQM-C
                defgabcdefgabcdefgabcdefgabcd


:: 2coh.0 -- 3 possible alignments
    defgabcdefgabcdefgabcd
  N-LKDLAQHLSQGLAEAYRRIERLATQRLKNRMAAALL-C
  N-LKDLAQHLSQGLAEAYRRIERLATQRLKNRMAAALL-C
    defgabcdefgabcdefgabcd
```

```
:: 2dfs.0 -- 9 possible alignments
                          abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdef >
  N-MNNLEITYSTETEKLRSDVERLRMSEEEAKNATNRVLSLQEEIAKLRKELHQTQTEKKTIEEWADKYKHETEQLVSELKEQNT >
  N-MNNLEITYSTETEKLRSDVERLRMSEEEAKNATNRVLSLQEEIAKLRKELHQTQTEKKTIEEWADKYKHETEQLVSELKEQNT >
                          abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdef >

> gabcdefga
> LLKTEKEELNRRIHDQAKEITETMEKKLVEETKQLELDLNDERL-C
> LLKTEKEELNRRIHDQAKEITETMEKKLVEETKQLELDLNDERL-C
> gabcdefga


:: 2e7s_4.0 -- 13 possible alignments
                          abcdefgabcdefgabcd                                      >
  N-LEEQLNKSLKTIASQKAAIENYNQLKEDYNTLKRELSDRDDEVKRLREDIAKENELRTKAEEEADKLNKEVEDLTASLFDEAN >
  N-LEEQLNKSLKTIASQKAAIENYNQLKEDYNTLKRELSDRDDEVKRLREDIAKENELRTKAEEEADKLNKEVEDLTASLFDEAN >
                          abcdefgabcdefgabcd                                      >

>
> NLVADARMEKYAIEILNKRLTEQLRE-C
> NLVADARMEKYAIEILNKR-C
>


:: 2fmy_1.0 -- 2 possible alignments
    defgabcdefgabcdefgabcd
  N-SLNMVKVLGDLLKNSLTIINGLVFKDARL-C
  N-SLNMVKVLGDLLKNSLTIINGLVFKDARL-C
    defgabcdefgabcdefgabcd


:: 2fxo_2.0 -- 9 possible alignments
    defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga                 >
  N-REKEMASMKEEFTRLKEALEKSEARRKELEEKMVSLLQEKNDLQLQVQAEQDNLADAEERCDQLIKNKIQLEAKVKEMNKRLE >
  N-REKEMASMKEEFTRLKEALEKSEARRKELEEKMVSLLQEKNDLQLQVQAEQDNLADAEERCDQLIKNKIQLEAKVKEMNKRLE >
    defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga                 >

>
> DEEEMNAELTAKKRKLEDECSELKRDIDDLELTL-C
> DEEEMNAELTAKKRKLEDECSELKRDIDDLELTL-C
>


:: 2fxo_2.1 -- 11 possible alignments
                                                                        abcdefgabcdef >
  N-SAEREKEMASMKEEFTRLKEALEKSEARRKELEEKMVSLLQEKNDLQLQVQAEQDNLADAEERCDQLIKNKIQLEAKVKEMNK >
  N-SAEREKEMASMKEEFTRLKEALEKSEARRKELEEKMVSLLQEKNDLQLQVQAEQDNLADAEERCDQLIKNKIQLEAKVKEMNK >
                                                                        abcdefgabcdef >

> gabcdefgabcdefgabcdefgabcdefgabcd
> RLEDEEEMNAELTAKKRKLEDECSELKRDIDDL-C
> RLEDEEEMNAELTAKKRKLEDECSELKRDIDDL-C
> gabcdefgabcdefgabcdefgabcdefgabcd


:: 2gau.0 -- 13 possible alignments
            abcdefgabcdefgabcdefgabcd                                           >
  N-EALLKGNTSFCRYFLKALAKELGYAERRTVTLTQKHVRGRLAETLLILKENFGFENDGATLSIYLSREELATLSNMTVSNAIR >
  N-EALLKGNTSFCRYFLKALAKELGYAERRTVTLTQKHVRGRLAETLLILKENFGFENDGATLSIYLSREELATLSNMTVSNAIR >
            abcdefgabcdefgabcdefgabcd                                           >

>
> TLSTFVSERMLALDGKRIKIIDCDRL-C
> TLSTFVSERMLALDGKRIKIIDCDRL-C
>
```

```
:: 2gzh.0 -- 2 possible alignments
    abcdefgabcdefgabcdefgabcdefga
  N-YEEVLQELVKHKELLRRKDTHIRELEDYIDNLLVRV-C
  N-YEEVLQELVKHKELLRRKDTHIRELEDYIDNLLVRV-C
    abcdefgabcdefgabcdefgabcdefga


:: 2hg4_1.0 -- 2 possible alignments
           abcdefgabcdefgabcd
  N-EEKLRRYLKRTVTELDSVTARLREV-C
  N-EEKLRRYLKRTVTELDSVTARLREV-C
           abcdefgabcdefgabcd


:: 2hl5.0 -- 2 possible alignments
    abcdefgabcdefgabcdefgabcd
  N-VNVLKLTVEDLEKERDFYFGKLRNIELICQEN-C
  N-VNVLKLTVEDLEKERDFYFGKLRNIELICQEN-C
    abcdefgabcdefgabcdefgabcd


:: 2hv8_2.0 -- 2 possible alignments
           abcdefgabcdefgabcdefga
  N-AEISSVSRDELMEAIQKQEEINFRLQDYIDRIIVAI-C
         N-RDELMEAIQKQEEINFRLQDYIDRIIVAI-C
           abcdefgabcdefgabcdefga


:: 2iw5.1 -- 8 possible alignments
                abcdefgabcdefgabcdefgabcd
              N-KHVKDEQIEHWKKIVKTQEELKELLNKMVNLKEKIKELHQQYKEAS-C
  N-VEAVSANATAATTVLRQLDMELVSVKRQIQNIKQTNSALKEKLDGG-C
                abcdefgabcdefgabcdefgabcd


:: 2ncd.0 -- 3 possible alignments
           abcdefgabcdefgabcdefgabcdefgabcd
  N-TEELLRCNEQQAAELETCKEQLFQSNMERKELHNTVMDLRGNIRVF-C
  N-TEELLRCNEQQAAELETCKEQLFQSNMERKELHNTVMDLRGNIRVF-C
           abcdefgabcdefgabcdefgabcdefgabcd


:: 2ocy.1 -- 18 possible alignments
                                                  defgabcdefgabcdefga                >
  N-QLIESVDKQSHLEEQLNKSLKTIASQKAAIENYNQLKEDYNTLKRELSDRDDEVKRLREDIAKENELRTKAEEEADKLNKEVE >
  N-QLIESVDKQSHLEEQLNKSLKTIASQKAAIENYNQLKEDYNTLKRELSDRDDEVKRLREDIAKENELRTKAEEEADKLNKEVE >
                                                  defgabcdefgabcdefga                >

>
> DLTASLFDEANNMVADARKEKYAIEILNKRLTEQLREKDTLLDTLTLQLKNLKKV-C
> DLTASLFDEANNMVADARKEKYAIEILNKRLTEQLREKDTLLDTLTLQLKNLKKV-C
>


:: 2ocy.2 -- 17 possible alignments
                                                                                     >
  N-STQLIESVDKQSHLEEQLNKSLKTIASQKAAIENYNQLKEDYNTLKRELSDRDDEVKRLREDIAKENELRTKAEEEADKLNKE >
  N-STQLIESVDKQSHLEEQLNKSLKTIASQKAAIENYNQLKEDYNTLKRELSDRDDEVKRLREDIAKENELRTKAEEEADKLNKE >
                                                                                     >

>                       abcdefgabcdefgabcdefgabcdefga
> VEDLTASLFDEANNMVADARKEKYAIEILNKRLTEQLREKDTLLDTLTLQLKNLKKVM-C
> VEDLTASLFDEANNMVADARKEKYAIEILNKRLTEQLREKDTLLDTLTLQLKNLKKVM-C
>                       abcdefgabcdefgabcdefgabcdefga


:: bbz2_C_EBPbeta+44_CEBPalpha -- 12 possible alignments
                                     abcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-SKAKKSVDKHSDEYKIRRERNNIAVRKSRDKAKMRNLETQHKVLELTAENERLQKKVEQLSRELSTLRNLFKQL-C
  N-GKAKKSVDKNSNEYRVRRERNNIAVRKSRDKAKQRNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRGIFRQL-C
                                     abcdefgabcdefgabcdefgabcdefgabcdefgabcd


:: bbz3_C_EBPgamma+35_ATF4 -- 12 possible alignments
                                     defgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-SSSMDRNSDEYRQRRERNNMAVKKSRLKSKQKAQDTLQRVNQLKEENERLEAKIKLLTKELSVLKDLFLEHAHNL-C
  N-AKVKGEKLDKKLKKMEQNKTAATRYRQKKRAEQEALTGECKELEKKNEALKERADSLAKEIQYLKDLIEEVRKAR-C
                                     defgabcdefgabcdefgabcdefgabcdefgabcdefga
```

```
:: bbz4_ATF_7+55_MAFK -- 14 possible alignments
                                                                    abcdefgabcdefgabcdef >
                                                        N-RAAASRCRQKRKLWVSSLEKKAEELTSQNIQLSN >
  N-VKKEAGENAAALSDDELVSMSVRELNQHLRGLTREEVVRLKQRRRTLKNRGYAASCRIKRVTQKEELERQRVELQQEVEKLAR >
                                                                    abcdefgabcdefgabcdef >

> gabcdefgabcdefgabcd
> EVTLLRNEVAQLKQLLLAH-C
> ENSSMRLELDALRSKYEALQTFARTV-C
> gabcdefgabcdefgabcd


:: bbz5_ATF_2+10_FOS -- 8 possible alignments
                           abcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-RKFLERNRAAASRCRQKRKVWVQSLEKKAEDLSSLNGQLQSEVTLLRNEVAQLKQLLLAH-C
  N-RIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAH-C
                           abcdefgabcdefgabcdefgabcdefgabcdefgabcd


:: bbz6_CREBPA+28_JUN -- 11 possible alignments
                                  abcdefgabcdefgabcdefgabcdefgabcd
          N-RKFLERNRAAATRCRQKRKVWVMSLEKKAEELTQTNMQLQNEVSMLKNEVAQLKQLLLTH-C
  N-SQERIKAERKRMRNRIAASKCRKRKLERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNHVNSGCQLMLTQQLQ-C
                                  abcdefgabcdefgabcdefgabcdefgabcdefgabcd


:: bbz7_ATF_1+52_CREM -- 8 possible alignments
                                  abcdefgabcdefgabcdefgabcdefgabcd
                    N-REAARECRRKKKEYVKCLENRVAVLENQNKTLIEELKTLKDLYSNK-C
  N-EEATRKRELRLMKNREAARECRRKKKEYVKCLENRVAVLENQNKTLIEELKALKDLYCHK-C
                                  abcdefgabcdefgabcdefgabcdefgabcd


:: bbz7_ATF_1+7_ATF_1 -- 3 possible alignments
                     abcdefgabcdefgabcdefgabcd
  N-REAARECRRKKKEYVKCLENRVAVLENQNKTLIEELKTLKDLYSNK-C
  N-REAARECRRKKKEYVKCLENRVAVLENQNKTLIEELKTLKDLYSNK-C
                     abcdefgabcdefgabcdefgabcdefgabcd
```

bZIPs

```
:: bz10_FOS+27_JUNB -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHR-C
  N-ERIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz10_FOS+28_JUN -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHR-C
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd
```

```
:: bz10_FOS+48_JUND -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcde
  N-ELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHR-C
  N-ERISRLEEKVKTLKSQNTELASTASLLREQVAQLKQKVLSHV-C
    fgabcdefgabcdefgabcdefgabcdefgabcde


:: bz11_FRA1+27_JUNB -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ELTDFLQAETDKLEDEKSGLQREIEELQKQKERLELVLEAHR-C
  N-ERIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz11_FRA1+28_JUN -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ELTDFLQAETDKLEDEKSGLQREIEELQKQKERLELVLEAHR-C
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz11_FRA1+48_JUND -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcde
  N-ELTDFLQAETDKLEDEKSGLQREIEELQKQKERLELVLEAHR-C
  N-ERISRLEEKVKTLKSQNTELASTASLLREQVAQLKQKVLSHV-C
    fgabcdefgabcdefgabcdefgabcdefgabcde


:: bz14_ATF_6+14_ATF_6 -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdef
  N-EYMLGLEARLKAALSENEQLKKENGTLKRQLDEVVSENQRLKV-C
  N-EYMLGLEARLKAALSENEQLKKENGTLKRQLDEVVSENQRLKV-C
    fgabcdefgabcdefgabcdefgabcdefgabcdef


:: bz14_ATF_6+19_XBP_1 -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-EYMLGLEARLKAALSENEQLKKENGTLKRQLDEVVSENQRLKV-C
  N-ARMSELEQQVVDLEEENQKLLLENQLLREKTHGLVVENQEL-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz15_CREB3+63_BBF2 -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-VYVGGLESRVLKYTAQNMELQNKVQLLEEQNLSLLDQLRKLQAMVIEISNKTSS-C
  N-EYVECLEKKVETFTSENNELWKKVETLENANRTLLQQLQKLQTLVTNKISR-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg


:: bz16_unknown_1+16_unknown_1 -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdef
  N-EYIDGLESRVAACSAQNQELQKKVQELERHNISLVAQLRQLQTLIAQTSNKAAQTST-C
  N-EYIDGLESRVAACSAQNQELQKKVQELERHNISLVAQLRQLQTLIAQTSNKAAQTST-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdef


:: bz16_unknown_1+63_BBF2 -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-EYIDGLESRVAACSAQNQELQKKVQELERHNISLVAQLRQLQTLIAQTSNKAAQTST-C
  N-EYVECLEKKVETFTSENNELWKKVETLENANRTLLQQLQKLQTLVTNKISR-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg


:: bz18_MAFB+41_C_MAF -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-QQKHHLENEKTQLIQQVEQLKQEVSRLARERDAYKVKCEKLANSG-C
  N-QQRHVLESEKNQLLQQVDHLKQEISRLVRERDAYKEKYEKLVSSGFREN-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
```

```
:: bz19_XBP_1+19_XBP_1 -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ARMSELEQQVVDLEEENQKLLLENQLLREKTHGLVVENQEL-C
  N-ARMSELEQQVVDLEEENQKLLLENQLLREKTHGLVVENQEL-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz19_XBP_1+43_HCF -- 5 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ARMSELEQQVVDLEEENQKLLLENQLLREKTHGLVVENQEL-C
  N-EYVMGLESRVRGLAAENQELRAENRELGKRVQALQEESRYLRAVLANETGL-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz1_C_EBPepsilon+1_C_EBPepsilon -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-RRILETQQKVLEYMAENERLRSRVEQLTQELDTLRNLFRQI-C
  N-RRILETQQKVLEYMAENERLRSRVEQLTQELDTLRNLFRQI-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz1_C_EBPepsilon+38_DDIT3 -- 5 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-RRILETQQKVLEYMAENERLRSRVEQLTQELDTLRNLFRQI-C
  N-EKEQENERKVAQLAEENERLKQEIERLTREVEATRRALIDRMVNLHQA-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz20_NFE2L3+42_MAFG -- 5 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-DIILNLEDDVCNLQAKKETLKREQAQCNKAINIMKQKLHDL-C
  N-TQKEELEKQKAELQQEVEKLASENASMKLELDALRSKYEALQTFARTVARS-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz20_NFE2L3+55_MAFK -- 5 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-DIILNLEDDVCNLQAKKETLKREQAQCNKAINIMKQKLHDL-C
  N-TQKEELERQRVELQQEVEKLARENSSMRLELDALRSKYEALQTFARTVAR-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz22_NFE2L1+42_MAFG -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-DTILNLERDVEDLQRDKARLLREKVEFLRSLRQMKQKVQSLYQEV-C
  N-TQKEELEKQKAELQQEVEKLASENASMKLELDALRSKYEALQTFARTVARS-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefga


:: bz22_NFE2L1+55_MAFK -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-DTILNLERDVEDLQRDKARLLREKVEFLRSLRQMKQKVQSLYQEV-C
  N-TQKEELERQRVELQQEVEKLARENSSMRLELDALRSKYEALQTFARTVAR-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefga


:: bz23_NFE2+23_NFE2 -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdef
  N-ETIVQLERELERLTNERERLLRARGEADRTLEVMRQQLTELYR-C
  N-ETIVQLERELERLTNERERLLRARGEADRTLEVMRQQLTELYR-C
    fgabcdefgabcdefgabcdefgabcdefgabcdef


:: bz24_NFE2L2+42_MAFG -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-ENIVELEQDLDHLKDEKEKLLKEKGENDKSLHLLKKQLSTLYLEV-C
  N-TQKEELEKQKAELQQEVEKLASENASMKLELDALRSKYEALQTFARTVARS-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
```

```
:: bz24_NFE2L2+55_MAFK -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-ENIVELEQDLDHLKDEKEKLLKEKGENDKSLHLLKKQLSTLYLEV-C
  N-TQKEELERQRVELQQEVEKLARENSSMRLELDALRSKYEALQTFARTVAR-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefga

:: bz26_BACH1+42_MAFG -- 5 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-DCIQNLESEIEKLQSEKESLLKERDHILSTLGETKQNLTGL-C
  N-TQKEELEKQKAELQQEVEKLASENASMKLELDALRSKYEALQTFARTVARS-C
    fgabcdefgabcdefgabcdefgabcdefgabcd

:: bz26_BACH1+55_MAFK -- 5 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-DCIQNLESEIEKLQSEKESLLKERDHILSTLGETKQNLTGL-C
  N-TQKEELERQRVELQQEVEKLARENSSMRLELDALRSKYEALQTFARTVAR-C
    fgabcdefgabcdefgabcdefgabcdefgabcd

:: bz27_JUNB+29_p21SNFT -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ERIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH-C
  N-QKADKLHEEYESLEQENTMLRREIGKLTEELKHLTEALKEHEKMC-C
    fgabcdefgabcdefgabcdefgabcdefgabcd

:: bz27_JUNB+47_BATF -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ERIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH-C
  N-QKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHE-C
    fgabcdefgabcdefgabcdefgabcdefgabcd

:: bz28_JUN+29_p21SNFT -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
  N-QKADKLHEEYESLEQENTMLRREIGKLTEELKHLTEALKEHEKMC-C
    fgabcdefgabcdefgabcdefgabcdefgabcd

:: bz28_JUN+47_BATF -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
  N-QKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHE-C
    fgabcdefgabcdefgabcdefgabcdefgabcd

:: bz29_p21SNFT+38_DDIT3 -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-QKADKLHEEYESLEQENTMLRREIGKLTEELKHLTEALKEHEKMC-C
  N-EKEQENERKVAQLAEENERLKQEIERLTREVEATRRALIDRMVNLHQA-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefga

:: bz29_p21SNFT+48_JUND -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcde
  N-QKADKLHEEYESLEQENTMLRREIGKLTEELKHLTEALKEHEKMC-C
  N-ERISRLEEKVKTLKSQNTELASTASLLREQVAQLKQKVLSHV-C
    fgabcdefgabcdefgabcdefgabcdefgabcde

:: bz2_C_EBPbeta+38_DDIT3 -- 5 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-MRNLETQHKVLELTAENERLQKKVEQLSRELSTLRNLFKQL-C
  N-EKEQENERKVAQLAEENERLKQEIERLTREVEATRRALIDRMVNLHQA-C
    fgabcdefgabcdefgabcdefgabcdefgabcd
```

```
:: bz31_TEF+31_TEF -- 2 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-LKENQITIRAAFLEKENTALRTEVAELRKEVGKCKTIVSKYETK-C
  N-LKENQITIRAAFLEKENTALRTEVAELRKEVGKCKTIVSKYETK-C
     fgabcdefgabcdefgabcdefgabcdefgabcdefg


:: bz31_TEF+60_DBP -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-LKENQITIRAAFLEKENTALRTEVAELRKEVGKCKTIVSKYETK-C
  N-LKENQISVRAAFLEKENALLRQEVVAVRQELSHYRAVLSRYQAQ-C
     fgabcdefgabcdefgabcdefgabcdefgabcdefg


:: bz32_E4BP4+32_E4BP4 -- 2 possible alignments
     fgabcdefgabcdefgabcdefgabcde
  N-LNDLVLENKLIALGEENATLKAELLSLKLKFGLIS-C
  N-LNDLVLENKLIALGEENATLKAELLSLKLKFGLIS-C
     fgabcdefgabcdefgabcdefgabcde


:: bz35_ATF4+44_CEBPalpha -- 5 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcd
  N-AEQEALTGECKELEKKNEALKERADSLAKEIQYLKDLIEEVRKARGKKRV-C
  N-QRNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRGIFRQL-C
     fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz38_DDIT3+44_CEBPalpha -- 5 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcd
  N-EKEQENERKVAQLAEENERLKQEIERLTREVEATRRALIDRMVNLHQA-C
  N-QRNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRGIFRQL-C
     fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz38_DDIT3+47_BATF -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcde
  N-EKEQENERKVAQLAEENERLKQEIERLTREVEATRRALIDRMVNLHQA-C
  N-QKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHE-C
     fgabcdefgabcdefgabcdefgabcdefgabcde


:: bz38_DDIT3+60_DBP -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-EKEQENERKVAQLAEENERLKQEIERLTREVEATRRALIDRMVNLHQA-C
  N-LKENQISVRAAFLEKENALLRQEVVAVRQELSHYRAVLSRYQAQ-C
     fgabcdefgabcdefgabcdefgabcdefgabcdefg


:: bz3_C_EBPgamma+33_ATF5 -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcdefgabc
  N-QKAQDTLQRVNQLKEENERLEAKIKLLTKELSVLKDLFLEHAHNLAD-C
  N-AEGEALEGECQGLEARNRELKERAESVEREIQYVKDLLIEVYKARSQ-C
     fgabcdefgabcdefgabcdefgabcdefgabcdefgabc


:: bz3_C_EBPgamma+35_ATF4 -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcdefgabc
  N-QKAQDTLQRVNQLKEENERLEAKIKLLTKELSVLKDLFLEHAHNLAD-C
  N-AEQEALTGECKELEKKNEALKERADSLAKEIQYLKDLIEEVRKARGKKRV-C
     fgabcdefgabcdefgabcdefgabcdefgabcdefgabc


:: bz3_C_EBPgamma+38_DDIT3 -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcdefgabc
  N-QKAQDTLQRVNQLKEENERLEAKIKLLTKELSVLKDLFLEHAHNLAD-C
  N-EKEQENERKVAQLAEENERLKQEIERLTREVEATRRALIDRMVNLHQA-C
     fgabcdefgabcdefgabcdefgabcdefgabcdefgabc
```

```
:: bz3_C_EBPgamma+47_BATF -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcde
  N-QKAQDTLQRVNQLKEENERLEAKIKLLTKELSVLKDLFLEHAHNLAD-C
  N-QKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHE-C
    fgabcdefgabcdefgabcdefgabcdefgabcde


:: bz3_C_EBPgamma+6_CREBPA -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-QKAQDTLQRVNQLKEENERLEAKIKLLTKELSVLKDLFLEHAHNLAD-C
  N-VWVMSLEKKAEELTQTNMQLQNEVSMLKNEVAQLKQLLLTHKDC-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefg


:: bz41_C_MAF+41_C_MAF -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcde
  N-QQRHVLESEKNQLLQQVDHLKQEISRLVRERDAYKEKYEKLVSSGFREN-C
  N-QQRHVLESEKNQLLQQVDHLKQEISRLVRERDAYKEKYEKLVSSGFREN-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcde


:: bz43_HCF+43_HCF -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-EYVMGLESRVRGLAAENQELRAENRELGKRVQALQEESRYLRAVLANETGL-C
  N-EYVMGLESRVRGLAAENQELRAENRELGKRVQALQEESRYLRAVLANETGL-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg


:: bz44_CEBPalpha+44_CEBPalpha -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-QRNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRGIFRQL-C
  N-QRNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRGIFRQL-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz47_BATF+48_JUND -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcde
  N-QKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHE-C
  N-ERISRLEEKVKTLKSQNTELASTASLLREQVAQLKQKVLSHV-C
    fgabcdefgabcdefgabcdefgabcdefgabcde


:: bz49_CREBH+49_CREBH -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdef
  N-EYIDGLETRMSACTAQNQELQRKVLHLEKQNLSLLEQLKKLQAIVVQSTS-C
  N-EYIDGLETRMSACTAQNQELQRKVLHLEKQNLSLLEQLKKLQAIVVQSTS-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdef


:: bz4_ATF_7+27_JUNB -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-LWVSSLEKKAEELTSQNIQLSNEVTLLRNEVAQLKQLLLAHKDC-C
  N-ERIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz4_ATF_7+28_JUN -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-LWVSSLEKKAEELTSQNIQLSNEVTLLRNEVAQLKQLLLAHKDC-C
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz4_ATF_7+48_JUND -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcde
  N-LWVSSLEKKAEELTSQNIQLSNEVTLLRNEVAQLKQLLLAHKDC-C
  N-ERISRLEEKVKTLKSQNTELASTASLLREQVAQLKQKVLSHV-C
    fgabcdefgabcdefgabcdefgabcdefgabcde
```

```
:: bz51_CREBdelta+51_CREBdelta -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdef
  N-EYVKCLENRVAVLENQNKTLIEELKALKDLYCHKSD-C
  N-EYVKCLENRVAVLENQNKTLIEELKALKDLYCHKSD-C
    fgabcdefgabcdefgabcdefgabcdef


:: bz51_CREBdelta+52_CREM -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdef
  N-EYVKCLENRVAVLENQNKTLIEELKALKDLYCHKSD-C
  N-EYVKCLENRVAVLENQNKTLIEELKALKDLYCHKVE-C
    fgabcdefgabcdefgabcdefgabcdef


:: bz52_CREM+52_CREM -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdef
  N-EYVKCLENRVAVLENQNKTLIEELKALKDLYCHKVE-C
  N-EYVKCLENRVAVLENQNKTLIEELKALKDLYCHKVE-C
    fgabcdefgabcdefgabcdefgabcdef


:: bz59_HLF+59_HLF -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-LKENQIAIRASFLEKENSALRQEVADLRKELGKCKNILAKYEARH-C
  N-LKENQIAIRASFLEKENSALRQEVADLRKELGKCKNILAKYEARH-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefga


:: bz5_ATF_2+28_JUN -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcd
  N-VWVQSLEKKAEDLSSLNGQLQSEVTLLRNEVAQLKQLLLAHKDC-C
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz63_BBF2+63_BBF2 -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-EYVECLEKKVETFTSENNELWKKVETLENANRTLLQQLQKLQTLVTNKISR-C
  N-EYVECLEKKVETFTSENNELWKKVETLENANRTLLQQLQKLQTLVTNKISR-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefg


:: bz6_CREBPA+27_JUNB -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-VWVMSLEKKAEELTQTNMQLQNEVSMLKNEVAQLKQLLLTHKDC-C
  N-ERIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz6_CREBPA+28_JUN -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcd
  N-VWVMSLEKKAEELTQTNMQLQNEVSMLKNEVAQLKQLLLTHKDC-C
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
    fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz6_CREBPA+48_JUND -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcde
  N-VWVMSLEKKAEELTQTNMQLQNEVSMLKNEVAQLKQLLLTHKDC-C
  N-ERISRLEEKVKTLKSQNTELASTASLLREQVAQLKQKVLSHV-C
    fgabcdefgabcdefgabcdefgabcdefgabcde


:: bz7_ATF_1+51_CREBdelta -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdef
  N-EYVKCLENRVAVLENQNKTLIEELKTLKDLYSNKSV-C
  N-EYVKCLENRVAVLENQNKTLIEELKALKDLYCHKSD-C
    fgabcdefgabcdefgabcdefgabcdef
```

```
:: bz7_ATF_1+52_CREM -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdef
  N-EYVKCLENRVAVLENQNKTLIEELKTLKDLYSNKSV-C
  N-EYVKCLENRVAVLENQNKTLIEELKALKDLYCHKVE-C
     fgabcdefgabcdefgabcdefgabcdef


:: bz7_ATF_1+7_ATF_1 -- 2 possible alignments
     fgabcdefgabcdefgabcdefgabcdef
  N-EYVKCLENRVAVLENQNKTLIEELKTLKDLYSNKSV-C
  N-EYVKCLENRVAVLENQNKTLIEELKTLKDLYSNKSV-C
     fgabcdefgabcdefgabcdefgabcdef


:: bz8_FOSB+27_JUNB -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHK-C
  N-ERIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH-C
     fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz8_FOSB+28_JUN -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHK-C
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
     fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz8_FOSB+48_JUND -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcde
  N-ELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHK-C
  N-ERISRLEEKVKTLKSQNTELASTASLLREQVAQLKQKVLSHV-C
     fgabcdefgabcdefgabcdefgabcdefgabcde


:: bz9_FRA2+27_JUNB -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ELTEKLQAETEELEEEKSGLQKEIAELQKEKEKLEFMLVAHG-C
  N-ERIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH-C
     fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz9_FRA2+28_JUN -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcd
  N-ELTEKLQAETEELEEEKSGLQKEIAELQKEKEKLEFMLVAHG-C
  N-ERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNH-C
     fgabcdefgabcdefgabcdefgabcdefgabcd


:: bz9_FRA2+48_JUND -- 4 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcde
  N-ELTEKLQAETEELEEEKSGLQKEIAELQKEKEKLEFMLVAHG-C
  N-ERISRLEEKVKTLKSQNTELASTASLLREQVAQLKQKVLSHV-C
     fgabcdefgabcdefgabcdefgabcdefgabcde


:: bzCC156+CC156 -- 2 possible alignments
     fgabcdefgabcdefgabcdefgabcdef
  N-QRMKQLEDKVEELLSKNYHLENEVARLKKLVGDAAR-C
  N-QRMKQLEDKVEELLSKNYHLENEVARLKKLVGDAAR-C
     fgabcdefgabcdefgabcdefgabcdef


:: bzCC163+CC163 -- 2 possible alignments
     fgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-LKEKELESSIHELTEIAASLQKRIHTLETENKLLKNLVLSSGET-C
  N-LKEKELESSIHELTEIAASLQKRIHTLETENKLLKNLVLSSGET-C
     fgabcdefgabcdefgabcdefgabcdefgabcdefg
```

```
:: bzCC40+CC40 -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-NKLQVLEETIESLSKVVKNYETKLNRLQNELQAKESENHALKQKLETLTLKQASV-C
  N-NKLQVLEETIESLSKVVKNYETKLNRLQNELQAKESENHALKQKLETLTLKQASV-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcd

:: bzCC66+CC163 -- 4 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefg
  N-QKNFENMNKLQNLNTQINKLRDRIEQLNKENEFWKAKLNDINEIKS-C
  N-LKEKELESSIHELTEIAASLQKRIHTLETENKLLKNLVLSSGET-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefg

:: bzSCbZ_1+SCbZ_1 -- 2 possible alignments
    fgabcdefgabcdefgabcdefgab
  N-ERLEELEKKEAQLTVTNDQIHILKKENELLHF-C
  N-ERLEELEKKEAQLTVTNDQIHILKKENELLHF-C
    fgabcdefgabcdefgabcdefgab

:: bzSCbZ_8+SCbZ_8 -- 2 possible alignments
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-TRIQVLEEKVEMLHNLVDDWQRKYKLLESEFSDTKENLQKSIALNNELQKAL-C
  N-TRIQVLEEKVEMLHNLVDDWQRKYKLLESEFSDTKENLQKSIALNNELQKAL-C
    fgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefga
```

Parallel Heterodimers

```
:: 1j1d_1.0 -- 6 possible alignments
                        abcdefgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-REKKKKILAERRKVLAIDHLNEDQLREKAKELWQTIYNLEAEKFDLQEKFKQQKYEINVLRNRI-C
                 N-LELAGLGFAELQDLARQLHARVDKVDEERYDIEAKVTKNITEIADLTQKI-C
                        abcdefgabcdefgabcdefgabcdefgabcdefgabcdefga

:: 1yke_1.1 -- 8 possible alignments
           abcdefgabcdefgabcd
  N-HQSRESLIMLLEEQLEYKRGEIREIEQVCKQV-C
  N-AEEQLRKIDMLQKKLVEVEDEKIEAIKKKEKLLRHVDSLIEDFVDG-C
           abcdefgabcdefgabcd

:: 1ytz.0 -- 5 possible alignments
    defgabcdefgabcdefgabcdefgabcdefgabcd
  N-LRDKAKELWDWLYQLQTEKYDFAEQIKRKKYEIVTLRNRIDQA-C
  N-LQELSKKLHAKIDSVDEERYDTEVKLQKTNKELEDLSQKLFDLRGKFKRP-C
    defgabcdefgabcdefgabcdefgabcdefgabcd

:: 2aze.0 -- 5 possible alignments
    abcdefgabcdefgabcdefgabcdefga
  N-CQNLEVERQRRLERIKQKQSQLQELILQQIAFKNLVQRNRHAE-C
  N-LEGLTQDLRQLQESEQQLDHLMNICTTQLRLLSEDT-C
    abcdefgabcdefgabcdefgabcdefga

:: 2iw5.1 -- 8 possible alignments
                  abcdefgabcdefgabcdefgabcd
                N-KHVKDEQIEHWKKIVKTQEELKELLNKMVNLKEKIKELHQQYKEAS-C
  N-VEAVSANATAATTVLRQLDMELVSVKRQIQNIKQTNSALKEKLDGG-C
                  abcdefgabcdefgabcdefgabcd
```

```
:: bbz3_C_EBPgamma+35_ATF4 -- 12 possible alignments
                                        defgabcdefgabcdefgabcdefgabcdefgabcdefga
  N-SSSMDRNSDEYRQRRERNNMAVKKSRLKSKQKAQDTLQRVNQLKEENERLEAKIKLLTKELSVLKDLFLEHAHNL-C
  N-AKVKGEKLDKKLKKMEQNKTAATRYRQKKRAEQEALTGECKELEKKNEALKERADSLAKEIQYLKDLIEEVRKAR-C
                                        defgabcdefgabcdefgabcdefgabcdefgabcdefga


:: bbz5_ATF_2+10_FOS -- 8 possible alignments
                             abcdefgabcdefgabcdefgabcdefgabcdefgabcd
  N-RKFLERNRAAASRCRQKRKVWVQSLEKKAEDLSSLNGQLQSEVTLLRNEVAQLKQLLLAH-C
  N-RIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAH-C
                             abcdefgabcdefgabcdefgabcdefgabcdefgabcd


:: gi|2119236|pir||I37984+gi|186772.0.0 -- 14 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcdefgabc                                    >
  N-IDDLKDQIVDLTVGNNKTLLDIDNTRMTLDDFRIKFEMEQNLRQGVDADINGLRQVLDNLTMEKSDLEMQYETLQEELMALKK >
  N-INNLRRRVDQLKSDQSRLDSELKNMQDMVEDYRNKYEDEINKRTNAENEFVTIKKDVDGAYMTKVDLQAKLDNLQQEIDFLTA >
    abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabc                              >

>
> NHKE-C
> LYQA-C
>


:: gi|547749|sp|P13645|K1CJ_+gi|186772.0.0 -- 14 possible alignments
    abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabc                             >
  N-IDDLKNQILNLTTDNANILLQIDNARLAADDFRLKYENEVALRQSVEADINGLRRVLDELTLTKADLEMQIESLTEELAYLKK >
  N-INNLRRRVDQLKSDQSRLDSELKNMQDMVEDYRNKYEDEINKRTNAENEFVTIKKDVDGAYMTKVDLQAKLDNLQQEIDFLTA >
    abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabc                              >

>
> NHEE-C
> LYQA-C
>
```