Computational Design of Orthogonal Antiparallel Homodimeric Coiled Coils

by

Christopher Negron

B.S. Physics
City College of New York, CUNY, 2008

Submitted to the Program of Computational and Systems Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[September 2014]
June 2014

Signature of Author: Signature redacted
Program of Computational and Systems Biology
June 16, 2014

Certified by: Signature redacted
Amy E. Keating
Associate Professor of Biology
Thesis Supervisor

Accepted by: Signature redacted
Christopher B. Burge
Professor of Biology and Biological Engineering
Director of Computational and Systems Biology Ph.D. program

Computational Design of Orthogonal Antiparallel Homodimeric Coiled Coils

by

Christopher Negron

Submitted to the Program of Computational and Systems Biology
on June 16, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Living cells integrate a vast array of protein-protein interactions (PPIs) to govern cellular functions. For instance, PPIs are critical to biosynthesis, nanostructural assembly, and in processing environmental stimuli through cell-signaling pathways. As fields such as synthetic biology and protein engineering mature they seek to mimic and expand the functions found in living systems that integrate PPIs. A critical feature to many PPIs that are integrated together to perform a complex function is orthogonality, i.e. PPIs that do not cross interact with each other. The engineering of orthogonal PPIs is thus an alluring problem. Since it not only tests our understanding of molecular specificity by having to stabilize and destabilize interactions simultaneously. The results of the design process can also have interesting applications in synthetic biology or bionanotechnology. The coiled coil, a rope-like structure made of helices, is a PPI ubiquitously found in biological systems and is an attractive fold for engineering orthogonal PPIs. Though the coiled coil is well studied, destabilization of undesired interactions still remains challenging. In this thesis I will discuss strategies for obtaining orthogonal PPIs, and describe the current sequence-to-structure relationships known about coiled coils. I will then introduce the computational multistate design framework, CLASSY, and explain how I applied it to the computational design of six orthogonal antiparallel homodimeric coiled coils. Five of these designed sequences were experimentally tested, of which only three of the sequences adopted the target antiparallel homodimer topology. All three of these sequences, as well as a previously designed antiparallel homodimer, were tested for cross reactivity in a pairwise manner. None of these sequences appeared to cross react. The sequences that failed to adopt the antiparallel topology highlight the need for improving our computational design framework. In the final chapter I will discuss strategies to improve our models, and applications for orthogonal antiparallel coiled coils.

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

Acknowledgments

There are many people who contributed to this work, and my overall well-being during graduate school, and owe my thanks to.

First and for most, I'd like to thank my advisor Amy Keating. Even before I came to MIT, Amy was supportive and helpful. During my years in her lab she has given me the freedom to try and learn a variety of techniques. They span from computational approaches such as molecular dynamics to experimental techniques like analytical ultracentrifugation. Amy has greatly shaped how think, and communicate, both in science and more broadly in life.

I'd like to thank my committee members Bruce Tidor, and Jeremy England. They have provided me with interesting ideas, and perspectives during my time at MIT. Additionally, I liked to thank Mark Bathe for taking the time to listen and think about my research.

I'd like to thank Debby Pheasant for all her help with the instruments in the BIF. She is incredibly patient, a great teacher, and always fun to talk to.

I'd like to thank the many people I have worked with in the Keating lab, who have now become my friends. I'd especially like to thank the former and current graduate students Aaron Reinke, Orr Ashenberg, Scott Chen, Jen Kaplan, Glenna Foight, and Vincent Xue. I have learned so many things from the each of you, and have enjoyed our Friday beer hours together that have often lasted much longer then an hour. Additionally, I'd like to thank all the post docs in the lab, particularly Vladimir Potapov, Karl Hauschild, and Raheleh Rezaei Araghi. They spent a lot of time teaching me or helping me trouble shoot my problems. And of course, I'd like to thank Christos Kougentakis. He has listened to me complain about many things, and helped me unwind after many trying times.

Lastly, I'd like to thank my friends and family, especially my mom and dad, Milagros and Arcangel Negron. They have provided me with a lot of love, which has kept me strong during the challenges of life. I'd like to thank my brothers Michael, Gabriel, and Anthony. They are the best set of brothers a guy could have, and it's always good to know they have my back. I'd also like to thank my friends Danny Park, Tracy Washington, and Joseph McNally for many good times outside of work. Lastly, I'd like to thank two really special people, Luisirene Hernandez and Faye-Marie Vassel. I shared many great times with you both, and that will always be with me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Fields such as synthetic biology and protein engineering are seeking to reengineer the molecules that carry out the work of living cells in order to design biomolecular systems intended to reduce the cost of manufacturing drugs, produce "green" fuels, and design targeted therapeutics (Khalil & Collins, 2010). However, biomolecules are incredibly complex, making the manipulation and construction of novel biomolecular systems exceedingly challenging. Thus these fields are turning their focus to the design of modular parts to facilitate the engineering of these systems (Bromley et al., 2009; Purnick & Weiss, 2009). Many fields of engineering rely on the design of modular parts to expedite the fabrication of complex devices. This is because modular parts aid in managing the complexity of a system by dividing it into separate components. This allows for the parallel development of the various modules that comprise the system, enabling progress to be achieved more rapidly. Perhaps even more importantly, modular parts help to reduce future uncertainty by providing robust behavior in a variety of contexts. Additionally, when integrating modular parts together to obtain more complex functions, orthogonality becomes a key property. A part is orthogonal if it does not cross-react with other components in the system, allowing components to be integrated together.

Primarily, the design of modular biomolecular parts has focused on the development of reagents encoded in genetic elements that are involved in transcriptional regulation (Tamsir et al., 2010). However, vital parts to many biomolecular systems within cells are protein-protein interactions (PPIs). The development and use of modular protein-protein interactions (PPIs) is thus particularly alluring, and has several advantages over other biomolecular reagents. Through recombinant DNA technology it is possible to express the components of a PPI within bacterial cells, simplifying synthesis. Also, the proteins and polypeptides in PPIs are made up of a chemically diverse set of 20 naturally occurring amino acids that provide a wide-range of functions, and are now being expanded on by the development of nonnatural amino acids (Hendrickson et al., 2004). Lastly, many cellular functions depend on PPIs, thus in order to mimic these behaviors and rapidly modify them it will be advantageous to develop modular PPI parts. However, in order to make PPI parts that can be integrated with each other, it will be necessary to develop orthogonal PPIs (Figure 1-1), PPIs that do not cross talk. This is because cross talk can be devastating to the functions of PPIs such as signal transduction in cell-signaling pathways, enzymatic scaffolding, and assembly of nanostructures (Kapp et al., 2012; Tsai et al., 2013; Gradišar et al., 2013) (Figure 1-2). In this chapter, I will discuss the strategies for developing orthogonal PPIs. Then I will provide a detailed description of the coiled-coil PPI. Explain how it is an ideal molecular reagent. Present the known sequence-to-structure relationships (design rules) of coiled coils, and lastly summarize the work in this thesis towards designing and testing orthogonal antiparallel coiled-coil interactions.

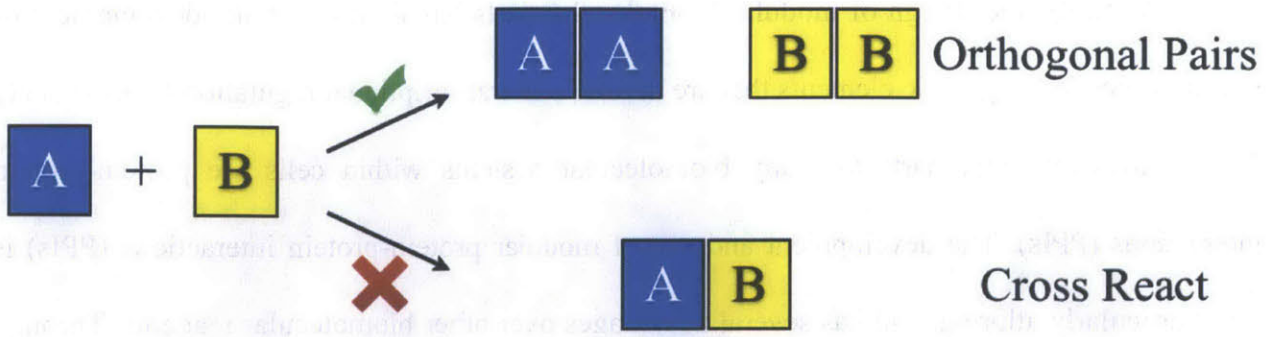


Figure 1-1. Concept of an Orthogonal PPI Pair. Boxes represent protein monomers. The example depicts the simplest orthogonal pair, two homodimers that do not cross-react to form a heterodimer. A true orthogonal pair would not form any type of A-B complex.

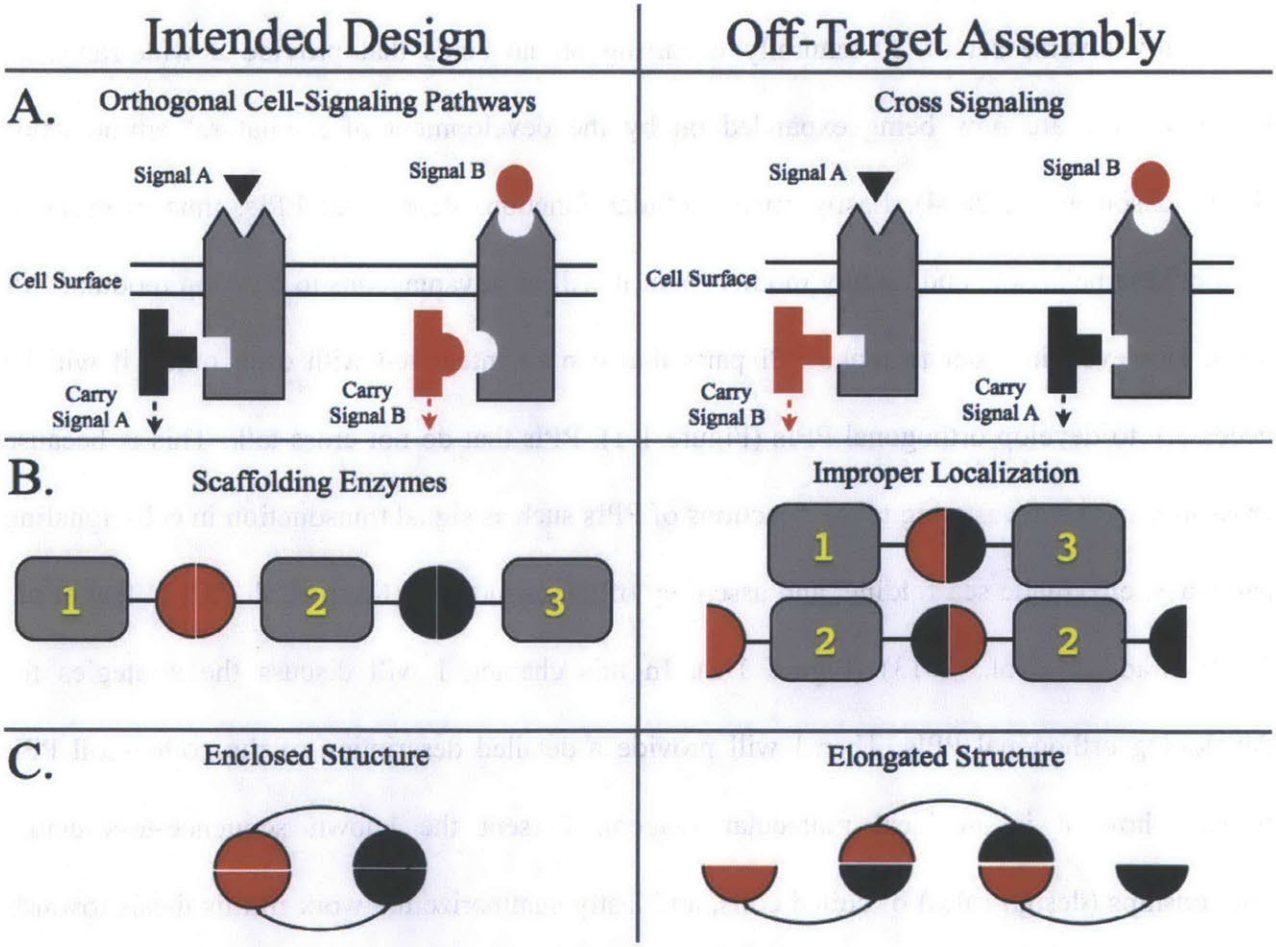


Figure 1-2. Consequences of PPI cross talk. Three possible applications of PPIs are shown in the left column, with improper assembly, due to cross talk, shown on the right. (A) The introduction of two signaling pathways, A and B, into a cell is shown. Binding of the signal molecule by a membrane receptor (gray rectangle) in both pathways is propagated into the cell via a PPI between a receptor and an intracellular protein. Signal propagation is improperly transduced into the cell if the PPIs can cross talk. (B) A pair of orthogonal PPIs is shown as

colored circles. Enzymes involved in an enzymatic pathway are shown as numbered rectangles. Improper localization of the enzymes will occur in the presence of cross talk shown by differentially colored circles. Similarly in (C), improper folding of an intended super structure will occur if the PPIs engage in cross talk.

1.1 Strategies for obtaining orthogonal PPIs

Four strategies for obtaining orthogonal PPIs will be presented in this section. Briefly, they involve using native PPIs that are structurally distinct, mining structurally similar PPIs for orthogonal PPIs in the native, or non-native sequence space, or computationally designing orthogonal PPIs. Examples from the literature that implement these strategies will be highlighted, along with the benefits and challenges of each of these techniques.

The first strategy involves engineering a network of proteins that use structurally distinct protein interaction domains. Key to this strategy is that structurally distinct domains are unlikely to cross react due to their geometrically distinct interfaces. This strategy was implemented by Lai et al. to engineer a proof-of-concept molecular cage made from PPIs (Lai et al., 2012). In brief, the approach taken by Lai et al. for engineering a molecular cage using PPIs involves fusing the domains of multimeric proteins together through a linker sequence (Figure 1-3A). An angle, Θ, is defined between the interfaces of the PPIs that compose the cage (Figure 1-3B). Depending on the oligomer adopted by each PPI, certain values of Θ will result in cage-like structures. The linker sequence between the PPIs must be designed to enforce the desired angle between the PPI interfaces to get the desired super structure (Padilla et al., 2001). If the PPIs cross talk, misfolding will occur due to inappropriate pairing, similar to Figure 1-2C. Lai et al. fused the trimerizing domain of bromoperoxidase with the dimerizing domain of M1 virus matrix protein,

through a rigid α-helical linker to obtain molecular cages. These domains did not engage in any detectable cross talk. The advantage of using structurally distinct protein interaction domains is that many unique PPI interfaces exist in nature. Estimates place the total number of structurally unique PPIs in nature at about ~4,000, allowing for many unique combinations (Garma et al.,

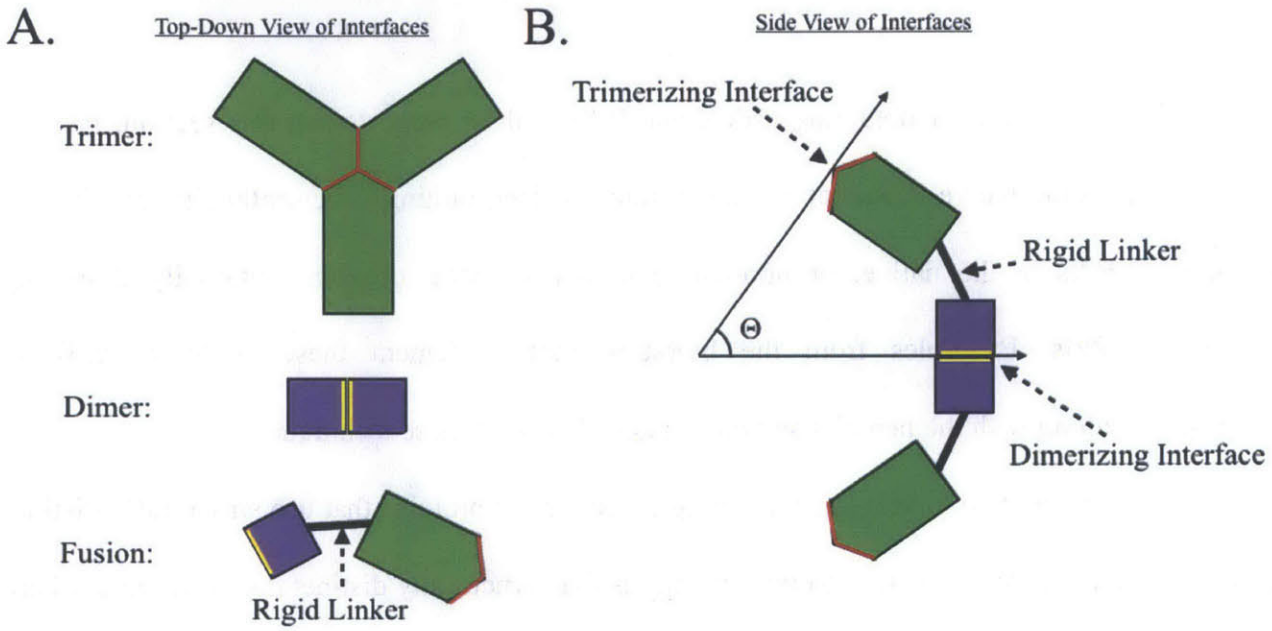


Figure 1-3. PPI applied to the design of molecular cages. (A) A trimerizing domain is shown in green. Its interface is colored red. A dimerizing domain is shown in purple, with its interface colored in yellow. These domains are then fused together through a linker sequence. These domains must not interact with each other for proper assembly. (B) The angle, Θ, between the interfaces is shown. A rigid linker must hold the PPI interfaces at the angle Θ, then upon oligomerization the desired structural assembly will occur.

2012). One disadvantage however is that if native PPI domains are to be used in application in cells, the constructs are likely to cross-react with their native counter parts. Complications can even arise when using native PPIs from one organism in a different organism. Zarrinpar et al. observed that one peptide fragment from yeast, Pbs2, known to specifically bind the Sho1 SH3 domain in yeast, cross-reacted with several non-yeast SH3 domains (Zarrinpar et al., 2003). Additionally, it is unclear how robust the assumption that structurally distinct domains will be

orthogonal to each other. Mechanisms such as domain swapping (Liu & Eisenberg, 2002) of beta-strands or other common secondary-structural elements may result in unwanted cross talk between PPIs with geometrically distinct interfaces.

A second strategy involves mining homologous families of PPIs for orthogonal pairs. Orthogonal PPI pairs have evolved in homologous families in order to maintain cell viability during coexpression in a cell. For example, orthogonal components have been observed in the two-component signal transduction pathway of prokaryotes (Laub & Goulian, 2007; Ashenberg et al., 2011). These pathways consist of a dimeric histidine kinase that can transduce a signal by binding to and phosphorylating a protein known as a response regulator. In prokaryotes, dozens to hundreds of these pathways coexist in the same organism. It has been observed that both the dimerization of the histidine kinase, and the binding of the response regulator to the histidine kinase occur with strong preferences for their cognate partner. In another case, a study by Reinke et al. uncovered a dozen or so orthogonal pairs among a set of bZIP transcription factors obtained from multiple organisms, which dimerize to perform their function (Reinke et al., 2013).

The benefit of using interaction domains from large families of homologs is that it takes advantage of the billions of years of evolutionary selection pressure for orthogonal PPIs. However, screening native biological components for orthogonal PPIs remains challenging due to the vast size of these spaces. In the study by Reinke et al., the largest orthogonal set was slightly less then a dozen pairs, however nearly 2900 binding curves were measured. Additionally, biological systems have not evolved to achieve the minimal amount of cross talk between two components. For instance, given enough time a histidine kinase will begin to phosphorylate non-cognate partners. Lastly, use of native orthogonal PPIs within cells may result

in cross talk with the native sequence it was derived from, or related homologs, as mentioned earlier.

A third strategy for obtaining orthogonal PPIs involves designing synthetic libraries of PPIs, and then screening them for orthogonality. Typically, one member of the PPI is varied. Variants are then selected for binding to the non-varied interaction partner (target). Specificity can be introduced by selecting for variants that only bind to the target in the presence of a competing homolog (off target) that does not interact with the target. Orthogonality can then be introduced by screening the library again; however in this screen the off target is now the target, and the original target is now the off target. For instance, Dutta et al. used a combination of SPOT arrays and yeast surface display experiments to search for a mutant variant of the BH3 peptide Bim that would specifically bind to the pro-apoptotic protein Bcl-x_L over its homolog Mcl-1, and then repeated this screen, but for specificity towards Mcl-1 (Dutta et al., 2010). One peptide showed 100-fold specificity for Bcl-x_L, and another peptide showed 1000-fold specificity for Mcl-1, making an orthogonal pair of PPIs. This strategy has also been applied to the search for orthogonal protein-DNA interactions. Temme et al. designed libraries of T7 RNA polymerases by combing fragment sequences from homologs of T7 RNA polymerase into four T7 RNA polymerase scaffolds known to reduce cytotoxicity to cells (Temme et al., 2012). These recombined T7 RNA polymerase variants were then screened for activity and preferential binding to the recombined fragment's cognate promoter site over non-cognate sites. One challenge with this strategy is the ability to develop a selection assay that properly destabilizes the off-target state. For instance, in yeast display, detection of binding involves the binding and washing of fluorescent antibodies. However, sequences that strongly bind the off-target state can

evade detection as long as the off rate of the binding to the off target is fast. Caveats like these are often hard to remove from selection-based assays. A second challenge is the ability to screen against many off-target states at once. In the work by Dutta et al., only one additional off-target state was screened against.

Lastly, computational design is a very attractive strategy for obtaining orthogonal PPIs. It provides a cheap way to screen a large number of sequences for orthogonality. Additionally, it provides a rigorous test of our understanding of PPIs. Kortemme et al. devised a structural modeling approach termed the computational second-site suppressor strategy for designing orthogonal pairs implemented in the Rosetta framework (Kortemme et al., 2004). This computational framework takes the structure of a known PPI as an input. It then screens the known PPI for mutations that are predicted to disrupt binding. This is then followed by a search for a compensatory mutation in the protein partner. The design pair is then thought to be orthogonal to the parent PPI. This was first applied to a DNase-inhibitor PPI, and has now been extended to several other PPIs (Sammond et al., 2010). In a recent hallmark paper, Kapp et al. applied the second-site suppressor strategy to the GTPase/GEF PPI (Kapp et al., 2012). GTPase/ GEF pairs are interesting due to their role as binary switches in many signaling pathways in mammalian cells. The synthetic GTPase/GEF pair was not only experimentally confirmed to be orthogonal to its wild-type parent PPI, but the synthetic pair functioned as an orthogonal signaling pathway within mammalian cells. Despite the success of this computational technique, this strategy has not been applied to the design of multiple orthogonal pairs at once, or the complete redesign of a PPI. Additionally, using computational structure-based models to predict whether a single-point mutation in a protein structure is stabilizing or destabilizing remains

challenging (Kellogg et al., 2010). One approach for overcoming the hurdles mentioned in this strategy is to use a well-understood PPI, where models can more reliably predict the properties of an interaction. The coiled coil is such a PPI. The characteristics and benefits of using coiled coils will be the focus of the following sections.

1.2 Introduction to coiled-coil structure and sequence

Over 50 years ago, Linus Pauling and Francis Crick proposed that the X-ray diffraction data of α-keratin could be explained by a structural motif they referred to as the coiled coil (Pauling & Corey, 1953; Crick, 1953). Since then, the coiled coil has been discovered to be a part of many protein-based macromolecular structures. It is predicted that 3-5% of the coding sequences in all genomes encode amino acids that are part of a coiled-coil structure (Wolf et al., 1997). As a result, the coiled coil has become a model system for studying PPIs (Woolfson, 2005).

The coiled-coil structure is made of α-helices that supercoil around each other, typically with a left-handed twist (Figure 1-4A). The amino-acid side chains between the helices arrange in an interlocking fashion that is defined as "knobs-into-holes" packing. Coiled coils adopt a wide range of topologies. For instance, coiled-coil structures can vary in the number of helices, the most prevalent of which are dimers, trimers, and tetramers (Figure 1-4B). Higher-order complexes like heptamers have also been observed (Liu et al., 2006). Additionally, adjacent helices can occupy two types of orientations with respect to each other. They can be either parallel, meaning their N-terminal ends pack against each other, or antiparallel, meaning the N-

terminal end of one helix packs against the C-terminal end of the other helix (Figure 1-4C). Finally, coiled coils can form homo- or heteroassemblies.

Interestingly, this diverse set of structural topologies can all be encoded by a repeating sequence pattern typically called the heptad repeat, with the seven positions denoted as *abcdefg* (Figure 1-4D). The *a* and *d* positions are located in the cores of coiled-coil structures, and are typically occupied by hydrophobic residues such as isoleucine or leucine. These residues provide the driving force for binding, as it is more energetically favorable for them to be occluded from water. The *e* and *g* positions are partially buried and are most frequently occupied by charged residues like lysine and glutamate. They often make salt-bridge interactions along the interface of the helices. Lastly, the *b*, *c*, and *f* positions reside on the surface of the coiled coil. They are occupied by polar residues such as glutamine and help promote solubility. These simple sequence features have been used to design several synthetic coiled-coil complexes, which have helped to further our understanding of coiled-coil sequence to structure relationships (design rules). Several examples of synthetic coiled-coil complexes, and the design rules learned from them, will be summarized in the next section.

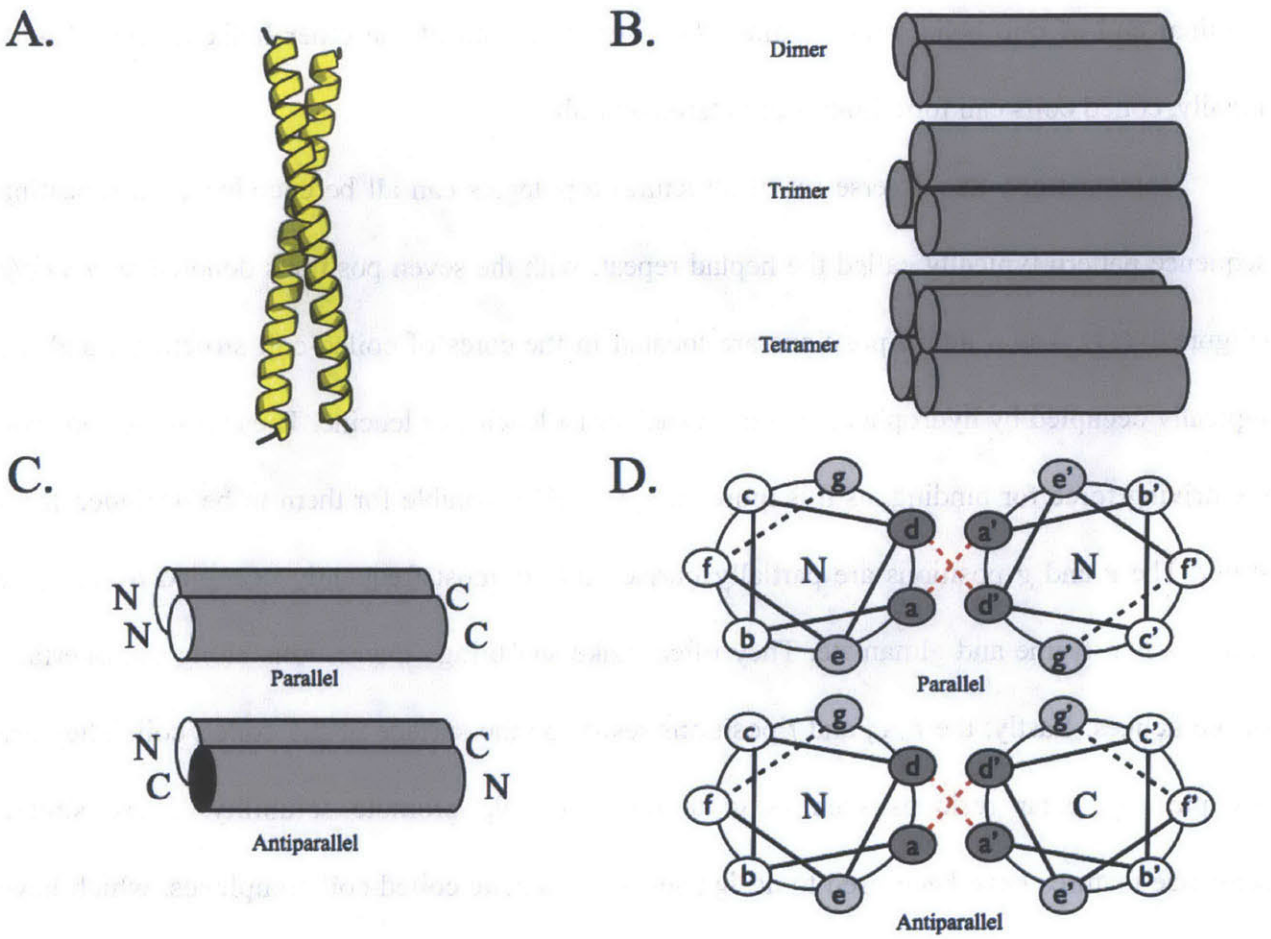


Figure 1-4. Coiled coil structural diversity. (A) Cartoon representation of the side view of a dimeric coiled-coil backbone. The cartoon shows that the helices wrap around each other along what is known as the superhelical axis. (B) Helices are represented as cylinders. The three most common oligomerization states for coiled coils are shown, i.e. dimer (top), trimer (middle), and tetramer (bottom). (C) The helices can have two orientations with respect to each other. They can either be parallel, with the N-terminus adjacent to the N-terminus of its partner strand, or antiparallel, where the N-terminus is adjacent to the C-terminus of its partner strand. (D) These cartoons represent a cross section taken while looking down the superhelical axis of a dimeric coiled coil. Coiled coils are made of a repeating unit known as the heptad repeat, denoted *a-g*. Heptad positions are colored based on the amount of exposure to solvent, with dark grey representing the most buried positions. Depending on the orientation of the helices (parallel vs. antiparallel), different inter-helical heptad positions make direct contacts. For example, in the parallel state *a* and *a'*, as well as *d* and *d'*, are adjacent to each other. In the antiparallel state *a* to *d'* are adjacent to each other.

1.3 Coiled coils as molecular reagents

Coiled coils have been applied to the manipulation of cell signaling pathways, the construction of molecular electronics, and the design of polyhedra, just to name a few applications (Bashor et al., 2008; Shlizerman et al., 2010; Gradišar et al., 2013). A list of several applications can be found in Table 1-1. There are many reasons why coiled coils have been widely used. They provide precise control over spatial arrangement. For instance, due to the linear structure of a coiled coil it is possible to reduce or increase the length of the interaction domain by removing or adding amino acids. Additionally, because coiled coils adopt different oligomerization states, coiled coils can bring a small or large number of molecules together. The surface positions of coiled coils can often play a minimal role in coiled-coil assembly (Mason et al., 2009). As a result, these positions can be mutated to non-natural amino acids, or altered in other ways, to confer novel functions on to the coiled coil (Li et al., 2008; Grigoryan et al., 2011). Surface positions even provide a way to tune the stability of an interaction (Dahiyat et al., 1997). Lastly, one of the major benefits of coiled-coil interactions is that with a small number of amino acids, typically in the range of 35-42 amino acids, it is possible to encode complex PPI networks (Reinke et al., 2010). This efficient encoding of information should facilitate the transporting of these PPI networks into various synthetic biology and protein engineering applications.

Table 1-1. Applications of coiled coils.

COILED-COIL TOPOLOGY USED	APPLICATION	REF.
Parallel homodimer fused to zinc fingers	Used in the design of an artificial transcription factor.	(Wolfe et al., 2003)
Parallel heterodimer	Used to modulate the signaling dynamics of the yeast MAP kinase pathway.	(Bashor et al., 2008)
Parallel homotrimer	Used in the design of hydrogels.	(Jing et al., 2008)
Antiparallel heterodimer and parallel heterodimers	Used to alter the surface electronic properties (Work function) of gold.	(Shlizerman et al., 2010)
Antiparallel homohexamer	Used to assemble gold around a single-walled carbon nanotube.	(Grigoryan et al., 2011)
Antiparallel heterodimer	Used to direct drug delivery to cancer cells in mice.	(Wu et al., 2011)
Parallel heterodimer	Used to create fibers and various nanostructures.	(Boyle et al., 2012)
Antiparallel homodimer fused to p53	Used to design a protein-based filter that had nanometer sized pores.	(Doles et al., 2012)
Parallel homotrimer	Designed to adopt a pre-determined symmetry upon crystallization	(Lanci et al., 2012)
Parallel heterodimer and a parallel homotrimer	Used to assemble into cages that were ~80 nm in diameter.	(Fletcher et al., 2013)
Antiparallel homodimers and parallel heterodimers	Used in the de novo design of a protein-based tetrahedron	(Gradišar et al., 2013)

1.4 Synthetic coiled coils and design rules

Many of the applications involving coiled-coil parts require they adopt a unique structural topology to function properly (Bashor et al., 2008; Shlizerman et al., 2010; Grigoryan et al., 2011; Gradišar 2013). For instance, Shlizerman et al. described how antiparallel dimers and parallel dimers alter the surface electronic properties of gold in different ways (Shlizerman et al., 2010). In another example, Grigoryan et al. sought to coat single-walled carbon nanotubes

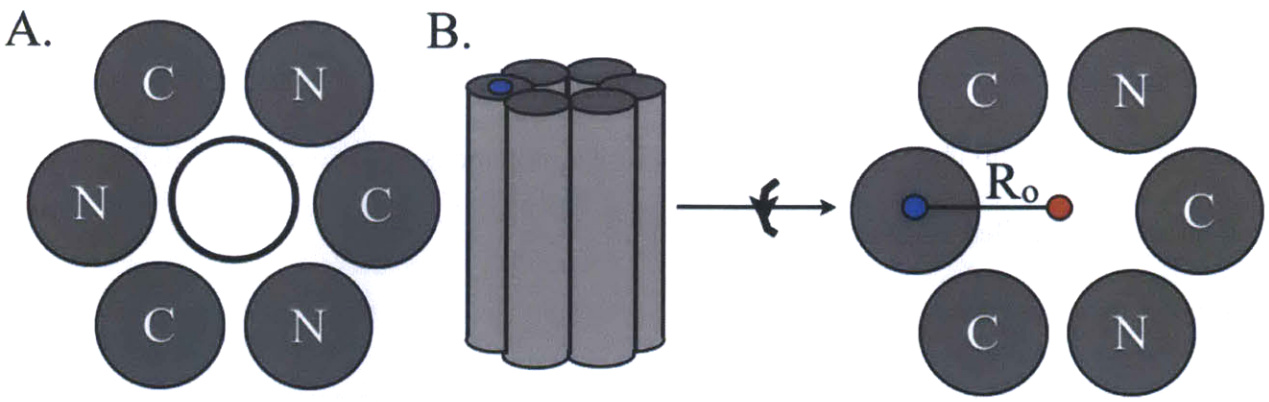


Figure 1-5. Coating SWNT with coiled coils. (A) Schematic looking down the axis of a SWNT (hollow black circle). An antiparallel hexameric coiled coil, shown as a collection of gray circles, wraps around the SWNT. (B) Shows the superhelical radius, Ro, proposed by Crick (Crick, 1953), for a hexameric coiled coil. It is measured from the point at the center of one helix, blue circle, to a point that is equidistant to the center of all helices in the complex, red circle. Ro has been measured for coiled coils of different oligomerization states (Grigoryan & DeGrado, 2011).

(SWNTs) with coiled coils (Figure 1-5A) (Grigoryan et al., 2011). When coiled coils form higher-order oligomers, they can form pores, and if these pores are large enough, the cylindrical SWNT can fit into it. The sizes of these pores depend on the oligomerization state of the coiled coil (Figure 1-5B) (Grigoryan & DeGrado, 2011). The authors determined that a coiled coil larger then a tetramer was needed to fully wrap around the SWNT, thus the authors designed an antiparallel hexamer (Figure 1-5A). Additionally, adopting a unique topology was important to the design of artificial transcription factors (Wolfe et al., 2003). The artificial transcription factors were designed by combining dimeric coiled coils with DNA binding zinc-finger domains. It was necessary that the zinc fingers remain on the same terminal ends of the coiled coil in order to bind specific DNA binding sites, thus the parallel dimer topology was crucial to function.

To design coiled coils that adopt unique structural topologies in order to carry out their function, it is necessary to understand the design rules that promote stability and specificity in a particular topology (Table 1-2). A description of the known design rules that have been

elucidated by the design of synthetic coiled coils, along with a list of some of the available synthetic coiled-coils parts, will help future endeavors hoping to exploit coiled coils as molecular reagents. More importantly it well help highlight the limitations of the existing coiled-coil parts and direct future coiled-coil studies.

Table 1-2. Known coiled-coil design rules.

Physical Property	Sequence-to-structure relationships (Design Rules)
Binding Affinity	• A decrease in chain length can weaken binding affinity. • Addition of residues with higher helical propensity at surface heptad positions *b*, *c*, and *f* can strengthen binding affinity.
Oligomerization	• Polar or charged residues at heptad positions *a* and *d* can favor the formation of dimers. • Beta-branch residues at heptad positions *a* and *d* can favor the formation of parallel trimers. • Leucine residues at heptad position *a* and beta-branch residues at heptad position *d* can favor the formation of parallel tetramers.
Orientation	• Placement of salt-bridge interactions (charge patterning) can preferentially stabilize one orientation over the other (parallel vs. antiparallel). • Asparagine residue pairs, that mismatch in the undesired orientation, will destabilize this state. • Beta-branched residues at heptad position *d* in the parallel dimer state can destabilize the parallel dimer orientation.
Interaction Partner	• Placement of salt-bridge interactions (charge patterning) can preferentially stabilize the desired interaction partner over an undesired partner. • Asparagine residue pairs, that mismatch when bound to an undesired partner, will destabilize this states.

The first few synthetic coiled coils were based on the consensus heptad sequence of tropomyosin, (KLESLES at *gabcdef* heptad positions respectively) and are thought to adopt a parallel homodimer configuration (St. Pierre & Hodges, 1976; Hodges et al., 1981; Lau et al., 1984; Zhou, 1992). These studies helped to reveal many of the key design rules of coiled coils. For instance, Zhou et al. studied how hydrophobicity in the core can stabilize the coiled coil (Zhou et al., 1992). This was done by taking a designed sequence with leucine residues at all

core *a* and *d* positions (except for one *a* site containing a cysteine) and making several new constructs that had a pair of leucine residues at the *a* and *d* sites mutated to a pair of alanine residues. The authors observed that each construct with an alanine pair was less stable then the construct with an all leucine core. Additionally, the authors observed that alanine mutations closer to the center of the coiled coil were more destabilizing than alanine mutations near the termini. Lau et al. noted how increasing chain length increased stability in their synthetic peptide (Lau et al., 1984). They worked with peptides that were 8, 15, 22, 29, and 36 residues long. Peptides of less then 29 residues were shown not to dimerize. Of the peptides that formed dimers, the 36-residue peptide was more stable than the one with only 29 residues. Thomas et al. further verified this observation by using chain length to modulate the dissociation constants of a set of parallel heterodimers (Thomas et al., 2013). The measured dissociation constants span the micromolar to sub-nanomolar range.

Harbury et al. did mutational studies involving the core residues of the dimerization domain of the GCN4 transcription factor, GCN4-p1 (Harbury et al, 1993). The native GCN4 coiled coil forms a parallel homodimer, the structure of which was solved by O'Shea at al. (O'Shea et al., 1991). These studies helped to reveal how side-chain packing of beta-branched residues like isoleucine in the core influence the preferred oligomerization state of a coiled coil through steric clashes (Figure 1-6). The sequences in this study were made up of various combinations of leucine, isoleucine, and valine in the core. They were designed in such a way that the *a* positions and *d* positions would be occupied by just one of these three types of residues. Through analytical ultracentrifugation and size-exclusion experiments, the authors learned that construct p-IL, with isoleucine at *a* and leucine at *d*, populated only the parallel

homodimeric state. The construct, p-II, with isoleucine both at *a* and at *d*, populated the parallel homotrimer state. The construct, p-LI, with leucine at *a* and isoleucine at *d*, populated the parallel homotetramer state.

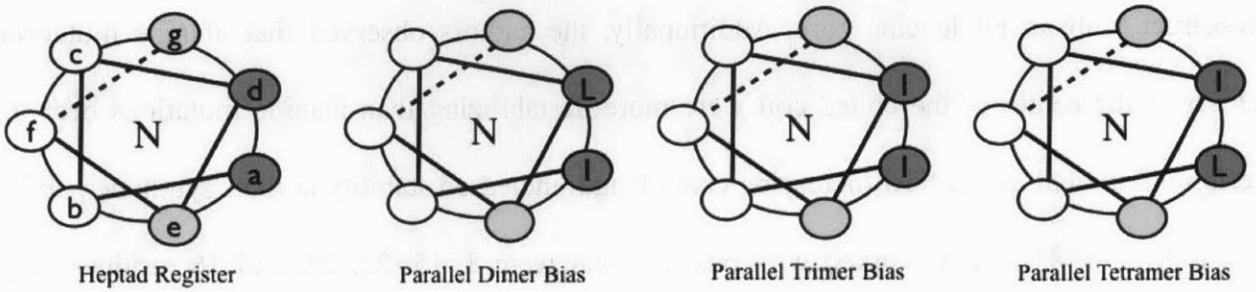


Figure 1-6. Rules that govern oligomerization state bias of parallel topologies. The left-most panel shows the heptad register mapped onto a helix. The remaining panels show how the differential placement of isoleucine at heptad positions *a* and *d* can alter the oligomerization state bias according to Harbury et al. (Harbury et al., 1993).

Fletcher et al. attempted to further validate Harbury's design rules by engineering another set of parallel homo- dimers, trimers, and tetramers with different *e*, *g*, *b*, *c*, and *f* residues from Harbury's designs (Fletcher et al., 2012). When isoleucine was placed at all *a* and all *d* positions, or when leucine was placed at all *a* positions and isoleucine at all *d* positions, the designs formed the expected oligomerization states of a trimer and tetramer, respectively. However, when isoleucine was used at all *a* positions, and leucine was used at all *d* positions, a trimer instead of a dimer was observed. The authors then looked at native parallel homo- dimeric and trimeric coiled-coil sequences, and observed that isoleucine at *a* was not significantly enriched in dimers over trimers. Trimer sequences however are enriched at isoleucine at *d*. In order to destabilize the trimer state, and recover the dimer they intended to design, the authors mutated one isoleucine in the core to an asparagine. Analytical ultracentrifugation studies, and X-ray crystallography confirmed that the new asparagine-containing sequence formed a dimer. Additionally, the authors noted that the original p-IL design formed a dimer and or a trimer in their analytical

ultracentrifugation experiments depending on the concentration used in the experiment, thus reconciling their work with that of Harbury et al. The work of Fletcher et al. is also consistent with a tetramer designed by Betz and DeGrado (Betz & DeGrado, 1996). Betz and DeGrado designed a series of antiparallel homotetramers. One of their designed sequences contained valine at all *a* positions and leucine at all *d* positions. This sequence formed a homotetramer according to analytical ultracentrifugation. The sequence was designed such that *b* and *e* positions, as well as the *g* and *c* positions on different helices could only interact in an antiparallel tetramer state. This provided another example that beta-branched residues in the core *a* position aren't sufficient to dictate dimerization.

O'Shea et al. studied the effect that electrostatics at the *e* and *g* positions have on coiled-coil formation (O'Shea et al., 1993). Two designed peptides, Acid-p1 and Base-p1, form a parallel heterodimer. The authors used charged residues at the *e* and *g* position to promote the heterodimer state over the two-homodimer state. Additionally, each of Acid-p1 and Base-p1 have an asparagine at an *a* position of the coiled coil that destabilizes the antiparallel state. The authors observed that charged residues at *e* and *g* positions had more of an impact on destabilizing the homodimer state than on stabilizing the heterodimer state. They concluded that in this system electrostatic repulsion is more important to interaction specificity than electrostatic attraction.

Monera et al. designed a model antiparallel heterodimeric coiled coil, to facilitate learning antiparallel dimer design rules (Monera et al., 1993). These authors used almost the same sequence for their design as the Zhou et al. study based on the consensus sequence of tropomyosin, as mentioned above. Disulphide-exchange studies revealed that this sequence

preferred the parallel configuration. The authors subsequently re-designed the charges at the *e* and *g* positions to be attractive in the antiparallel state and repulsive in the parallel state, which improved the preference of the antiparallel sequence. Oakley et al. designed another antiparallel heterodimer (Oakley & Kim 1998). The authors made a variant of Acid-p1/Base-p1. They shifted the asparagine in both the Acid and Base sequence such that it could only form an interchain hydrogen bond in the antiparallel state. This single asparagine stabilized the antiparallel state by 2.3 kcal/mol relative to the parallel state. The new sequence was named Acid-a1/Base-a1. McClain et al. went on to make additional mutations of the Acid-a1/Base-a1 construct (McClain et al., 2001). This variant was named Acid-Kg/Base-Eg. The electrostatics were altered so that stabilizing interactions were formed only in the antiparallel state, with many predicted repulsions in the parallel state. The combined use of charge patterning and an asparagine-asparagine interaction stabilized the antiparallel state by at least 4.4 kcal/mol relative to the parallel state.

Gurnon et al. and Pagel et al. each designed antiparallel coiled-coil homodimers (Gurnon et al., 2003; Pagel et al., 2005). Each design used charged residues at the *e* and *g* position to destabilize the parallel state, as was done for the anti-parallel heterodimers discussed above. However, Gurnon et al. used two more design rules to stabilize the antiparallel state over off-target states. The authors used a charged residue in a core *d* position to destabilize higher-order states (McClain et al., 2002). It is thought that *d* positions are completely buried in higher-order states, preventing the charge residues from being properly solvated. However, the *d* sites in dimeric coiled coils are more easily accessible by solvent. This may allow charged residue at *d* sites to interact with water, thus biasing the formation of dimers. Gurnon et al. also placed an

isoleucine in the core *d* position such that it would sterically clash with itself in the parallel state, as mentioned earlier.

Though many design rules have been elucidated that preferentially stabilize one topology over another (Table 1-2, Figure 1-7), a model that can quantitatively compare and appropriately weight the influence of these different interactions does not exist. And it is still unclear whether

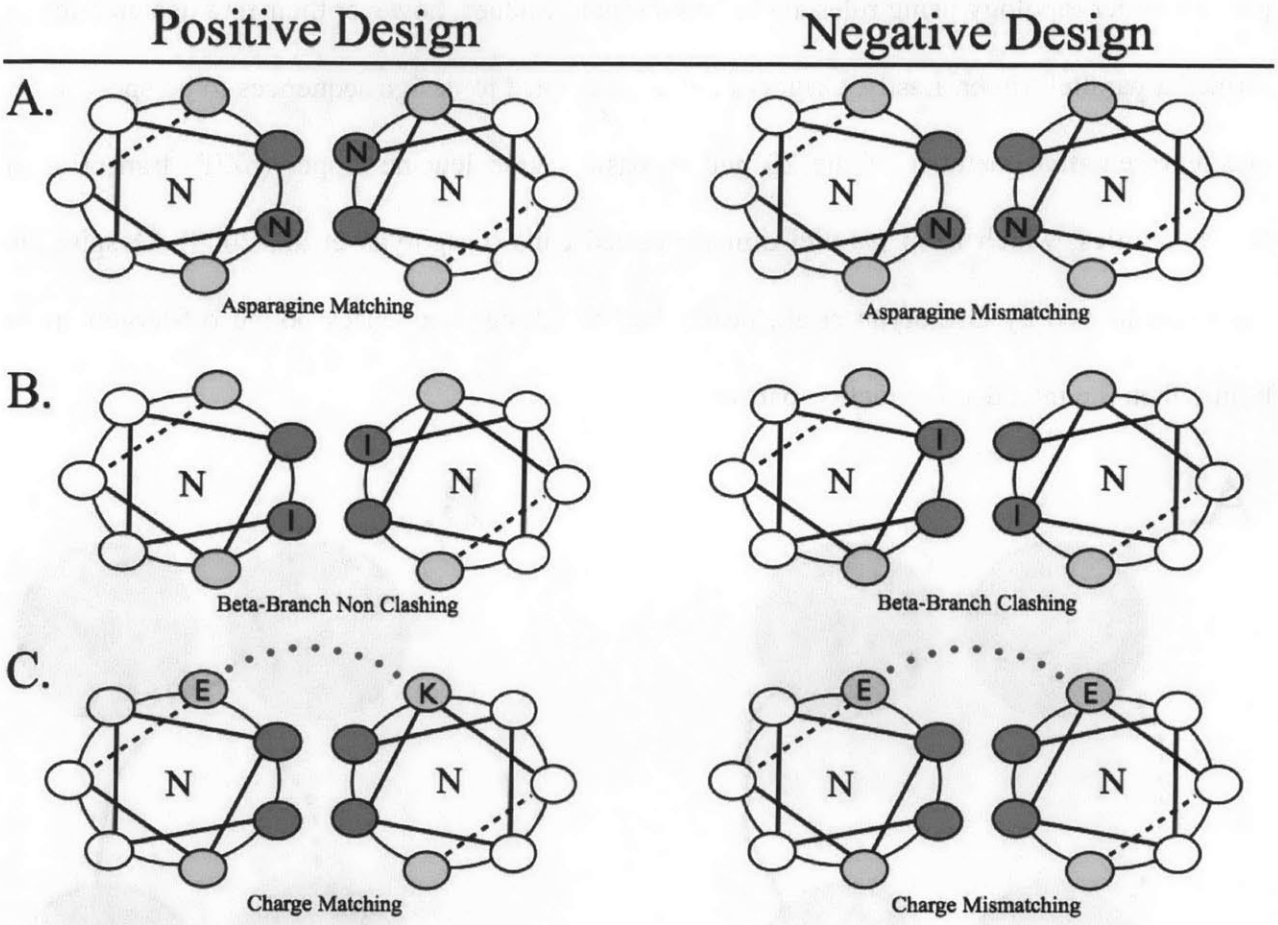


Figure 1-7. Negative design rules shown on parallel dimers. Three design rules that can preferentially stabilize a target state over off-target states are shown with helical wheel representations of parallel dimeric coiled coils. The left column shows examples of when these interactions stabilize a state. The right column shows examples of when these interactions destabilize a state. (A) Shows asparagine matching and mismatching. (B) Shows non-clashing and clashing beta-branch residues. (C) Shows both attractive (blue) and repulsive (red) electrostatic interactions.

all design rules have been uncovered, as several attempts to design coiled coils have met with complications . For example, Hill and DeGrado attempted to design a coiled coil made up of a helix-loop-helix that dimerizes into an up-down-up-down helical bundle (Figure 1-8A) (Hill & DeGrado, 1998). The construct the authors designed however obtained a bisecting U motif (Figure 1-8B). Another example by Fletcher et al., as mentioned earlier, attempted to design a parallel dimer topology using rules about beta-branch residues, however their first design attempt formed a parallel trimer. Lastly, Grigoryan et al. attempted to design sequences to be specific for one representative member of the 20 human basic region leucine zipper (bZIP) transcription factor families, which form parallel dimeric coiled coils (Grigoryan et al., 2009). Despite the success achieved by Grigoryan et al., nearly half the design sequences bound off-targets more tightly then the intended interaction partner.

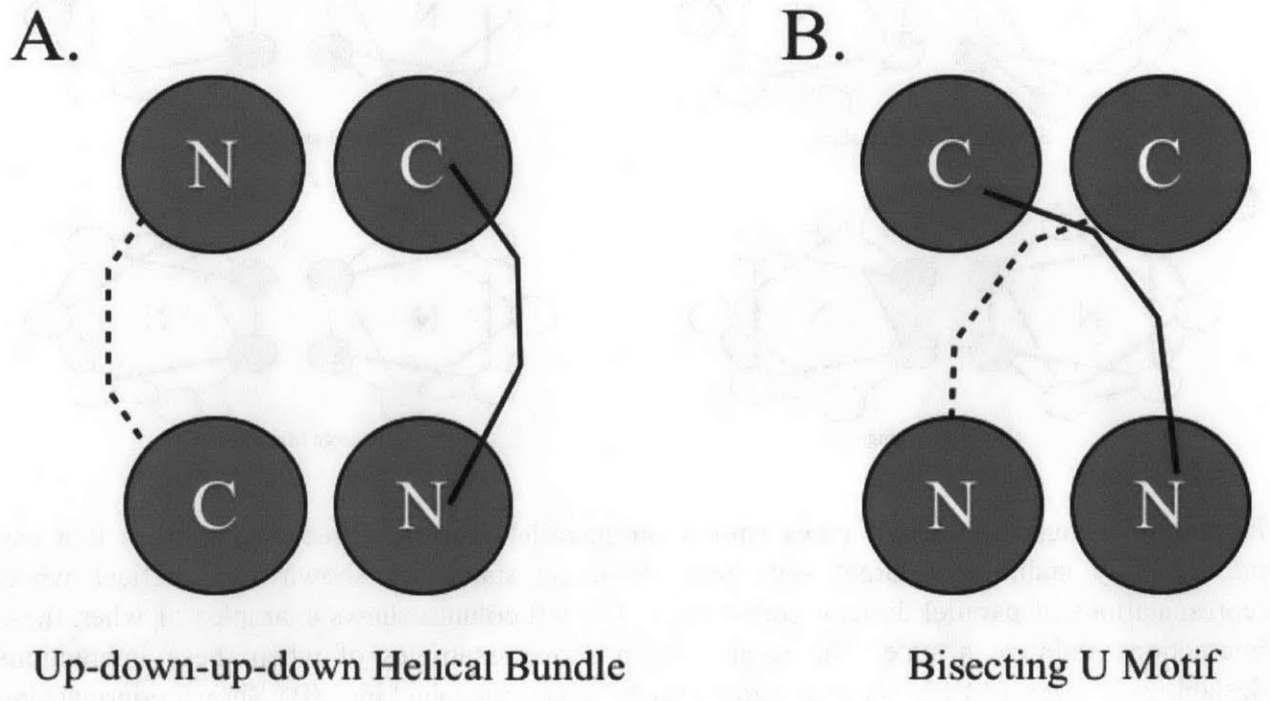


Figure 1-8. Schematics of two four-helix bundle topologies. Gray circles represent helices, with terminal ends labeled N or C. The loop connectivities are shown with dashed lines for loops

into the page, and solid lines for loops out of the page. (A) Shows an up-down-up-down helical bundle. (B) Shows a bisecting U motif.

1.5 Coiled-coil databases for synthetic biology

As the sequence to structure relationship of coiled coils has become more established, several groups have begun to design libraries of coiled-coil reagents containing various types of interaction networks, to be used for synthetic biology and protein engineering. In a noteworthy study by Reinke et al., all interactions among 22 peptides that form 27 synthetic heterodimers were determined using a protein microarray (Reinke et al., 2010). These synthetic coiled-coil peptides were referred to as SYNZIPs. Thompson et al. further characterized the biophysical properties of SYNZIPs by measuring their helix orientation bias, oligomerization states, and affinities, as well as by evaluating whether the SYNZIPs oligomerize in cells to down-regulate the expression of a reporter gene (Thompson et al., 2012).

As large data sets of coiled-coil parts emerge, several groups have created depositories to aid synthetic biologists and protein engineers in search of existing coiled-coil parts. For instance, specification sheets listing the properties of the SYNZIP sequences can be found at the SYNZIP website [http:// keatingweb.mit.edu/SYNZIP/]. Additionally, information on the constructs engineered by the Woolfson group has been deposited in the *Pcomp* database [http:// coiledcoils.chm.bris.ac.uk/pcomp/]. It should be noted that both these databases are dominated by parallel dimers. 96% of the SYNZIP sequences form parallel heterodimers. ~63% of the

Pcomp database form parallel homo- or heterodimers. Designing coiled coils that are not parallel dimers will thus expand the coiled-coil toolkit in a meaningful way.

1.6 Design of orthogonal coiled-coil interactions

As the design of sequences that can adopt a single coiled-coil topology in solution has become more standard, several groups have begun to attempt the design of coiled-coil interaction networks. Of particular interest to synthetic biologists and protein engineers are orthogonal PPIs, i.e., sets of PPIs that do not interact with each other (Kapp et al., 2012; Gradišar et al, 2013). Gradišar and Jerala considered hydrophobic packing, buried polar residues, and electrostatics in the *de novo* design of four orthogonal parallel heterodimers that are four heptads long (Gradišar & Jerala, 2010). All 36 possible pairs of the eight sequences designed to fold into four parallel heterodimers were measured using circular dichroism (CD). Sequences designed not to interact often gave CD spectra characteristic of a random coil. The four pairs of sequences designed to interact gave CD spectra indicative of α –helical structure, providing good evidence of success. Bromley et al. carried out a similar successful study. Their orthogonal designs were a set of parallel heterodimers as well, but their designs were three heptads long (Bromley et al., 2009).

Despite the successful design of these orthogonal parallel heterodimer sets, Gradišar et al. have argued that a current limitation to using coiled coils as molecular reagents is still the number of orthogonal coiled-coil pairs that are available (Gradišar et al., 2013). The authors argued this based on a strategy they put forward to design polyhedra of arbitrary shape from concatenating orthogonal coiled-coil dimers onto a single chain. The coiled-coils are intended to

dimerize on the chain, forming the edges of the polyhedra. Using graph theory they determined that the design of most polyhedrons require both parallel and antiparallel dimers. Yet, not only are the current sets of orthogonal coiled coils small, but none of the synthetic coiled coil sets contain orthogonal antiparallel dimers. This highlights the importance of expanding the number of orthogonal antiparallel coiled-coil dimers for constructing protein polyhedra of arbitrary shape.

1.7 Summary of thesis content

This thesis describes the computational design and experimental characterization of several orthogonal antiparallel homodimeric coiled coils. The multi-state computational design framework known as CLASSY was used for this design problem. One of the benefits of the CLASSY framework is that it can design multiple orthogonal PPIs simultaneously. This is in contrast to the second-site suppressor strategy developed by Kortemme, which only designs one orthogonal pair at a time (Kortemme et al., 2004). Antiparallel homodimers are underrepresented in the toolkit of coiled-coil parts, and as described above, modular-orthogonal parts have been shown to be in great demand by the synthetic biology and protein engineering communities.

1.8 References

Ashenberg, O., Rozen-Gagnon, K., Laub, M. T., & Keating, A. E. (2011) Determinants of homodimerization specificity in histidine kinases. *Journal of Molecular Biology*, *413* (1), 222-235.

Bashor, C. J., Helman, N. C., Yan, S., & Lim, W. A. (2008) Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. *Science, 319*, 1539-1543.

Betz, S. F., & DeGrado, W. F. (1996) Controlling topology and native-like behavior of de novo-designed peptides: Design and characterization of antiparallel four-stranded coiled coils. *Biochemistry, 35*, 6955-6962.

Boyle, A. L., Bromley, E. H. C., Bartlett, G. J., Sessions, R. B., Sharp, T. H., Williams, C. L., Curmi, P. M. G., Forde, N. R., Linke, H., & Woolfson, D. N. (2012) Squaring the circle in peptide assembly: From fibers to discrete nanostructures by de novo design. *Journal of the American Chemical Society, 134*, 15457-15467.

Bromley, E. H. C., Sessions, R. B., Thomson, A. R., & Woolfson, D. N. (2009) Designed α-helical tectons for constructing multicomponent synthetic biological systems. *Journal of the American Chemical Society, 131* (3) 928-930.

Crick, F. H. C. (1953) The fourier transform of a coiled-coil. *Acta Crystallographica, 6*, 685-689.

Dahiyat, B. I., Gordon, D. B., & Mayo, S. L. (1997) Automated design of the surface positions of protein helices. *Protein Science, 6*, 1333-1337.

Doles, T., Bozic, S., Gradisar, H., & Jerala, R. (2012) Functional self-assembling polypeptide bionanomaterials. *Biochemical Society Transactions, 40* (4), 629-634.

Dutta, S., Gullá, S., Chen, T. S., Fire, E., Grant, R. A., & Keating, A. E. (2010) Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL. *Journal of Molecular Biology, 398* (5), 747-762.

Fletcher, J. M., Boyle, A. L., Bruning, M., Bartlett, G. J., Vincent, T. L., Zacci, N. R., Armstrong, C. T., Bromley, E. H. C., Booth, P. J., Brady, R. L., Thomson, A. R., & Woolfson, D. N. (2012) A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology. *ACS Synthetic Biology, 1*, 240–250.

Garma, L., Mukherjee, S., Mitra, P., & Zhang, Y. How many protein-protein interactions types exist in nature?. *PLoS ONE, 7* (6), e38913 1-9.

Gradišar, H., Božič, S., Doles, T., Vengust, D., Hafner-Bratkovič, I., Mertelj, A., Webb, B., Šali, A., Klavžar, S., & Jerala, R., (2013) Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nature Chemical Biology, 9* (6), 362-366.

Gradišar, H., & Jerala, R. (2010) De novo design of orthogonal peptide pairs forming parallel coiled-coil heterodimers. *Journal of Peptide Science*, *17* (2), 100-106.

Grigoryan, G., & DeGrado, W. F. (2011) Probing designability via a generalized model of helical bundle geometry. *Journal of Molecular Biology*, *405* (4), 1079-1100.

Grigoryan, G., Kim, Y. H., Acharya, R., Axelrod, K., Jain, R. M., Willis, L., Drndic, M., Kikkawa, J. M., & DeGrado, W. F. (2011) Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science*, *332* (6033), 1071-1076.

Grigoryan, G., Reinke, A. W., & Keating, A. E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature*, *458*, 859–864.

Gurnon, D. G., Whitaker, J. A., & Oakley, M. G. (2003) Design and characterization of a homodimeric antiparallel coiled coil. *Journal of the American Chemical Society*, *125*, 7518-7519.

Harbury, P. B., Zhang, T., Kim, P. S., & Alber, T. (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science*, *262* (5138), 1401-1407.

Hendrickson, T. L., de Cre´cy-Lagard, V., & Schimmel, P.(2004) Incorporation of nonatural amino acids into proteins. *Annual Review of Biochemistry*, *73* (1), 147-176.

Hill, R. B., & Degrado, W. F. (1998) Solution structure of α2D, a nativelike de novo designed protein. *Journal of the American Chemical Society*, *120,* 1138-1145.

Hodges, R. S., Saund, A. K., Chong, P. C. S., St.-Pierre, S. A., & Reid, R. E. (1981) Synthetic model for two-stranded alpha-helical coiled-coils. *Journal of Biological Chemistry*, *256* (3), 1514-1224.

Jing, P., Rudra, J. S., Herr, A. B., & Collier, J. H. (2008) Self-assembling peptide-polymer hydrogels designed from the coiled coil region of fibrin. *Biomacromolecules*, *9* (9), 2438-2446.

Kapp, G. T., Liu, S., Stein, A., Wong, D. T., Reményic, A., Yeh, B. J., Fraser, J. S., Taunton, J., Lim, W. A., & Kortemme, T. (2012) Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. *Proceedings of the National Academy of Science*, *109* (14), 5277–5282.

Kellogg, E. H., Leaver-Fay, A., & Baker, D. (2010) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics*, *79* (3), 830-838.

Khalil, A. S., & Collins, J. J. (2010) Synthetic biology: Applications come of age. *Nature Reviews Genetics, 11* (5) 367-379.

Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L., & Baker, D. (2004) Computational redesign of protein-protein interaction specificity. *Nature Structural Biology and Molecular Biology, 11* (4), 371-379.

Lai, Y. T, Cascio, D., & Yeates, T. O. (2012) Structure of a 16-nm cage designed by using protein oligomers. *Science, 336*, 1129.

Lanci, C. J., MacDermaid, C. M., Kang, S., Acharya, R., North, B., Yang, X., Qiu, X. J., DeGrado, W. F., & Saven, J., G. (2012) Computational design of a protein crystal. *Proceedings of the National Academy of Science, 109*, 7304–7309.

Laub, M. T., & Goulian, M. (2007) Specificity in two-component signal transduction Pathways. *Annual Review of Genetics, 41* (1), 121-145.

Li, Y., Kaur, H., & Oakley, M.G. (2008) Probing the recognition properties of the antiparallel coiled coil motif from PKN by protein grafting. *Biochemistry, 47*, 13564-13572.

Liu, J., Zheng, Q., Deng, Y., Cheng, C. S., Kallenbach, N.R., & Lu, M. (2006) A seven-helix coiled coil. *Proceedings of the National Academy of Science, 103*, 15457-15462.

Liu, Y., & Eisenberg, D. (2002) 3D domain swapping: As domains continue to swap. *Protein Science*, 11 (6), 1285-1299.

Mason, J.M., Hagemann, U. B., & Arndt, K.M. (2009) Role of hydrophobic and electrostatic interactions in coiled coil stability and specificity. *Biochemistry, 48* (43), 10380-10388.

McClain, D. L., Gurnon, D. G., & Oakley, M. G. (2002) Importance of potential interhelical salt-bridges involving interior residues for coiled-coil stability and quaternary structure. *Journal of Molecular Biology, 324* (2), 257-270.

McClain, D. L., Woods, H. L., & Oakley, M. G. (2001) Design and characterization of a heterodimeric coiled coil that forms exclusively with an antiparallel relative helix orientation. *Journal of the American Chemical Society, 123* pp. 3151-3152.

Monera, O. D., Zhou, N. E., Kay, C. M., & Hodges, R. S. (1993) Comparison of antiparallel and parallel two-stranded alpha-helical coiled-coils. *Journal of Biological Chemistry, 268* (26), 19218-19227.

Oakley, M. G. & Kim, P.S (1998) A buried polar interaction can direct the relative orientation of helices in a coiled coil. *Biochemistry*, 37, 12603-12610.

O'Shea, E. K., Klemm, J. D., Kim, P. S., & Alber, T. (1991) X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science, 254* (5031), 539-544.

O'Shea, E. K., Lumb, K. J., & Kim, P.S. (1993) Peptide 'Velcro*': Design of a heterodimeric coiled coil. *Current Biology, 3*, 658-667.

Padilla, J. E., Colovos, C., & Yeates, T. O. (2001) Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proceedings of the National Academy of Science*, 98 (5), 2217–2221.

Pagel, K., Seeger, K., Seiwert, B., Villa, A., Mark, A. E., Berger, S., & Koksh, B. (2005) Advanced approaches for the characterization of a de novo designed antiparallel coiled coil peptide. *Organic and Biomolecular Chemistry, 3* (7), 1189-1194.

Pauling, L., & Corey, R. B. (1953). Compound helical configurations of polypeptide chains: Structure of proteins of the alpha-keratin type. *Nature, 171*, 59-61.

Purnick, P. E., & Weiss, R. (2009) The second wave of synthetic biology: From modules to systems *Nature Reviews Molecular Cell Biology, 10*, 410–422.

Reinke, A.W., Baek, J., Ashenberg, O., & Keating, A.E. (2013) Networks of bZIP protein-protein interactions diversified over a billion years of evolution. *Science, 340* (6133), 730-734.

Reinke, A. W., Grant, R. A., & Keating, A. E. (2010) A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. *Journal of the American Chemical Society, 132*, 6025-6031.

Sammond, D. W., Eletr, Z. M., Purbeck, C., & Kuhlman, B. (2010) Computational design of second-site suppressor mutations at protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics, 78*, 1055-1065.

Shlizerman, C., Atanassov, A., Berkovich, I., Ashkenasy, G., & Ashkenasy, N. (2010) De novo designed coiled-coil proteins with variable conformations as components of molecular electronic devices. *Journal of the American Chemical Society, 132*, 5070–5076.

St-Pierre, S. A., & Hodges, R. S. (1976) A sequential polyheptapeptide as a model for the double-stranded alpha-helical coiled-coil structure of tropomyosin. *Biochemical and Biophysical Research Communications, 72* (2), 581-588.

Tamsir, A., Tabor, J. J., & Voigt, C. A. (2011) Robust multicellular computing using genetically encoded NOR gates and chemical 'wires', *Nature, 469(7329)*, 212-215.

Temme, K., Hill, R., Segall-Shapiro, T. H., Moser, F., & Voigt, C. A., (2012) Modular control of multiple pathways using engineered orthogonal T7 polymerases. *Nucleic Acids Research, 40* (17), 8773-8781.

Thompson, K. E., Bashor, C. J., Lim, W. A., & Keating, A. E. (2012) SYNZIP protein interaction toolbox: In vitro and in vivo specifications of heterospecific coiled-coil interaction domains. *ACS Synthetic Biology, 1* (4), 118-129.

Thomas, F., Boyle, A. L., Burton, A. J., & Woolfson D. N. (2013) A set of de novo designed parallel heterodimeric coiled coils with quantified dissociation constants in the micromolar to sub-nanomolar regime. *Journal of the American Chemical Society, 135*, 5161−5166.

Tsai, S. L., DaSilva, N. A., & Chen, W. (2013) Functional display of complex cellulosomes on the yeast surface via adaptive assembly. *ACS Synthetic Biology, 2*, 14−21.

Wolf, E., Kim, P.S., & Berger, B., (1997) MultiCoil: A program for predicting two- and three-stranded coiled coils. *Protein Science, 6*, 1179-1189.

Wolfe, S. A., Grant, R. A., & Pabo, C. O. (2003) Structure of a designed dimeric zinc finger protein bound to DNA. *Biochemistry, 42*, 13401−13409.

Woolfson, D. N. (2005) The design of coiled-coil structures and assemblies. *Advances in Protein Chemistry, 70*, 79-112.

Wu, K., Yang, J., Liu, J., & Kopeček, J. (2011) Coiled-coil based drug-free macromolecular therapeutics: In vivo efficacy. *Journal of Controlled Release, 157* (1), 126-131.

Zarrinpar, A., Park, S. H., & Lim, W. A. (2003) Optimization of specificity in a cellular protein interaction network by negative selection. *Nature, 426*, 676-680.

Zhou, N. E., Kay, C. M., & Hodges, R. S. (1992) Synthetic model proteins. *Journal of Biological Chemistry, 267* (4), 2664-2670.

Chapter 2

Multistate Protein Design Using CLEVER and CLASSY

Reproduced with permission from: Negron, C., & Keating, A. E. (2013) Multistate protein design using CLEVER and CLASSY. *Methods in Enzymology*, *523*, 171-190.

Structure-based protein design is a powerful technique with great potential. Challenges in two areas limit performance: structure scoring and sequence-structure searching. Many of the functions used to describe the relationship between protein sequence and energy are computationally expensive to evaluate, and the spaces that must be searched in protein design are enormous. Here, we describe computational tools that can be used in certain situations to provide enormous accelerations in protein design. Cluster expansion is a technique that maps a complex function of three-dimensional atomic coordinates to a simple function of sequence. This is done by expanding the sequence-energy relation as a linear function of sequence variables, which are fit using training examples. Generating a simpler function speeds up scoring dramatically, relative to all-atom methods, and facilitates the use of new types of search strategies. The application of cluster expansion in protein modeling is new but has shown utility for design problems that require simultaneous consideration of multiple states. In this chapter, we describe cases where cluster expansion can be useful, outline how to generate a cluster-expanded version

of any existing scoring procedure using the software CLEVER, and describe how to apply a cluster-expanded potential to multistate protein design using the CLASSY method.

2.1 Introduction: Accomplishments and limitations of structure-based design

Nearly 30 years ago, Drexler suggested that proteins had the potential to be manipulated to create molecular machines with predefined functions (Drexler, 1981). At that time, a realistic strategy for designing proteins rationally could not be envisioned in detail. Several groups subsequently tackled protein design using computational structure-based methods, culminating in the first fully automated design of a folding protein sequence in 1997 (Dahiyat, Gordon, & Mayo, 1997). Many researchers have now demonstrated impressive accomplishments in this area, including the engineering of protein inhibitors of therapeutically relevant targets (Fleishman et al., 2011), the creation of novel enzymes (Jiang et al., 2008; Rothlisberger et al., 2008), and the assembly of molecular structures that incorporate proteins and other materials (Grigoryan et al., 2011). Modern structure-based design requires two things: an energy function for evaluating candidate sequences and an algorithm that can search the enormous space of sequence-structure possibilities. The requirements for each are linked because the nature of the scoring function dictates what kinds of searches are possible. One of the many limitations of commonly used scoring functions is that they can be costly to calculate. For example, all-atom scoring functions must, at a minimum, evaluate interactions between all pairs of atoms that lie within a prescribed distance. Computing electrostatic interactions can be particularly expensive, depending on the method used. Several techniques have been developed to increase the speed of

energy evaluation. For example, Leaver-Fay et al. implemented a tree data structure in RosettaDesign to eliminate redundant calculation of atom–atom interactions, and this gave a four-fold speedup in the calculation of pairwise energy terms (Leaver-Fay, Kuhlman, & Snoeyink, 2005). Many groups have also worked on speeding up the search component of design. Early recognition that optimal search strategies using algorithms such as dead-end elimination (DEE) are often too slow for real design problems led to widespread adoption of stochastic sampling methods such as Monte Carlo optimization with simulated annealing and genetic algorithms (Havranek & Harbury, 2003; Kuhlman et al., 2003; Voigt, Gordon, & Mayo, 2000). FASTER is a particularly noteworthy stochastic sampling method that was shown to be 100–1000 times faster than DEE (Desmet, Spriet, & Lasters, 2002) and subsequently improved further (Allen & Mayo, 2006). Despite these innovations, design problems involving large proteins, extensive structural sampling, or a large number of states can still be computationally intractable. In this chapter, we focus particularly on challenges posed by multistate design. In a multistate design problem, the designer is concerned not just with a single desired structure or function of interest but with numerous states either desired or undesired. For example, when designing dominant-negative inhibitors, it is important to avoid self-interaction or interactions with other proteins in the cell (Chen, Reinke, & Keating, 2011). Harbury was among the first to treat multistate design, designing topologically specific coiled-coil structures using multiple backbone templates (Harbury, Plecs, Tidor, Alber, & Kim, 1998). Since then, several groups have proposed different approaches (Allen & Mayo, 2010; Bolon, Grant, Baker, & Sauer, 2005; Havranek & Harbury, 2003; Humphris & Kortemme, 2007; Leaver-Fay, Jacak, Stranges, & Kuhlman, 2011; Yanover, Fromer, & Shifman, 2007; Kortemme et al., 2004; Sammond et al.,

2010), and several excellent reviews cover this topic (Erijman, Aizner, & Shifman, 2011; Havranek, 2010; Karanicolas & Kuhlman, 2009). In one example from our laboratory, Grigoryan et al. used multistate design to engineer specific binding partners for representative members of 20 human basic-region leucine zipper (bZIP) transcription factor families (Grigoryan, Reinke, & Keating, 2009). For this purpose, a novel computational solution provided both a dramatic acceleration of energy evaluation and an efficient way to search a complex, multistate design landscape. The approach used a method called cluster expansion (CE) to convert structure-based models of protein energetics into sequence-based models. Grigoryan et al. showed that CE can speed up energy calculations by seven orders of magnitude (Grigoryan et al., 2006). Furthermore, the use of cluster-expanded energy functions allowed a novel application of integer linear programming (ILP) to solve the multistate design problem. With the expectation that this approach can be applied to other problems in protein design, we illustrate the use of the open-source program CLEVER 1.0 to generate CE scoring functions, and discuss how such energy functions can be used in conjunction with ILP in the multistate design method CLASSY.

2.2 Theory

CE is a general technique for deriving a simple linear function that approximates a complex mathematical expression. It involves fitting a set of coefficients to describe a space covered by a user-defined set of relevant variables. CE is used extensively in modeling alloys (de Fontaine, 1994; Sanchez, Ducastelle, & Gratias, 1984), and Zhou et al. demonstrated how to use CE to score the fitness of a protein sequence for a given protein structure (Zhou et al., 2005).

That is, these authors showed how to apply CE when the complex mathematical expression to be described is the energy of a protein sequence adopting a particular three-dimensional structure. In this application, the energy of a protein sequence is written as an expansion around a reference sequence, with energies from specific amino acids and groups of amino acids contributing to the expansion. Two key assumptions are that lower order terms such as interactions between pairs of amino acids at pairs of sites contribute more to the energy than higher order clusters involving many residues, and that, consistent with this, a limited number of residue interaction terms are sufficient to approximate the energy of a protein. These assumptions are well aligned with the physical intuition of structural biologists, who expect short-range pairwise interactions to dominate an energy expression. A brief description of the theory of CE as used for protein energetics is presented here. For a more detailed description, see Grigoryan et al. (2006). Let the variable σ^i index the amino acid at site *i*. If there are M allowed amino acids at site *i* then σ^i can take the values from 0 to (M-1). For a protein of length L amino acids, values of *i* range from 1 to L, and an amino acid sequence is represented by the vector $\sigma = [\sigma^1 \ldots \sigma^L]$. The energy of a sequence, $E(\sigma)$, based on an expansion around a reference sequence, is expressed as:

$$E(\sigma) = J_0 + \sum_{\substack{i \\ \sigma^i \neq 0}} J^i_{\sigma^i} + \sum_{\substack{i \\ \sigma^i \neq 0}} \sum_{\substack{j>i \\ \sigma^j \neq 0}} J^{ij}_{\sigma^i \sigma^j} + \ldots \qquad (1)$$

The J parameters are effective cluster interaction (ECI) values. J_0 is a constant term that reflects the energy of the reference sequence, and the other J values give the contributions of amino acids and amino acid combinations relative to this reference. $J^i_{\sigma^i}$ represents the energetic contribution of a single amino acid σ^i at site *i*, and $J^{ij}_{\sigma^i \sigma^j}$ represents the energetic contribution from a pair of amino acids σ^i and σ^j at sites *i* and *j*, etc. All higher order interactions, up to L-body terms, would

be needed to obtain an exact expansion. The goal in deriving a CE is to find a minimal set of ECI values that provide an accurate estimate of the energy. ECI values are eliminated by truncating the expansion so that it does not include high-order terms, and by testing ECI values that capture low-order terms to confirm that they make important contributions (and, if they do not, removing them as described below). ECI values are determined by fitting, using a training set of sequences for which the correct function value according to some model or experiment is known. In our application, this is the protein energy $E(\sigma)$ computed using an all-atom model. Based on Eq. (1), for any training set with N sequences we can write

$$\boldsymbol{E} = \boldsymbol{XJ} \qquad (2)$$

where $\boldsymbol{E}$ is an N-dimensional column vector of energies for the training sequences, $\boldsymbol{J}$ is a P-dimensional column vector of ECI values, and $\boldsymbol{X}$ is an NP binary matrix indicating which residues and combinations of residues contribute to the energy in each training sequence. The presence/absence of different sets of residues is stored in "cluster functions" (or CFs) that compose matrix $\boldsymbol{X}$. For example, a CF indicating the presence of alanine at site 1 and leucine at site 2 in a protein would evaluate to 1 for a given sequence only if that sequence had that combination of amino acids and to 0 otherwise. Each CF has an associated ECI. For a given set of CFs, the training-set sequences and their energies define matrix $\boldsymbol{X}$ and vector $\boldsymbol{E}$ in Eq. (2). When there are more training sequences than ECI values, the system is over-determined, which allows techniques such as least-squares fitting to be used to find the optimal values for the unknown parameters $\boldsymbol{J}$. The procedure used by CLEVER 1.0 to select CFs/ECIs to be included is described in Figure 2-1 and more details can be found in Hahn et al., (Hahn et al., 2010). At the outset, the user defines a set of candidate CFs (candidate amino acid combinations) likely to

contribute significantly to the total energy. An iterative procedure is then used to determine which of these should be included in the final expansion, using a leave-sequence-out cross-validation procedure to avoid overfitting. After fitting, the accuracy of the expansion can be evaluated by scoring another set of sequences, known as the test set, with both the original energy function and the cluster-expanded version of it.

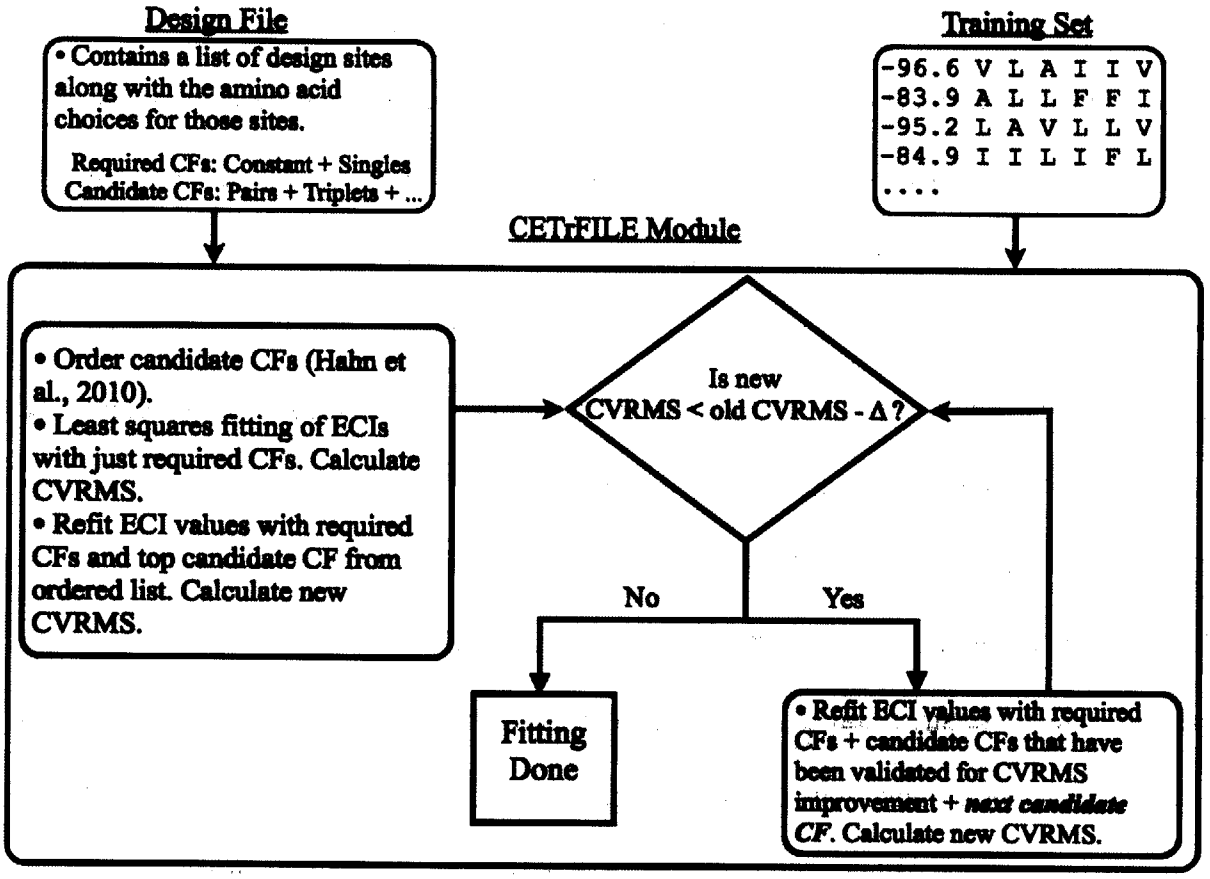


Figure 2-1. Procedure for fitting a cluster expansion. Two inputs are required to train a cluster expansion for protein design using the CLEVER package: the design file (see Figure 2-2) and the training-set file. The design file lists a set of candidate cluster functions (CFs) and defines the sequence space the user is interested in describing. The training-set file provides a set of sequences with associated energies. The fitting procedure fits a subset of the variables listed in the design file to reproduce the training-set data. In our implementation, the constant and point CFs are always included in the fitting process. To avoid overfitting, pair and higher order CFs are incrementally added, following an order that is predetermined at an early stage of the fitting routine. The progress of the fitting is monitored using the cross-validated root mean square

(CVRMS) error. When a new CF is added, all terms are refit, and a new CVRMS is calculated. If the CVRMS score improves by at least Δ, the new CF is accepted and used in the final expansion. If the CVRMS does not improve, the fitting process ends. The goal is to find the smallest number of CFs that must be included to give a good model.

2.3 Benefits offered by cluster expansion in protein modeling and design

There are several reasons a protein modeler or designer might want to develop a cluster-expanded version of their energy function of interest. To understand these, it is necessary to focus on what CE delivers, which is an approximate version of a scoring procedure that is extremely rapid and convenient to evaluate. The cost of obtaining this benefit is the diminished accuracy of the CE function and the time required to develop it. Also, it should be emphasized that in all cases tested so far, the relationships derived between sequence and energy using CE have assumed conservation of the underlying protein backbone structure for all sequences. That is, CE delivers a structure-specific scoring function. Apgar et al. observed good performance when cluster expanding structure-based models that included some treatment of backbone flexibility, but these structural changes were very small (Apgar, Hahn, Grigoryan, & Keating, 2009). CE provides a significant speedup to energy evaluation and thus will be of greatest benefit when the energy evaluation confronted in design is especially challenging. For example, electrostatic energies are often treated in a very crude way in protein design, in order to make the resulting functions expressible as a sum over residue pairs. More accurate functions can be much more costly to compute (Lippow, Wittrup, & Tidor, 2007). CE provides an attractive solution in such cases, and Grigoryan et al. explored the expansion of various scoring methods including a generalized Born treatment of electrostatics (Grigoryan et al., 2006). In another, more extreme

example, CE can be used when protein energies are determined experimentally for a training set of interest. The time and expense of experimental protein characterization means that only a part of sequence space will ever be covered this way. But if sufficient examples are available, it may be possible to train a CE expression that can be used to guide protein design. Hahn et al. presented an example of using experimental data to train a CE for SH3 domain protein–peptide interactions (Hahn et al., 2010). Importantly, CE energy expressions also help address the search problem in design. Beyond just speeding up standard Monte Carlo searches, CE energy functions can be used to formulate protein design as an integer linear program in sequence space. The ILP provides provably optimal solutions and flexibility in optimization (see below). Another advantage stems from the fact that in multistate design, the best approach for combining the scores of many states into one objective may not be clear. Deriving CE functions for all of the states allows facile searching and researching using a variety of different objectives and also allows tradeoffs to be rigorously explored, for example, between optimizing stability and specificity. This is discussed further below, where we illustrate one way to do this. Overall, the suitability of CE for a particular problem will depend on many things, including the accuracy with which a desired scoring method can be approximated by its expansion. Previous work has shown that this varies considerably for different scoring functions and different structures. When a cluster-expanded scoring method has low error, it provides a tremendous advantage to the search part of the problem and is worth the cost of fitting. When the error is moderate, CE may still provide a useful filter to help identify promising parts of the sequence space that can then be examined in more detail with more expensive calculations. As more work is done, it will become easier to judge those problems for which CE will provide the greatest benefit.

2.4 How to run a cluster expansion with CLEVER 1.0

Hahn et al. created the open-source package CLEVER 1.0, available at http:// web.mit.edu/biology/keating/software/, to aid users in developing their own CE models (Hahn et al., 2010). Here, we provide an overview of how to use CLEVER to cluster expand an arbitrary scoring method provided by the user. The discussion is geared toward a new user of the program. More details can be found in the original papers. Instructions for installing CLEVER 1.0 can be found in the clever1.0 manual at http://web.mit.edu/biology/keating/software/, or in the docs subdirectory that is created when unzipping clever1.0-package.zip.

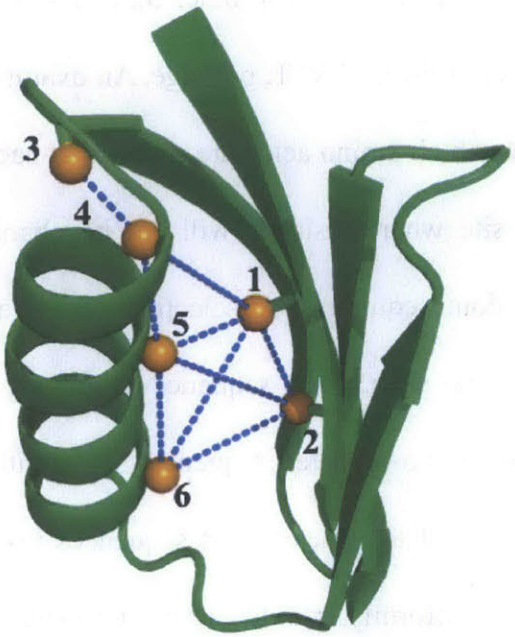
There are three executable modules in the CLEVER 1.0 package. First is the GenSeqs module, which can help the user generate sequences for both training and testing a CE. The second executable is the CETrFILE module. This program executes the crucial step of fitting the ECI values, as described in the theory section. The third module is the CEEnergy module, which uses the CE trained by CETrFILE to score sequences. This module can be combined with other data to assess CE performance.

2.5 GenSeqs

The GenSeqs, or Generate Sequences, module helps with the preparation of unique training and test-set sequences. It requires two inputs. The first is the desired number of training sequences the user would like GenSeqs to return. A rule of thumb is to use at least 2.5 times the number of training sequences as the number of CFs, although a larger number can reduce the

error, as discussed in some detail in Hahn et al. (2010). The other input for the module is the design file, which is crucial to many aspects of the CLEVER package. An example design file is shown in Figure 2-2. The design file states which amino acids are allowed at each of the design sites. It is not necessary to include any site where residues will not be changed. Given the appropriate input, GenSeqs generates random sequences by selecting amino acids uniformly from the allowed amino acids at each of the sites. Each sequence generated is checked for uniqueness such that the training set contains no repeated sequences. In addition, the -o flag combined with the training-set file can be used to generate test sequences not present in the training set. It should be noted that sampling uniformly from the set of amino acids at each of the positions may not provide a good description of the sequence space the user wants the CE to cover. For example, some amino acid substitutions at certain sites may be considered much more common or important, and a user might want to include these with higher frequency. In such a case, the user should generate sequences using their own distribution for the amino acids at each of the sites. An example command line for GenSeqs that would generate 3000 random sequences based on information in design.file is

GenSeqs -n 3000 -d design.file




```
# Design_start
1 A V L I F
2 A V L I F
3 A V L I F
4 A V L I F
5 A V L I F
6 A V L I F
# Design_end
# Cluster_start
1
2
3
...
1 2
1 3
1 4
...
# Cluster_end
```

Figure 2-2. Example design file. On the left is a structure of streptococcal protein Gβ1from crystal structure 1PGA (Gallager, Alexander, Bryan, & Gillard, 1994). The Cβ atoms of the six sites chosen for variation in design are shown as spheres. For clarity, only subsets of the possible pair interactions between the sites are shown as dashed lines. On the right is a sample design file. It is composed of two parts. The first half lists sites where the sequence will vary, and these sites are known as design sites. Each design site line lists the amino acids that will be allowed at that site. In the second half of the design file, the user lists the cluster functions to consider for inclusion in the expansion. A user must list all single sites. Only a subset of the cluster functions used for fitting in this example is shown.

2.6 CETrFILE

The CETrFILE, or CE Training File, module uses the procedure in Figure 2-1 to fit ECI values. This is the heart of the CE method. The module requires two inputs. The CETrFILE

module, like GenSeqs, requires a design file. As mentioned earlier, the design file states which amino acids are allowed at each of the design sites. The first amino acid in each design site position is taken as the reference amino acid for that position. The design file also lists which single and higher order interactions among the design sites should be considered during fitting. This is an important choice, made by the user, which can strongly influence performance. There must be a single body term for each of the design sites in order for the code to run; inclusion of pair or higher terms is optional. The second required input is the training-set file, which includes a list of energies paired with sequences. An example of the format for this file can be found in the clever 1.0 manual. Briefly, this file should have a column of energies that can come from any source. Each energy is followed by the sequence of residues at the design site positions. All sequences should be the same length because all sequences should have the same number of design sites. CETrFILE outputs several things to "standard out" such as a table containing all of the ECI values for all of the CFs. CETrFILE also outputs a binary file containing ECI values trained from the input data. An example command line for CETrFILE is

CETrFile -d design.file -s sequence.file –r training.result

2.7 CEEnergy

CEEnergy scores sequences with the derived CE. This module uses the binary output of CETrFILE, for example, training.result, and a sequence file with a list of test sequences. The format of the test-sequence file is the same as the training-sequence file, except the energy column is not used. To get a good idea of expected performance on new problems, test sequences

should not overlap with sequences that the CE was trained on. Test sequences are specified using only the design site residues. An example command line for CEEnergy is

CEEnergy -r training.result -s sequence.file

2.7 Cluster expansion case study

In this section, we provide a simple illustrative example of how the RosettaDesign conformational energy of selected sites in streptococcal protein Gβ1 can be cluster expanded using CLEVER 1.0. The Gβ1 structure is composed of an alpha helix lying across a beta sheet made up of two beta hairpins. Dahiyat et al. choose the Gβ1 domain as one of the first targets for redesign using automated software (Dahiyat & Mayo, 1997). Unlike coiled coils, for which CE has been used in numerous published examples (Apgar et al., 2009; Grigoryan et al., 2006; Hahn et al., 2010; Zhou et al., 2005), Gβ1 lacks any structural or sequence symmetry and thus represents a generic globular fold. Here, we select only a few residues for modeling, to keep the example very simple.

The first step is to create a design file as shown in Figure 2-2. We selected six positions in and around the core of the Gβ1 structure to vary and designated these as design sites, labeled 1 through 6, in the design file. Our choices correspond to residue positions 5, 7, 20, 26, 30, and 34 in PDB structure 1PGA (Gallager et al., 1994). We allowed the same small set of hydrophobic residues (A, V, L, I, and F) at each of these mostly buried positions. Alanine is listed first for each design position and serves as the reference at each site. In the bottom half of the design file, we specified the interactions between the design sites that should be considered for CE. Only

single and pair interactions between residues were considered, for simplicity, though it is possible to include triplets of amino acids or even higher-order terms. Higher-order terms may improve the accuracy of a CE, and suggested techniques for choosing them can be found in (Grigoryan et al., 2006). In this case study, we considered 15 pair clusters and 25 possible amino acid combinations for each pair, resulting in 375 ECI values to be fit. To fit these terms, 2000 random training sequences, drawn from the possible design space of $5^6 = 15{,}625$ sequences, were generated using the GenSeqs module. The training sequences were modeled on the 1PGA structure and scored using RosettaDesign with the "soft-potential" and "minimize side-chain" flags. For simplicity, only the side chains of the design site residues were optimized. All other side chains were fixed in their crystal-structure coordinates. The corresponding RosettaDesign energy was then paired with each of the 2000 training sequences to make the training-set file, which spanned an energy range over 70 Rosetta energy units (also referred to as kcal/mol). The training-set file and the design file were used to train a CE using the CETrFILE module. Once a CE is trained, it is crucial to evaluate its accuracy, as this can vary widely based on the type of problem, the selection of candidate CFs, and the underlying scoring method being approximated (Apgar et al., 2009; Grigoryan et al., 2006; Hahn et al., 2010). The CETrFILE module reports the cross-validated root mean square (CVRMS) error. This is a metric for assessing how well the results of a predictive model generalize to an independent data set. In this specific example, it is a measure of how well the CE would be expected to predict the RosettaDesign energy of a sequence threaded onto the Gβ1 structure. CE of the RosettaDesign energy function on the 1PGA structure gave a CVRMS score of 1.5 kcal/mol, as shown in Figure 2-3A. An additional test of the error is to generate a test set of sequences, independent of the training set, and score

them using both the structure-based method and the newly derived CE. For our example, the GenSeqs module was used to generate 2000 test-set sequences nonoverlapping with the training set that spanned an energy range of nearly 80 kcal/mol. The root mean square deviation (RMSD) between the scores for the test-set sequences derived from the CE model and the structure-based model was 1.6 kcal/mol, which is shown in Figure 2-3B. Overall, this CE performed very well at approximating the structure-based method it was derived from. The example shown here is very simple. Often, good performance requires refinement of the expansion protocol. There are several techniques for reducing error discussed by Hahn et al. (Hahn et al., 2010). For example, introducing higher order CFs such as triplets, increasing training-set size, or decreasing the number of variable amino acids are approaches for reducing error. Additionally, Hahn et al. presented techniques for identifying and removing from the design space CFs that are particularly poorly fit.

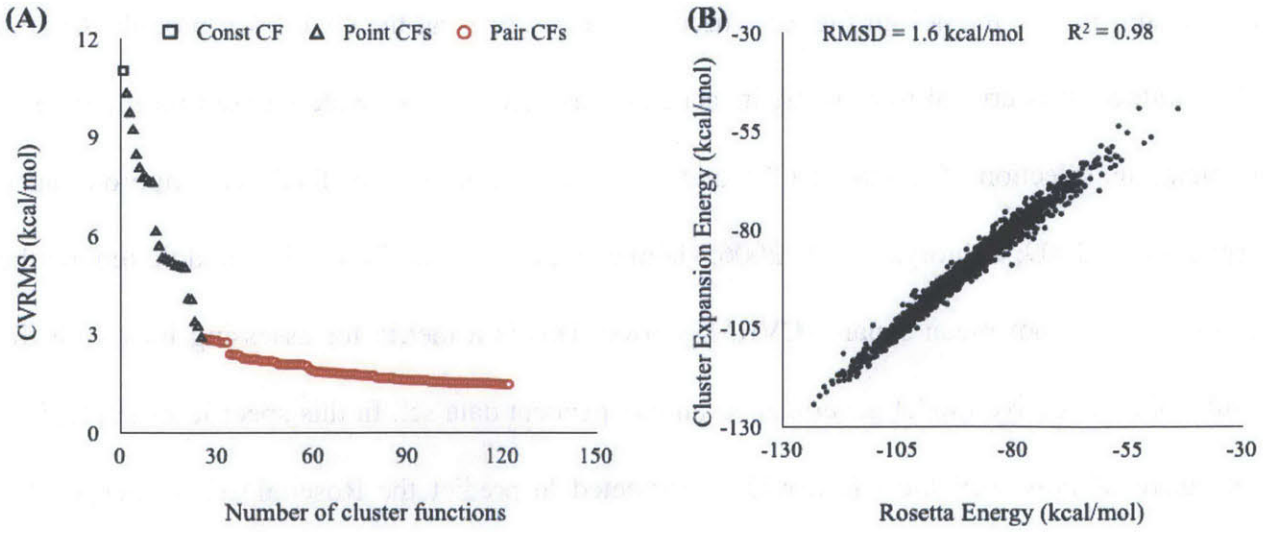


Figure 2-3. Cluster expansion error in the Gβ1 example. (A) Evolution of the CVRMS as CFs were added to the model. The type of CF added in each iteration is indicated by shape, as shown in the legend. A total of 122 CFs were added, giving a CVRMS of 1.5 Rosetta energy units (also referred to as kcal/mol). (B) The performance of the cluster expansion on 2000

randomly generated sequences not included in the training set. The RMSD between the CE test-set energies and the RosettaDesign test-set energies was 1.6 kcal/mol.

2.9 Using cluster expansion with integer linear programming

As mentioned earlier, CE energy functions are amenable to linear optimization techniques. Grigoryan et al. combined CE with ILP, resulting in the computational protocol called CLASSY (Chen et al., 2011; Grigoryan et al., 2009). The use of ILP for protein design was described previously by Kingsford et al. and is most easily explained for single-state design using the graph shown in Figure 2-4. Here, each cluster of nodes represents a design site with the associated nodes corresponding to design choices (Kingsford, Chazelle, & Singh, 2005). For Kingsford et al., these choices represented different conformations of one or more residues; in our case, they represent different residues because we are designing at the sequence level. A design solution corresponds to selecting one node at each site and connecting the selected nodes into a fully connected graph. A brief description of the ILP protocol for protein design based on this graph now follows.

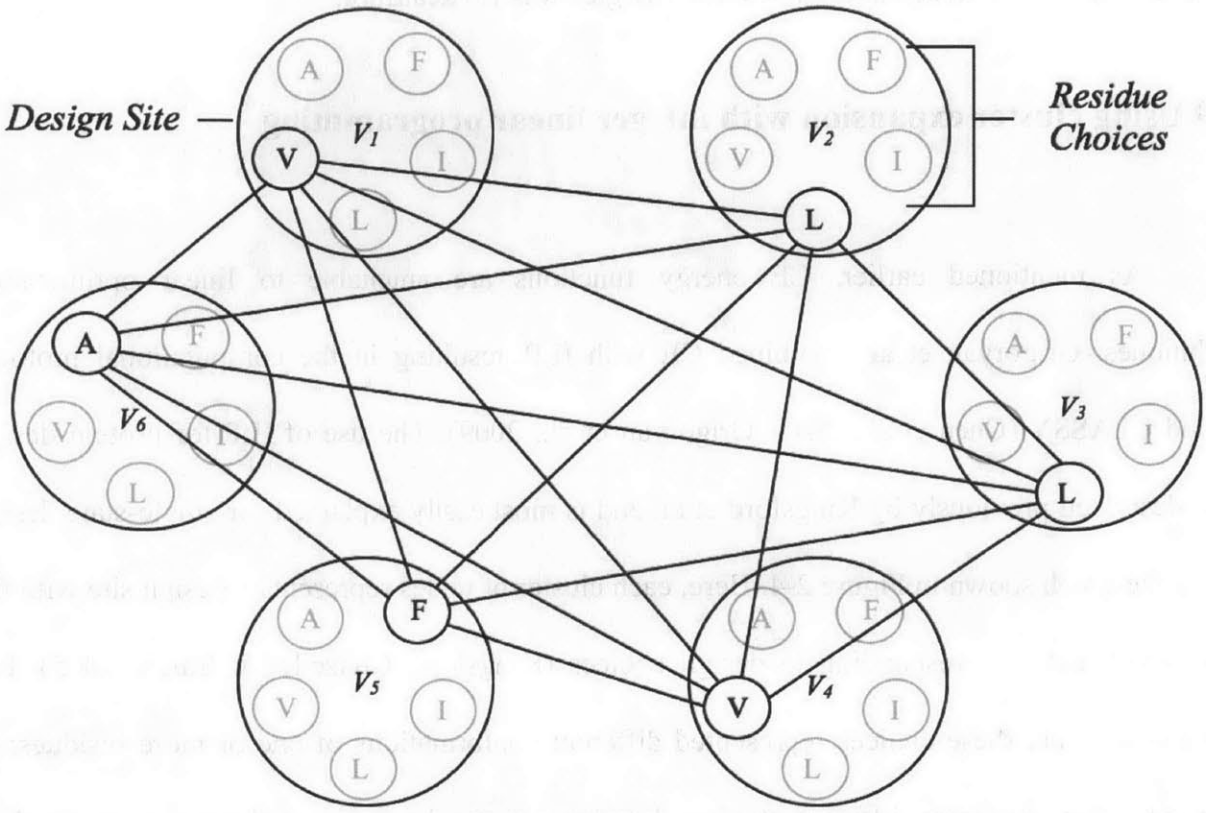


Figure 2-4. Integer linear programming (ILP) formulation for protein design. Each design site consists of nodes representing the allowed residue choices at that position. Edges between nodes represent interactions between those nodes. Our Gβ1 example had six design positions with five choices at each position. One set of nodes and the corresponding edges are highlighted to show one possible design solution.

Using notation similar to Kingsford et al., the sequence space for designing a protein L amino acids long can be represented with a node set $V = V_1 \cup \ldots \cup V_L$. Each subset, V_i, contains a set of nodes that represent the possible amino acids at site *i*. Nodes (u) of V_i have a weight (E^T_{uu}) representing the energetic contribution of that node to the target structure, *T*. Edges between nodes of the graph $D = \{(u,v): u \in V_i \text{ and } v \in V_j,\ i \neq j\}$ also have a weight corresponding to the energetic contribution of that edge (E^T_{uv}) to the target structure. E^T represents the energy of

a sequence as evaluated by the CE model. E^T is obtained by summing over the node energies (E^T_{uu}) and edge energies (E^T_{uv}). When using a cluster-expanded scoring method, both types of contributions can readily be written as sums of linear terms in sequence variables. Thus, minimization of the total energy (Eq. 3) can be done in sequence space using ILP, the only requirement being addition of constraints that enforce a unique and consistent choice of amino acid at each site; this can be done using Eqs. (4)–(6). Equation (4) forces the design solution to have only one amino acid at each design site. Equation (5) can then be used such that only the edges from the amino acid being chosen at each design site are used for edge energies. In Eq. (6), the terms x_{uu} and x_{uv} are the optimization variables and can have values of 0 or 1 corresponding to the absence or presence of a node or an edge, respectively.

$$\textbf{Minimize: } E^T = \sum_{u \in V} E^T_{uu} x_{uu} + \sum_{\{u,v\} \in D} E^T_{uv} x_{uv} \qquad (3)$$

subject to:

$$\sum_{u \in V_j} x_{uu} = 1 \text{ for } j = 1,\ldots,L \qquad (4)$$

$$\sum_{u \in V_j} x_{uv} = x_{vv} \text{ for } j = 1,\ldots,L \text{ and } v \in V \setminus V_j \qquad (5)$$

$$x_{uu}, x_{uv} \in \{0,1\} \qquad (6)$$

ILP has several attractive features for protein design. First, it is an optimal search technique and thus ensures, if any solution is returned, that it will be the global minimum energy according to the cluster-expanded scoring method. Second, we have found in practice that for protein design problems of the type described here, the ILP optimization converges reliably and quickly. Further, ILP readily accommodates the addition of arbitrary constraints that are linear in the optimization variables. Such constraints can include limits on the sequence composition or total

charge or helical propensity. Grigoryan et al. constrained designed sequences to have at least a minimum score based on a position-specific scoring matrix; this was used to ensure that designed sequences resembled natural sequences in their overall characteristics (Grigoryan et al., 2009). In our Gβ1 example, ILP can be used to find the lowest energy sequence on the Gβ1 template. In this case, the values E^{T}_{uu} correspond to point ECI values, while the values for E^{T}_{uv} come from the ECI values for pair interactions. For a multicriterion problem, a user can add constraints, for example, restricting solutions that are similar to the wild-type sequence. An example of such a constraint can be seen in Eq. (7). Here, WT_u takes on the value: 0.16 (one out of six sites) for wild-type residues at their respective positions and 0 for all other residues at those positions. A user can then define the maximum allowed sequence identity between the design solution and the wildtype sequence of Gβ1.

$$\sum_{u \in V} WT_u x_{uu} < Allowed\_SeqID \tag{7}$$

An open-source tool kit for solving ILP problems can be found at http://www.gnu.org/software/glpk/ and can be used with any CE of the type described here.

2.10 CLASSY applied to multistate design

As mentioned in Section 2.1, Grigoryan et al. used CE to design specific peptide inhibitors for human bZIP proteins (Grigoryan et al., 2009). The bZIPs are transcription factors that can homo- and/or heterodimerize by forming a parallel coiled coil. They provide an interesting design challenge because, due to the extensive sequence similarity between different

bZIPs, designing a peptide to specifically interact with one bZIP but not others is challenging (Mason, Muller, & Arndt, 2007). Grigoryan et al. selected one member of each of the 20 human bZIP families as a target for design and used members of the remaining families as examples of off-targets, to which binding of the design was not desirable. This was accomplished by using CLASSY.

The advantage of using an ILP framework for this design problem is that it enabled optimization of the design for interaction with the target with the addition of linear constraints enforcing simultaneous consideration of the additional competing states. This is because both the objective function, E^T, and the energies of the undesired states, E^i, were written as linear functions of design variables x_{uu} and x_{uv}. Thus, the difference in energy between E^T and each E^i could also be written this way. A series of equations of the form $E^i - E^T > \Delta$ was constructed and used as constraints to enforce an energy gap between the designed target and undesired competitors. Figure 2-5 shows how the constrained ILP optimization was used in a protocol known as a "specificity sweep." In the first step of a specificity sweep, the design is optimized for binding to a target with no constraints on interaction energies with off-target partners. Due to the high sequence similarity between bZIP families, this can often lead to sequences predicted to interact more favorably with off-target sequences than with the target. Therefore, in a subsequent round of optimization, a constraint can be imposed with a given value of Δ, such that the design is required to bind the target at least this much better than the off-targets. In subsequent rounds, Δ can be systematically increased until no more solutions can be found. The complete specificity sweep protocol generates an extensive and systematic search of sequences with different predicted stabilities and target versus off-target specificities. A user can then select any sequence

or set of sequences along the specificity sweep for experimental testing. This protocol was successful in generating experimentally validated specific binders, and the interested reader is referred to the original work for details (Grigoryan et al., 2009).

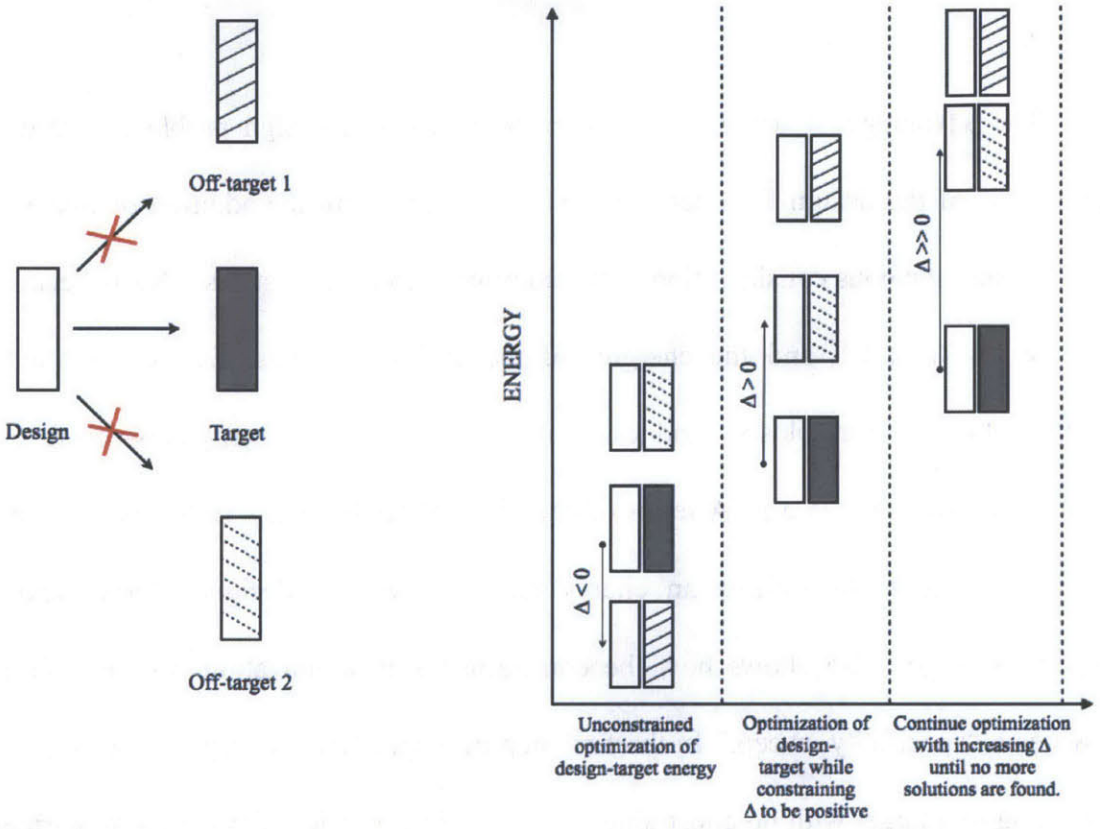


Figure 2-5. A CLASSY specificity sweep, illustrated using bZIP coiled-coil design. On the left is a cartoon representation of the bZIP multistate design problem. The goal in this problem is to design a sequence (white rectangle) that will interact with a target (gray rectangle), yet avoid interactions with off-target sequences (striped rectangles). On the right is a plot of the energies of the various states. Initially, the design–target interaction is predicted not to be the most stable state for the design (far left). Constraints are added in subsequent rounds of design (moving to the right) that impose specificity for the target at the price of the stability of the design–target complex. The constraint on specificity can be increased until the most specific sequence in the space defined for the search is found.

A multistate design criterion can be introduced to our example of redesigning Gβ1. For example, a quadruple mutant of Gβ1 has been shown to form a domain-swapped homodimer (Byeon, Louis, & Gronenborn, 2003). Three of the design sites chosen for our CE case study overlap with the four positions that can bring about the domain swap. To disfavor sequences likely to adopt this domain-swap dimer in design, the domain-swap dimer can be introduced as an explicit undesired state. This requires rescoring the training-set sequences with the RosettaDesign energy function on the domain-swapped homodimer backbone 1Q10 (Byeon et al., 2003). These energies can then be used to derive a new CE that describes the undesired dimer state. An expression just like Eq. (3) can be written for this state, E^i, and a linear constraint like that in Eq. (8) can be imposed to require that the energy gap between the undesired state and target state be greater than Δ. Similar to Grigoryan et al., Δ could be varied and a specificity sweep conducted to give multiple solutions with different values of Δ.

$$E^i - E^T > \Delta, \text{where } E^i = \sum_{u \in V} E^i_{uu} x_{uu} + \sum_{\{u,v\} \in D} E^i_{uv} x_{uv} \quad (8)$$

2.11 Conclusion

CE provides a way of converting a complex nonlinear function into a simple approximation of that function that has a linear dependence on sequence variables. This not only allows protein engineers to convert structure-based models into sequence-based models but also opens the door to new search protocols that operate in sequence space for protein design. In particular, CE combined with ILP promises to be a powerful tool for multistate design, and

previous analyses have suggested that even single-state problems can benefit (Grigoryan et al., 2006). With the rapid acceleration of sequence-based energy evaluation, and the flexibility that it affords in searching sequence space, designing protein–protein interaction networks using protocols where numerous possible interactions are considered may soon be possible. Excitingly, CE is not limited to expanding stabilities or binding energies resulting from theoretical structure-based models. Hahn et al. demonstrated that it is possible to cluster-expand experimental data directly (Hahn et al., 2010), and it may also be possible to cluster expand other protein properties like association and dissociation rates. As larger data sets emerge from high-throughput experiments linking sequences to protein properties, methods like CE will prove to be increasingly powerful tools.

2.12 Acknowledgments

We thank members of the Grigoryan and Keating labs, especially G. Grigoryan, J. D. Curuksu, and O. Ashenberg, for helpful comments. This work was funded by the National Science Foundation Graduate Research Fellowship Program to C. N., and by NSF CAREER award MCB-0347203 and NIH award GM67681 to A. K. We used computer resources provided by National Science Foundation award 0821391.

2.12 References

Allen, B. D., & Mayo, S. L. (2006). Dramatic performance enhancements for the FASTER optimization algorithm. Journal of Computational Chemistry, 27, 1071–1075.

Allen, B. D., & Mayo, S. L. (2010). An efficient algorithm for multistate protein design based on FASTER. *Journal of Computational Chemistry, 31*, 904–916.

Apgar, J., Hahn, S., Grigoryan, G., & Keating, A. E. (2009). Cluster expansion models for flexible-backbone protein energetics. *Journal of Computational Chemistry, 30*, 2401–2413.

Bolon, D. N., Grant, R. A., Baker, T. A., & Sauer, R. T. (2005). Specificity versus stability in computational protein design. *Proceedings of the National Academy of Sciences, 102*, 12724–12729.

Byeon, I. J., Louis, J. M., & Gronenborn, A. M. (2003). A protein contortionist: Core mutations of Gβ1 that induce dimerization and domain swapping. *Journal of Molecular Biology, 333*, 141-152.

Chen, T. S., Reinke, A. W., & Keating, A. E. (2011). Design of peptide inhibitors that bind the bZIP domain of Epstein–Barr virus protein BZLF1. *Journal of Molecular Biology, 408*, 304–320.

Dahiyat, B. I., Gordon, B., & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Science, 6*, 1333–1337.

Dahiyat, B. I., & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences, 94*, 10172–10177.

de Fontaine, D. (1994). Cluster approach to order-disorder transformations in alloys. In H. Ehrenreich & D. Turnbull (Eds.), Solid state physics: advances in research and applications. 47. (pp. 33–176). New York: Academic Press.

Desmet, J., Spriet, J., & Lasters, I. (2002). Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins: Structure, Function, and Bioinformatics, 48*, 31–43.

Drexler, K. E. (1981). Molecular engineering an approach to the development of general capabilities for molecular manipulation. *Proceedings of the National Academy of Sciences, 78*, 5275–5278.

Erijman, A., Aizner, Y., & Shifman, J. M. (2011). Multispecific recognition: Mechanism, evolution, and design. *Biochemistry, 50*, 602–611.

Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M., et al. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science, 332*, 816–821.

Gallager, T., Alexander, P., Bryan, P., & Gillard, G. L. (1994). Two crystal structures of the B1 immunoglobulin binding domain of streptococcal protein G and comparison with NMR. *Biochemistry*, *33*, 4721–4729.

Grigoryan, G., Kim, Y. H., Acharya, R., Axelrod, K., Jain, R. M., Willis, L., et al. (2011). Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science*, *332*, 1071–1076.

Grigoryan, G., Reinke, A. W., & Keating, A. E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature*, *458*, 859–864.

Grigoryan, G., Zhou, F., Lustig, S. R., Ceder, G., Morgan, D., & Keating, A. E. (2006). Ultra-fast evaluation of protein energies directly from sequence. *PLoS Computational Biology*, *2*, 551–563.

Hahn, S., Ashenberg, O. A., Grigoryan, G., & Keating, A. E. (2010). Identifying and reducing error in cluster-expansion approximations of protein energies. *Journal of Computational Chemistry*, *31*, 2900–2914.

Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*, *282*, 1462–1467.

Havranek, J. J. (2010). Specificity in computational protein design. *The Journal of Biological Chemistry*, *285*, 31095–31099.

Havranek, J. J., & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nature Structural Biology*, *10*, 45–52.

Humphris, E. L., & Kortemme, T. (2007). Design of multi-specificity in protein interfaces. *PLoS Computational Biology*, *3*, 1591–1604.

Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., et al. (2008). De novo computational design of retro-aldol enzymes. *Science*, *319*, 1387–1391.

Karanicolas, J., & Kuhlman, B. (2009). Computational design of affinity and specificity at protein–protein interfaces. *Current Opinion in Structural Biology*, *19*, 458–463.

Kingsford, C. L., Chazelle, B., & Singh, M. (2005). Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, *21*, 1028–1036.

Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L., & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nature Structural & Molecular Biology*, *11*, 371–379.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, *302*, 1364–1368.

Leaver-Fay, A., Jacak, R., Stranges, P. B., & Kuhlman, B. (2011). A generic program for multistate protein design. *PLoS One*, *6*, 1–17.

Leaver-Fay, A., Kuhlman, B., & Snoeyink, J. (2005). Rotamer-pair energy calculations using a trie data structure. *Lecture Notes in Computer Science*, *3692*, 389–400.

Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibodyaffinity improvement beyond in vivo maturation. *Nature Biotechnology*, *25*, 1171–1176.

Mason, J. M.,Muller, K. M.,&Arndt, K. M. (2007). Positive aspects of negative design: Simultaneous selection of specificity and interaction stability. *Biochemistry*, *46*, 4804–4814.

Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., Dechancie, J., Betker, J., et al. (2008). Kempelimination catalysts by computational enzyme design. *Nature*, *453*, 190–195.

Sammond, D., Eletr, Z. M., Purbeck, C., & Kuhlman, B. (2010). Computational design of second-site suppressor mutations at protein–protein interfaces. Proteins: Structure, Function, and *Bioinformatics*, *78*, 1055–1065.

Sanchez, J. M., Ducastelle, F., & Gratias, D. (1984). Generalized cluster description of multicomponent systems. *Physica A: Statistical Mechanics and its Applications*, *128*, 334–350.

Voigt, C. A., Gordon, B., & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology*, *299*, 789–803.

Yanover, C., Fromer, M., & Shifman, J. M. (2007). Dead-end elimination for multistate protein design. *Journal of Computational Chemistry*, *28*, 2122–2129.

Zhou, F., Grigoryan, G., Lustig, S. R., Keating, A. E., Ceder, G., & Morgan, D. (2005). Coarse-graining protein energetics in sequence variables. *Physical Review Letters*, *95* (148103), 1–4.

Chapter 3

A Set of Computationally Designed Orthogonal Antiparallel Homodimers that Expands the Synthetic Coiled-Coil Toolkit

Submitted paper to: American Chemical Society.

Molecular engineering of protein assemblies, including the fabrication of nanostructures and synthetic signaling pathways, relies on the availability of modular parts that can be combined to give different structures and functions. Protein interactions are an important modular part, yet a limited number of well-characterized interaction components are available. Coiled coil protein interaction modules have been demonstrated to be useful for biomolecular design, and many parallel homodimers and heterodimers are available in the coiled-coil toolkit. In this work, we sought to design a set of orthogonal antiparallel homodimeric coiled coils using a computational approach. There are very few antiparallel homodimers described in the literature, and none have been measured for cross-reactivity. We tested the ability of the distance-dependent statistical potential DFIRE to predict orientation preferences for coiled-coil dimers of known structure. The DFIRE model was then combined with the CLASSY multi-state protein design framework to

engineer sets of three orthogonal antiparallel homodimeric coiled coils. Experimental measurements confirmed the successful design of three peptides that preferentially formed antiparallel homodimers that, furthermore, did not interact with one additional previously reported antiparallel homodimer. Two designed peptides that formed higher-order structures suggest how future design protocols could be improved. The successful designs represent a significant expansion of the existing protein-interaction toolbox for molecular engineers.

3.1 Introduction

Modular design is used for engineering complex devices in electronics, mechanics, nanotechnology and other fields. Recently, biologists have begun to exploit modular parts as a way to build novel synthetic biological systems.[1] Many types of parts are required to implement diverse structural, binding and catalytic functions. Here, we focus on the alpha-helical coiled coil, which is a protein-interaction domain highly suitable for inclusion in the growing molecular parts toolkit.[2,3] Coiled coils are prevalent in native proteins and are useful interaction motifs due to their capacity to encode complex protein interaction patterns in a short protein sequence.[4,5,6]

Coiled coils form a rod-like structure composed of α-helices that wrap around each other with a superhelical twist. Coiled-coil sequences have a characteristic motif commonly referred to as a heptad repeat, denoted as $[abcdefg]_n$. The *a* and *d* positions are dominated by hydrophobic residues, and are found at the core of the structure; we refer to *a* and *d* as core positions. The *e* and *g* positions are typically occupied by charged residues and form the boundary between the core and the surface of the coiled coil. The *b*, *c*, and *f* positions are located on the surface of the

coiled coil and are most often polar or charged. Lastly, in coiled-coil notation, a prime on a heptad position indicates a residue on an opposing chain.

The relationship between coiled-coil sequence and structure is incompletely understood, even after decades of study of native, mutant and de novo-designed coiled coils. This is partly due to the many topologies accessible to coiled-coil sequences. For example, coiled coils can fold into dimers, trimers, tetramers, and even higher-order oligomers. Additionally, oligomers can be homo- or heteroassemblies. Lastly, the orientations (parallel vs. antiparallel) and axial alignments of the constituent helices can vary.[7,8] The general problem of predicting detailed coiled-coil structure from sequence has not been solved, although progress has been made developing methods to predict oligomerization state from sequence, and in particular to discriminate parallel dimers from parallel trimers.[9-14]

Coiled coils have been used in a wide range of applications. They have been applied to the design of artificial transcription factors and used to manipulate cell-signaling pathways.[15,16] They have also been used to build engineered crystals, and to modulate the charge-transfer properties of electronic devices.[17,18] In many of these studies, controlling the orientation of the helices in the coiled coil was important. For example, Shlizerman et al. modulated the conductance between two monolayers of gold using coiled-coil dimers and showed that parallel and antiparallel coiled coils differentially impacted the electronic properties of the system. Coiled coils of different orientations have net molecular dipoles of different magnitude and direction, and can thereby confer different electronic properties.[18]

Recently, an exciting strategy was developed to design polypeptide polyhedra based around coiled-coil dimers. Gradišar et al. used a set of parallel and antiparallel dimeric coiled

coils as building blocks to engineer a nanoscale single-chain tetrahedron with coiled coils forming each edge.[19] The design strategy involved concatenating a series of 12 sequence segments coding for different coiled-coil helices into a single chain. The artificial protein sequence was designed such that folding of the chain, driven by pairing each coiled-coil helix with its appropriate intra-chain partner helix, would generate a pre-specified three-dimensional structure. A crucial aspect of the design strategy was the use of coiled-coil components that were orthogonal to one another, i.e. that had low potential to cross-interact. The designed tetrahedron was based on 4 parallel and 2 antiparallel coiled-coil dimers previously reported in the literature.[20-23] As part of their work, the authors computed the number and type of coiled coils that would be needed to build different polyhedra. Interestingly, they found that most polyhedra require both orthogonal antiparallel and parallel dimers. For example, of the 6 polyhedra considered by the authors, only an octahedron could be built without using antiparallel dimers.

Despite the clear benefits of having reagents that allow manipulation of orientation in a molecular assembly, most designed coiled coils adopt a parallel orientation. Very few antiparallel coiled-coil dimers have been characterized or designed, and none have been tested for orthogonality. In contrast, dozens of native and synthetic parallel coiled coils have been tested for interactions and orthogonality.[6,23,24] There are currently two databases maintained for designed coiled coils, the SYNZIP database, and the *Pcomp* database.[2,3] Currently 96% of the SYNZIP sequences and ~63% of the sequences in the *Pcomp* database form parallel dimers. Between these two databases, the biophysical properties of only one antiparallel coiled coil (a heterodimer) is reported.[2] Thus, designing sets of orthogonal antiparallel homodimers would expand the available coiled-coil parts in a meaningful way.

Because coiled-coil sequences can encode many different structures, negative design to destabilize undesired states is crucial when making peptides intended to assemble into a single topology.[25] Several negative design strategies have been used in the past that involve placing charged, beta-branched or polar asparagine residues such that they form unfavorable interactions in undesired states.[26-28] A recent study relied on all three of these strategies to design a parallel homodimer, homotrimer, and homotetramer.[3] The orientations of the helices were engineered to be parallel by placing lysines at all *e* positions and glutamates at all *g* positions, which leads to electrostatic attraction in parallel assemblies but repulsion in antiparallel states. Oligomerization states were specified by the differential placement of beta-branched residues in core *a* and *d* heptad positions, a strategy first discovered by Harbury et al., and by the use of asparagine residues to specify dimer formation, which was originally reported by Lumb and Kim.[27,28] Including charged residues in core *a* or *d* positions has also been observed to destabilize non-dimer states.[29]

Designing sets of orthogonal coiled-coil homodimers presents additional challenges related to encoding interaction specificity. This is due to the increased number of undesired, off-target states associated with forming hetero-oligomeric species. The number of possible hetero species increases dramatically as the number of designed orthogonal coiled coils grows, such that three orthogonal antiparallel homodimers have the potential to form six possible off-target parallel or antiparallel heterodimers; other undesired structures are also possible. To design sets of orthogonal antiparallel coiled-coil dimers, we therefore turned to computational methods to keep track of the numerous desired and undesired structures in this design problem.

Despite the many successes of structure-based approaches for modeling and designing protein-protein interactions, treating multiple states is difficult with these techniques.[30,31] The computational costs of modeling each structure can be large, and current optimization functions used with structure-based models do not provide efficient routines for optimizing one set of states while simultaneously destabilizing many off-target states. The multi-state design framework CLASSY addresses these issues by carrying out design in protein sequence space, without the need to explicitly model all protein structures.[32,33] By using a transformation of structure-based models to sequence-based models, CLASSY addresses both the search and scoring problems of multi-state design, and the method has previously been applied to design parallel coiled coils specific for a target sequence over closely related off-target states.[32,34,35]

This paper describes our work applying CLASSY in conjunction with the DFIRE[36] statistical potential to the *de novo* design of sets of coiled coils consisting of three orthogonal antiparallel homodimers. We designed two sets of three proteins, and used biophysical techniques to determine the oligomerization state, helix orientation and thermal stability of structures formed by the designed sequences. Some designed peptides formed trimers or higher-order assemblies, but we identified 3 peptides (APH2, APH3, and APH4) that formed orthogonal antiparallel homodimers. In addition, we showed that these proteins homodimerize in preference to binding to APH, a previously reported antiparallel homodimer.[21] Thus, we provide evidence for four sequences that preferentially form antiparallel homodimers that can be used for protein engineering applications.

3.2 Materials and method

3.2.1 *Building and Scoring Structures with DFIRE**

Structures were modeled on idealized coiled-coil backbones using Rosetta and scored using a modified version of the DFIRE statistical potential that is described below and referred to as DFIRE*. To construct libraries of parallel and antiparallel backbones, a set of 214 canonical coiled coils (i.e. left-handed coiled coils with uninterrupted heptad registers, *abcdefg*) with 2 helices each longer then 20 residues were culled from the CC+ database as of August 18, 2010.[37] Within the parallel and antiparallel sets, examples were filtered to have $\leq$ 50% sequence identity. This set of structures is referred to as the filtered CC+ set. Seven geometrical parameters defined by Crick to describe a coiled coil were fit to each structure using the CCCP Structure Fitter.[38,39] This set of backbones was then filtered to give 25 parallel and 23 antiparallel backbones with parameters within one standard deviation of the average value for each parameter. Averages and standard deviations are reported in Table 3-1. Idealized versions of these 48 structures were generated using the CCCP Structure Generator.[39] Coiled-coil sequences to be scored were modeled on each idealized backbone using the fixed-backbone packing protocol of Rosetta 3.2.[40] The soft-potential flag and expansion of the first and second dihedral angles of the rotamer library were used, along with the side-chain minimization flag. All surface heptad positions (*b*, *c*, and *f*) were modeled as alanine. Structures were scored using a modified version of DFIRE, a distance-dependent pairwise statistical potential based on the distance-scaled, finite ideal-gas reference state.[36] Two modifications were made to the published energy function. The cutoff

distance, r_{cut}, was set to 5.8 Å, and inter-atomic energies were evaluated only between residues on opposite helices in the coiled coil. We refer to this modified version of DFIRE as DFIRE*. DFIRE* outperforms DFIRE on certain interaction prediction tests for parallel coiled coils (V. Potapov, personal communication). The lowest DFIRE* energy for each sequence over all 25 parallel or 23 antiparallel backbones was used as the parallel or antiparallel energy, respectively.

Table 3-1. Averages and standard deviations for Crick parameters fit to coiled-coil crystal structures using CCCP.

Geometric Parameter[a]	Antiparallel		Parallel	
	Average	Stdev	Average	Stdev
R_0 (Å)	4.77	0.25	4.90	0.19
R_1 (Å)	2.77	0.02	2.27	0.02
ω_0 (deg/res)	-3.34	0.80	-3.69	0.64
ω_1 (deg/res)	102.79	0.39	102.79	0.42
α (deg)	-10.78	2.03	-12.17	2.27
$Z_{aa'}$ (Å)	2.64	0.99	0.06	0.36
φ (deg)	352.53	6.81	353.20	7.39

[a] For detailed definitions of the parameters that describe the superhelix geometry, see reference (39). Briefly:

R_0, superhelix radius

R_1, alpha-helix radius

ω_0, superhelical frequency

ω_1, alpha-helical frequency

α, pitch angle, i.e. the angle between a tangent to the super-helical curve and the super-helical axis.

$Z_{aa'}$, helical axial shift, i.e. the offset between the two alpha helices along the super-helical axis.

φ, minor helical phase; defines the rotation of each alpha helix around its own axis, relative to the superhelix interface. Defined using *f*-position residues.

3.2.2 *Deriving cluster expansion models*

Two cluster-expanded functions based on DFIRE* were derived to score the propensity of sequences to form antiparallel and parallel coiled coils. For an outline of the protocol, see Figure 3-1, and for an in-depth discussion of performing cluster-expansion calculations using CLEVER 1.0 see Negron et al..[33] In the present application, the cluster-expanded models express energy as a sum of terms corresponding to weights for single amino acids at *a*, *d*, *e*, and *g* heptad positions and pairs of amino acids at these positions. As in Grigroyan et al., only pairs of positions within the same or adjoining heptads were considered.[41] Weights were fit to training data using the CLEVER 1.0 package.[33,42] The training data consisted of DFIRE* energies for a central two-heptad unit within a six-heptad structure, calculated using the scoring protocol described in the previous section for 30,000 sequences. Another 8,000 sequences, non-overlapping with the training set, were generated in the same way to be used as a test set. Training sequences were 42 residues (six heptads) long and composed of a repeating two-heptad unit. Training sequences were generated randomly but with heptad-specific single-residue frequencies matching those of known coiled-coil dimers (both parallel and antiparallel). Antiparallel frequencies were obtained from antiparallel structures in the filtered CC+ set.[37] Parallel frequencies were obtained from the NPS database.[14] Once determined, the cluster expansion (CE) weights can be used to score antiparallel and parallel coiled-coil dimers of arbitrary length.

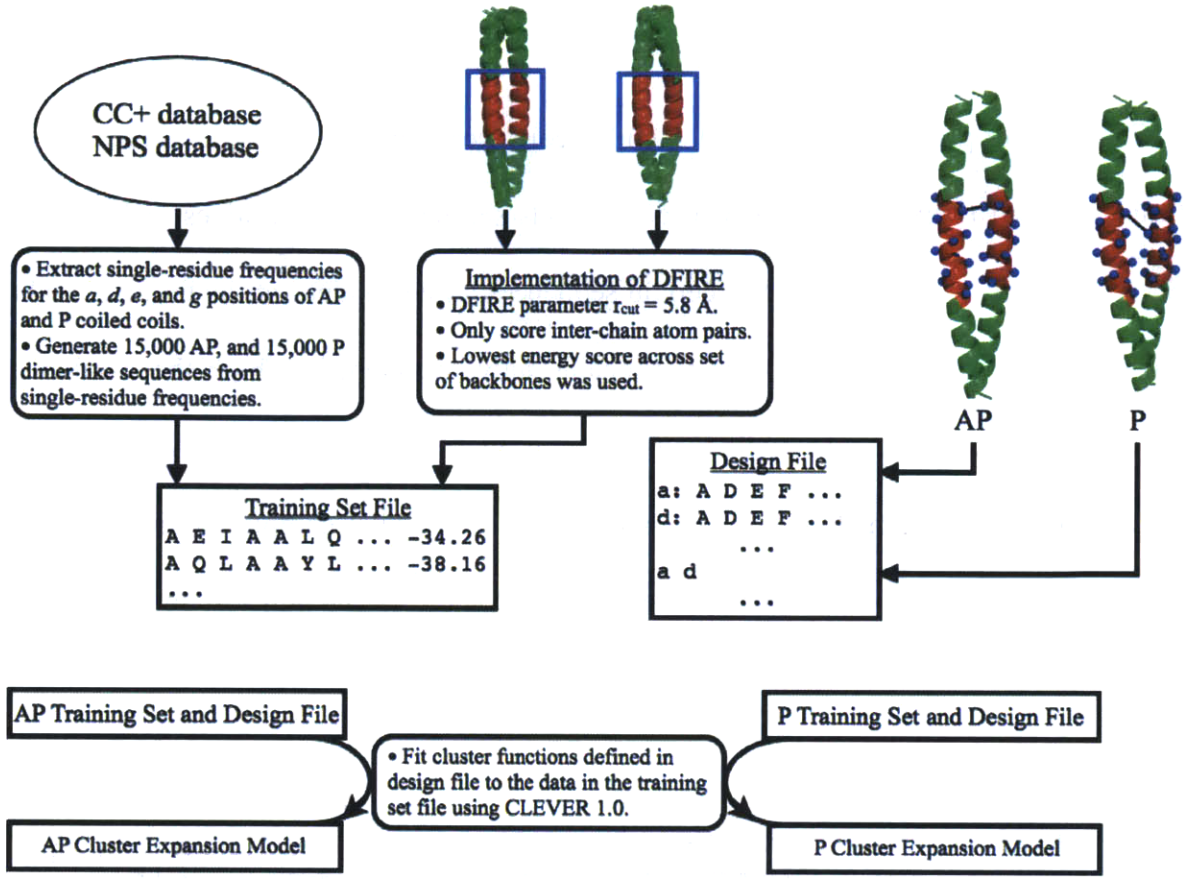


Figure 3-1. Schematic for deriving the antiparallel (AP) and parallel (P) cluster expansion models. The training set files for antiparallel and parallel coiled coils consisted of a list of randomly generated dimer-like sequences two heptads long, with energies computed using DFIRE*. The same set of 30,000 sequences was used to derive both parallel and antiparallel CE models; the set consisted of 15,000 sequences with residue frequencies matching those in P coiled-coil dimers and 15,000 with frequencies matching AP coiled-coil dimers. Energies were obtained by modeling a two-heptad sequence (red) as the central unit in a six-heptad long structure consisting of three repeats of the two-heptad unit. The CE design files specified important parameters for model fitting. They defined the sequence space for which the CE was relevant by listing allowed amino acids for each position, and also listing which pairs of heptad positions were included in the fitting procedure (an example of a pair that contributed to the CE energy is shown on the AP and P structures as a dashed line). This entire procedure is described in more detail in (33).

3.2.3 *Orientation test set*

Examples of parallel and antiparallel coiled coils were obtained from the filtered CC+ set and further filtered to exclude those with coiled coils shorter then 28 residues and those that contained non-natural amino acids. For certain sequences, three residues at the terminal ends of the two chains were removed until the two chains fully overlapped in both the parallel and antiparallel orientations, i.e. the coiled coils that were modeled were blunt-ended in both orientations. The final orientation test set contained 30 antiparallel complexes and 48 parallel complexes. PDB IDs with chain and residue numbers for the orientation test set are given in Table 3-2.

Table 3-2. Orientation test set.

Antiparallel			Parallel		
PDB ID	Chain(s)	Residue Numbers	PDB ID	Chain(s)	Residue Numbers
1cii	A	394-436, 232-274	1ci6	A, B	303-338, 247-282
1ecm	A, B	7-38, 7-38	1d7m	A, B	244-339, 244-339
1ek9	A	330-358, 372-400	1deb	A, B	6-41, 6-41
1exj	A, B	77-116, 77-116	1fos	E, F	158-190, 282-314
1few	A	34-65, 74-105	1gd2	E, F	97-125, 97-125
1fxk	C	94-132, 7-45	1go4	G, H	494-526, 494-526
1hf9	A, B	6-38, 6-38	1h88	A, B	299-331, 299-331
1io1	A	408-439, 64-95	1ik9	A, B	123-169, 123-169
1l8d	A	460-495, 402-437	1j1d	B, C	226-268, 90-132
1qoy	A	117-148, 226-257	1jcc	A, C	13-48, 12-47
1t3j	A, B	688-726, 688-726	1jnm	A, B	273-305, 273-305
1x03	A, B	212-244, 212-244	1joc	A, B	1306-1337, 1306-1337
1ybz	A, B	4-35, 4-35	1kd9	C, D	2-33, 2-33
2b5u	C	393-432, 333-372	1n6m	A, B	300-345, 300-345
2ch7	A, B	444-489, 262-307	1no4	A, B	36-71, 36-71
2d8e	A, B	5-36, 5-36	1pl5	A, S	1285-1341, 1285-1341

2hko	A	434-462, 483-511
2hld	G	227-258, 3-34
2jee	A, B	46-78, 11-43
2oto	A, D	158-190, 137-169
2p4w	A, B	127-155, 131-159
2q0o	C	28-56, 67-95
2vkl	A, B	14-45, 14-45
2zqm	A	72-110
3ggy	A	16-44, 53-81
3a8p	A	671-699, 631-659
3htk	A, B	310-355, 751-796
3i9w	A	99-130, 262-293
3i9y	A	111-146, 278-313
3aei	A	9-37, 63-91

1r05	A, B	47-79, 47-79
1tu3	F, G	807-835, 807-835
1uii	A, B	99-145, 99-145
1uix	A, B	978-1041, 978-1041
1wu9	A, B	193-224, 193-224
2aze	A, B	202-230, 203-231
2c9l	Y, Z	193-221, 193-221
2dfs	A, M	972-1042, 972-1042
2fxm	A, B	852-957, 852-957
2gd7	A, B	32-60, 32-60
2gzh	B, C	453-481, 453-481
2no2	A, B	535-577, 535-577
2o1k	A, B	98-133, 99-134
2ocy	A, B	123-151, 123-151
2oqq	A, B	5-40, 5-40
2q6q	A, B	69-121, 69-121
2v4h	A, B	291-333, 291-333
2v4h	A, B	253-281, 253-281
2v66	C, D	62-97, 62-97
2w6a	A, B	435-480, 435-480
2w83	C, D	396-445, 396-445
2yy0	A, B	62-90, 62-90
3a2a	A, B	228-263, 228-263
3a7p	A, B	64-127, 64-127
3bas	A, B	846-916, 846-916
3cl3	D, E	209-244, 209-244
3dkw	A, B	127-155, 127-155
3e1r	A, B	170-205, 170-205
3he4	A, B	17-52, 3-38
3he5	A, B	3-45, 3-45
3mud	A, B	239-274, 239-274
3nwh	B, D	116-148, 116-148

3.2.4 *CLASSY peptide design*

A detailed description of how integer linear programming (ILP) can be applied as part of the CLASSY multi-state design method is given in Negron *et al.*.[33] In this work, the objective function for ILP was the total energy (E^T), given by the sum of the energies of three antiparallel homodimers ($E^T = E^1 + E^2 + E^3$). All energies were obtained from either the antiparallel or parallel cluster-expanded models. The ILP solver of the IBM ILOG CPLEX optimizer was used to minimize this objective function under a set of constraints.[43] The constraints included energy gaps to off-target dimer states (Figure 3-3A, 3-3B), as well as constraints on the number of polar residues allowed at *a* and *d* heptad positions (maximum of 2 charged residues at *a*, and 1 Lys residue at *d* per design sequence). A constraint was included on the energy gap between every antiparallel homodimer and every off-target state (of those types considered in the calculation) that the constituent peptide could participate in. The constraints were of the form $E^{OT} - E^x > \Delta$, where E^{OT} represents the energy of a single off-target state, of which there were several as shown in Figure 3-3. E^x represents the energy of a single antiparallel homodimer, i.e. E^1, E^2, or E^3. Δ is a user-defined specificity gap, and different values of Δ were used as shown in Figures 3-3C and 3-3D. A solution, representing three sequences, was obtained for each Δ. Two sets of design calculations were done, one including glutamate as an option at *a* positions (sequence space 1) and one not allowing glutamate (sequence space 2). One solution was chosen manually for experimental testing from each calculation, based on predicted stabilities and specificities.

3.2.5 *Cloning, protein expression, and purification*

Synthetic genes encoding computationally designed coiled-coil sequences, and control sequences, were constructed by PCR amplification from two 258-base pair oligonucleotides and one 157-base pair oligonucleotide (gblocks) purchased from Integrated DNA Technologies. DNA sequences were codon-optimized for expression in *Escherichia coli* using DNAWorks.[44] Low-frequency *E. coli* codons selected by DNAWorks were manually switched with synonymous high-frequency codons.

Following amplification with primers to provide appropriate vector overlap, Gibson cloning (New England Biolabs) was used to clone synthetic genes into pENTR vectors. The products of the Gibson reactions were then recombined into pMAL (New England Biolabs) destination vectors using LR Clonase II (Invitrogen) in 2.5 μL reactions. pMAL encodes MBP followed by a TEV protease cleavage site (not used), a Gateway linker region, and a C-terminal His_6 tag. The LR Clonase II reaction inserted the synthetic gene between the Gateway linker region and the C-terminal His_6 site. The pMAL vectors were transformed into BL21 (DE3) cells (Agilent). BL21 cells were grown in liquid LB cultures (1 L) at 37 °C to an OD_{600} of ~0.4-0.6. Protein expression was then induced with 1 mM IPTG for 4.5-5.5 h. Cells were pelleted, resuspended, and then lysed by sonication. MBP-fused proteins were purified from the supernatant using NiNTA (Qiagen) column purification under native conditions. The elution buffer contained 0.3 M imidazole, 20 mM Tris base, and 0.5 M NaCl at a pH of 7.91. The approximate sizes of MBP-fused proteins were confirmed using protein gels with size ladders.

A second set of constructs was made by amplifying from gblocks using primers encoding a cysteine either at the N-terminal or C-terminal end, as well as flanking *BamHI/XhoI* restriction sites. The genes were cloned by means of the *BamHI/XhoI* restriction sites into a modified version of the pDEST17 vector. This vector encodes an N-terminal His_6 tag as well as a GESKEYKKGSGS linker shown to improve the solubility of recombinant proteins.[34] Cysteine-containing constructs were expressed in RP3098 cells grown, induced and lysed as described above for BL21. However, these proteins were purified from the supernatant using NiNTA (Qiagen) under denaturing conditions. The elution buffer consisted of 60% acetonitrile (HPLC-grade) and 0.1% trifluoroacetic acid (TFA). Ni-affinity purification was followed by reverse-phase HPLC with a water/acetonitrile gradient in the presence of 0.1% TFA. Masses were confirmed by MALDI-TOF mass spectrometry.

Concentrations of all constructs were determined using the Edelhoch method,[45] measuring UV absorbance of aromatic residues at 280 nm in 6 M guanidinium chloride. Amino-acid sequences of all constructs are given in Table 3-3.

Table 3-3 Protein and peptide constructs.

Construct	Sequence
Coiled coils fused to MBP	
APH	MBP-TEV-GL-KQLEKELKQLEKELQAIEKQLAQLQWKAQARKKKLAQLKKKLQA-GL-His_6
APH_i	MBP-TEV-GL-KEEKQIEKELKQIEKELQAIEWRLAQLRKRLQALRKRKAQKRE-GL-His_6
APH_{ii}/APH2[a]	MBP-TEV-GL-KRLKQLEKRLKQLRKRKQAKRWEEAQIEKELQAIEKQLAQIRE-GL-His_6
APH_{iii}/APH3[a]	MBP-TEV-GL-KRKKQKRKRAKQLRKRLQALEWQLAQIRKELQAAEKEEAQIEE-GL-His_6
APH_{iv}	MBP-TEV-GL-KEKKQLRKELKQLEKELQALRWRLAQIEKRLQAIRKRLAQKEE-GL-His_6
APH_v	MBP-TEV-GL-KRLKQKEKRKKQLRKRLQALRWQLAQIEKELQAAEKEAAQLRE-GL-His_6
APH_{vi}/APH4[a]	MBP-TEV-GL-KQLKQIEKRLKQIEKRLQAKEWEKAQLRKELQALRKKLAQLRE-GL-His_6
Peptides with N-terminal cysteine	
APH	SHHHHHHGESKEYKKGSGSCGGKQLEKELKQLEKELQAIEKQLAQLQWKAQARKKKLAQLKKKLQA
APH_{iii}	SHHHHHHGESKEYKKGSGSCGGKRKKQKRKRAKQLRKRLQALEWQLAQIRKELQAAEKEEAQIEE
APH_{ii}/APH2[a]	SHHHHHHGESKEYKKGSGSCGGKRLKQLEKRLKQLRKRKQAKRWEEAQIEKELQAIEKQLAQIRE
APH_{iii}/APH3[a]	SHHHHHHGESKEYKKGSGSCGGKEKKQLRKELKQLEKELQALRWRLAQIEKRLQAIRKRLAQKEE
APH_{vi}/APH4[a]	SHHHHHHGESKEYKKGSGSCGGKQLKQIEKRLKQIEKRLQAKEWEKAQLRKELQALRKKLAQLRE
Peptides with C-terminal cysteine	
APH	SHHHHHHGESKEYKKGSGSKRLKQLEKELKQLEKELQAIEKQLAQLQWKAQARKKKLAQLKKKLQAGGCYY
APH_{ii}/APH2[a]	SHHHHHHGESKEYKKGSGSKRLKRLKQLEKRLKQLRKRKQAKRWEEAQIEKELQAIEKQLAQIREGGCYY
APH_{iii}/APH3[a]	SHHHHHHGESKEYKKGSGSKRKKRKKQKRKRAKQLRKRLQALEWQLAQIRKELQAAEKEEAQIEEGGCQW
APH_{vi}/APH4[a]	SHHHHHHGESKEYKKGSGSKQLKQLKQIEKRLKQIEKRLQAKEWEKAQLRKELQALRKKLAQLREGGCYY

[a] Some peptides were given two names for clarity of exposition; the official names are APH2-4.

3.2.6 *Sedimentation equilibrium analytical ultracentrifugation*

Proteins were dialyzed with three changes of reference buffer (40 mM Tris base, 150 mM NaCl, pH 7.91) over the course of 24 hours. Sedimentation equilibrium runs were performed with a Beckman XL-I analytical ultracentrifuge using an An-50 Ti rotor at 20 °C. Proteins were spun at three speeds and at least two protein concentrations. Constructs fused to MBP were spun at concentrations ranging from 4 to 40 μM at 10,200, 16,300 and 20,400 rpm. These spins were monitored either using UV absorbance at 280 nm, or with interference optics when multiple MBP constructs were mixed. For protein constructs containing cysteine, 1 mM TCEP was added to the reference buffer prior to dialysis. These constructs were spun at concentrations of 20 and 40 μM at 28,000, 35,000 and 42,000 rpm and monitored using interference optics. For each speed, equilibrium was confirmed by negligible differences between the sample distributions in the cells over sequential scans. Data sets for each construct were globally fit to a model for a single ideal species using the program SEDPHAT.[46,47] Values for v-bar, solvent density, and viscosity were obtained from SEDNTERP.[48]

3.2.7 *Disulphide-exchange experiments*

Cysteine-containing proteins in varying states of oxidation/reduction (depending on construct) were placed in a redox buffer (500 μM reduced glutathione, 250 μM oxidized glutathione, 40 mM Tris base, 150 mM NaCl, pH 7.91) at 20 μM of each protein at room temperature. Redox reactions were quenched at different time points using a drop of 6 M

hydrochloric acid. The products of the reactions were then run on an analytical Vydac C_{18} reverse-phase column with absorbance monitored at 220 nm using a linear water/acetonitrile gradient containing 0.1% TFA. Equilibrium was confirmed by monitoring changes in HPLC profiles as a function of time. Retention times for the reduced proteins and for the oxidized states for each of the 6 cysteine-containing proteins were assigned by HPLC analysis of the constructs in TBS (40 mM Tris base, 150 mM NaCl, pH 7.91) alone, in TBS with TCEP added for an incubation time of 30 minutes (to generate the fully reduced species), or in TBS solution left exposed to air and stirring overnight (to generate the fully oxidized species). Glutathione adduct peaks were assigned by the appearance, following incubation in redox buffer, of a peak with a retention time not consistent with the reduced or oxidized states of each of the six individual protein constructs. Antiparallel peaks were assigned by monitoring the appearance of a peak only observed after mixing two constructs that encoded the same coiled coil, but with cysteine residues at opposing ends.

3.2.8 *Circular dichroism (CD) spectroscopy*

CD spectra and thermal-denaturation curves were measured on an AVIV 400 spectrometer. Peptides were equilibrated in PBS buffer (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4, pH 7.4) containing 1 mM of dithiothreitol (DTT) at ~25 °C for at least 1.5 hours prior to measurement. Measurements were made in a 1 mm quartz cuvette at a protein concentration of 20 μM using the N-terminal cysteine-containing constructs. CD spectra were measured at 25 °C. For each sample, three wavelength scans were measured and then averaged.

For each wavelength scan, data were collected from 190 to 280 nm, in 1 nm steps, averaging for 5 s at each wavelength. Thermal denaturation curves were generated by monitoring θ_{222} using a 30 s averaging time, 3 minute equilibration time, and temperature increments of 2.5 °C from 0 to 98 °C. Melting temperatures, T_m, were obtained by fitting the change of the CD signal over the change in temperature to the equation below.[49] Fitting was performed using the non-least squares method in Matlab 7.8.

$$\frac{D*\exp(H/T_m - H/T)}{(1+\exp(H/T_m - H/T))^2} * \frac{H}{T^2}$$

The fit parameters are D, H, and T_m. D is the difference between the upper and lower baseline, H is the change in enthalpy, ΔH, over the gas constant R, T is the temperature, and T_m is the melting temperature.

3.3 Results

3.3.1 *Benchmarking DFIRE* on orientation prediction preference*

Computational design of orthogonal antiparallel homodimers requires an energy function capable of scoring antiparallel vs. parallel dimers. To assess whether our design energy function could predict helix orientation for coiled-coil dimers of known structure, we implemented a test similar to that in Apgar et al..[50] We created a database of 30 antiparallel and 48 parallel dimer structures based on the CC+ database of Testa et al.;[37] we refer to this database as the orientation test set (see Methods). The orientation test set in this study differed from that used by Apgar et al. due to its higher stringency on length, ≥28 residues vs. ≥18 residues.[50] This more stringent

cutoff has the effect of removing examples of short coiled-coil sequences embedded in large structures, for which the helix orientation is less likely to be determined by the sequence of the coiled-coil region alone. Furthermore, sequence features of antiparallel coiled coils in the PDB are a function of their lengths, e.g. shorter coiled coils have a 16% higher frequency of hydrophobic residues at the *g* position (Table 3-4).

Table 3-4 Frequencies of polar residues at different heptad positions in antiparallel coiled coils.[a]

	Heptad Positions						
Length[b]	a	b	c	d	e	f	g
≥28	0.32	0.71	0.74	0.20	0.71	0.74	0.73
≥21	0.31	0.71	0.67	0.21	0.67	0.72	0.57

[a] Polar residues included: D, E, H, K, N, Q, R, S, and T.
[b] Antiparallel coiled-coil sequences were culled from the CC+ database with ≤ 50% sequence identity, with varying cutoff lengths as indicated in the table (37).

A modified version of DFIRE, DFIRE*, which includes only inter-chain energy terms was used for scoring. The orientation test-set sequences were modeled in both parallel and antiparallel orientations using Rosetta and scored using DFIRE*, as described in the Methods. The DFIRE* energy gap between the antiparallel and parallel state for each sequence is plotted in Figure 3-2A. We report energies in arbitrary units (AU), as we have no information at this time about how predicted energies from this procedure correlate with experimental free energies. The ability of DFIRE* to predict orientation preference on the test set was measured using the area under the curve (AUC) when plotting the fraction of parallel test-set sequences predicted correctly vs. the fraction of antiparallel sequences predicted correctly, as a function of the score

cutoff used to discriminate parallel from antiparallel sequences. As seen in Figure 3-2B, DFIRE* predicts orientation preference in this test with an AUC value of 0.91 (random predictions would result in an AUC of 0.5).

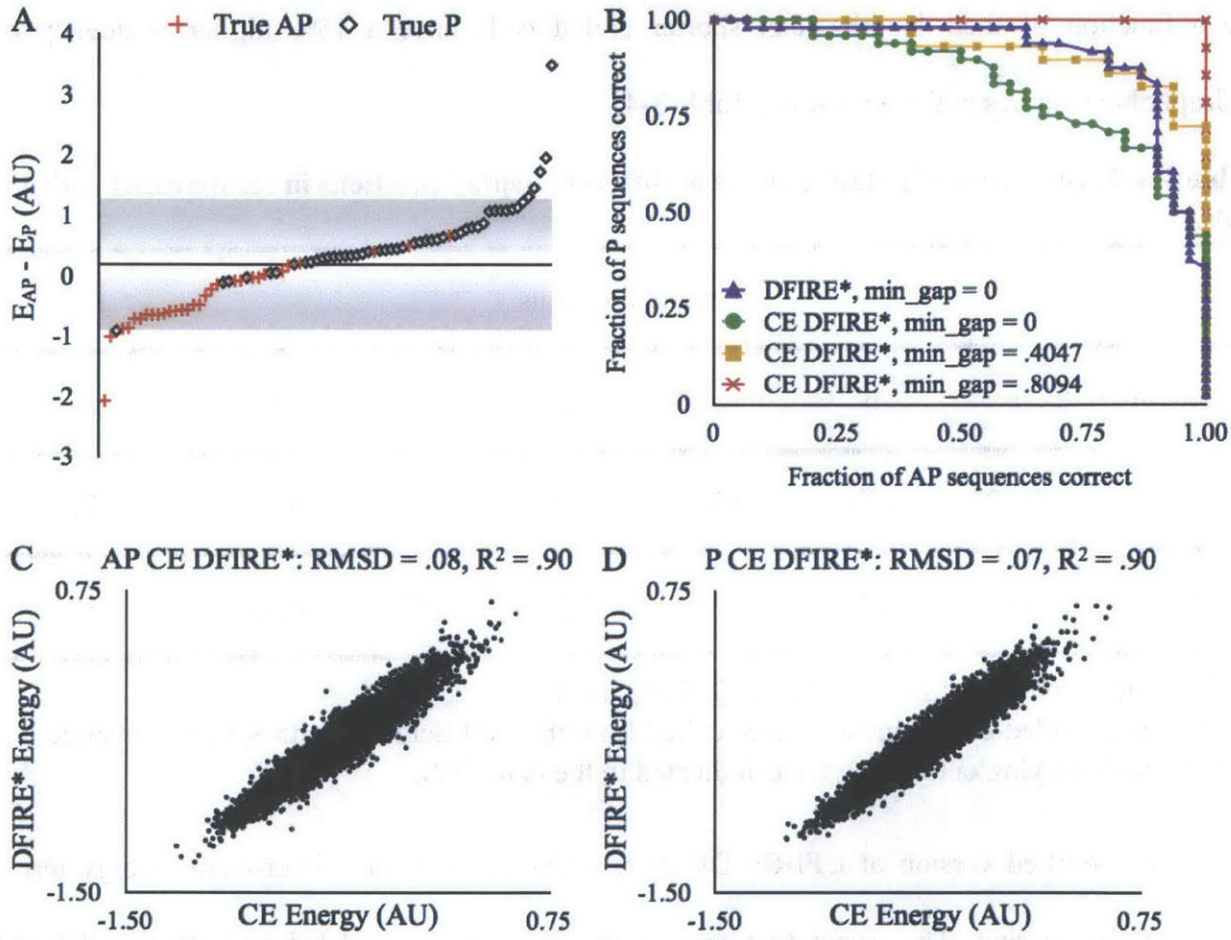


Figure 3-2. Predicting coiled-coil orientation preference and testing cluster-expanded DFIRE*. (A) E_{AP} and E_P are the antiparallel (AP) and parallel (P) DFIRE* energies for each orientation test set coiled coil. Antiparallel coiled coils (according to PDB structure) are plotted with red crosses; parallel with black diamonds. The line at E_{AP} - E_P = 0.18 AU gives optimal separation of parallel and antiparallel examples. Min_gap was used to remove examples with small DFIRE* orientation preferences (see text); shading indicates increasing min_gap from the line of optimal separation. (B) The fraction of antiparallel sequences predicted correctly vs. the fraction of parallel sequences predicted correctly, as the cutoff value for E_{AP} - E_P was changed, is plotted for DFIRE* and the CE model of DFIRE*. Curves for data sets with different min_gap are shown for the CE model of DFIRE*. (C, D) DFIRE* energies vs. the CE model of DFIRE* energies for randomly generated dimer-like test structures in the antiparallel (C) and parallel (D) states.

3.3.2 *Cluster expansion of DFIRE**

Cluster expansion (CE) is a computational method for generating a sequence-based scoring function that approximates energies calculated using structure-based techniques.[41,42,51] Once generated, a CE model eliminates the need for computationally costly structure building in protein design. Two CE models were built to approximate DFIRE* energies for antiparallel and parallel coiled-coil dimers, as described in the Methods, and the models were used to score 8,000 test sequences (Figures 3-2C, 3-2D). Both models showed good correlation with DFIRE*, $R^2 = .90$, indicating that the approximation of structure-based modeling with a sequence-based function introduced relatively little error within the sequence space explored.

We benchmarked the orientation prediction performance of the CE models using the orientation test set. Every pair of sequences in the set was scored with the antiparallel CE model and the parallel CE model. The energy difference between the two CE models was used to predict the orientation preference of each sequence. The AUC value, using the CE approximation of DFIRE*, was 0.84 (Figure 3-2B), demonstrating that the faster, yet more approximate model gave reduced performance, as expected. However, the AUC value significantly improved as coiled coils with small energy gaps were removed from the orientation test set. For 44 coiled coils for which the predicted difference in CE energy between parallel and antiparallel orientation was greater than 0.4047, the prediction performance (0.93) was similar to the performance of DFIRE* on the entire orientation test set. For 20 examples with predicted energy gaps greater than 0.8094, prediction performance was perfect. This information was used to set energy gap requirements for off-target states during the sequence-design stage of CLASSY.

3.3.3 *Computational design of orthogonal antiparallel homodimers using CLASSY*

CLASSY is a protein-design method that uses integer linear programming (ILP) to optimize a protein sequence using a CE scoring function. Importantly, the method allows a user to impose numerous constraints on the designed sequence. These can include constraints on sequence composition or properties (e.g. total charge). In multi-state design, it is convenient to impose a constraint on the energy of a designed sequence adopting an undesired structure, to disfavor formation of that structure.

In our application, the antiparallel and parallel CE models were combined with ILP to do CLASSY design of six-heptad long antiparallel homodimers. Only residues at *a*, *d*, *e*, and *g* positions were designed; these residues are thought to be most critical for interaction specificity.[52,53] The *b*, *c*, and *f* surface positions were taken from APH, which is one of the few characterized antiparallel homodimers reported in the literature. The surface of APH mainly consists of an oscillating pattern of glutamine and alanine residues at *b* and *c* positions, and lysine residues at *f* positions. This patterning has been used in both parallel and antiparallel designed coiled coils, and is thought to play a minimal role in interaction specificity.[21,26]

We used the CE model of DFIRE* to design the globally best-scoring antiparallel homodimer in a sequence space without cysteine, proline, or glycine and found that the designed sequence was highly charged and contained no hydrophobic residues in any heptad position. This peptide would not be expected to fold into a coiled-coil structure. The unrealistic design sequence is not inconsistent with the good performance of DFIRE* and the CE model of DFIRE* on the orientation prediction test above. In the orientation test, each of two compared

structures had the same sequence. In contrast, without constraints on sequence composition, optimization using the CE model of DFIRE* has the freedom to build a sequence entirely from charged pairs that have highly favorable CE weights. Surprisingly, the 20 most favorable weights in the CE DFIRE* model are all core-to-edge, or core-to-core charge-charge residue interactions. The weight of the most stabilizing hydrophobic-hydrophobic interaction is two-fold weaker then the most stabilizing charge-charge interaction. To use CE DFIRE* in protein design, we therefore imposed a native-like sequence composition on all sequences and restricted the design calculations to subsets of sequence space, as described below.

Two separate sequence spaces, sequence space 1 and sequence space 2, were chosen to search for antiparallel homodimer sequences (Figures 3-3A, 3-3B). Both sequence spaces included residues known to influence coiled-coil structural specificity through mechanisms such as electrostatic attraction/repulsion and beta-branch residue packing/clashing.[26,27] Sequence space 1 differed from sequence space 2 by the addition of glutamate as a choice at *a* positions. Statistics from the coiled-coil databases we analyzed show a three-fold frequency enrichment of glutamate in *a* sites of antiparallel dimers relative to parallel dimers (Table 3-5); this difference has also been noted by Straussman et al..[54]

To design three non-interacting coiled coils, we optimized the sum of the CE energies of three antiparallel homodimers using CLASSY. Constraints were added to allow no more than two hydrophilic residues at *a* positions and no more than one at *d* positions. This maintained the hydrophobicity of the design solutions at these positions close to that of known antiparallel dimers of lengths greater than four heptads. Constraints were also placed on the predicted energies of competing states. In particular, all design calculations treated all three possible

antiparallel heterodimer states as undesired states. Without these constraints, the global energy minimum would correspond to three copies of the lowest-energy antiparallel homodimer. Constraints on the off-target states were imposed as an energy gap, by requiring the energy of each antiparallel homodimer to be lower than the energy of each of the off-target states that sequence could participate in, by a fixed amount (Figure 3-3). Excluding parallel trimers, very few structures of higher-order states of any specific topology passed the orientation test set filters of <= 50% sequence identity and > 27 residues (0 antiparallel trimers, 6 antiparallel tetramers, and 9 parallel tetramers). Thus it was not possible to accurately benchmark DFIRE* ability to discriminate oligomerization preference. As a result, we did not include these states in the modeling process.

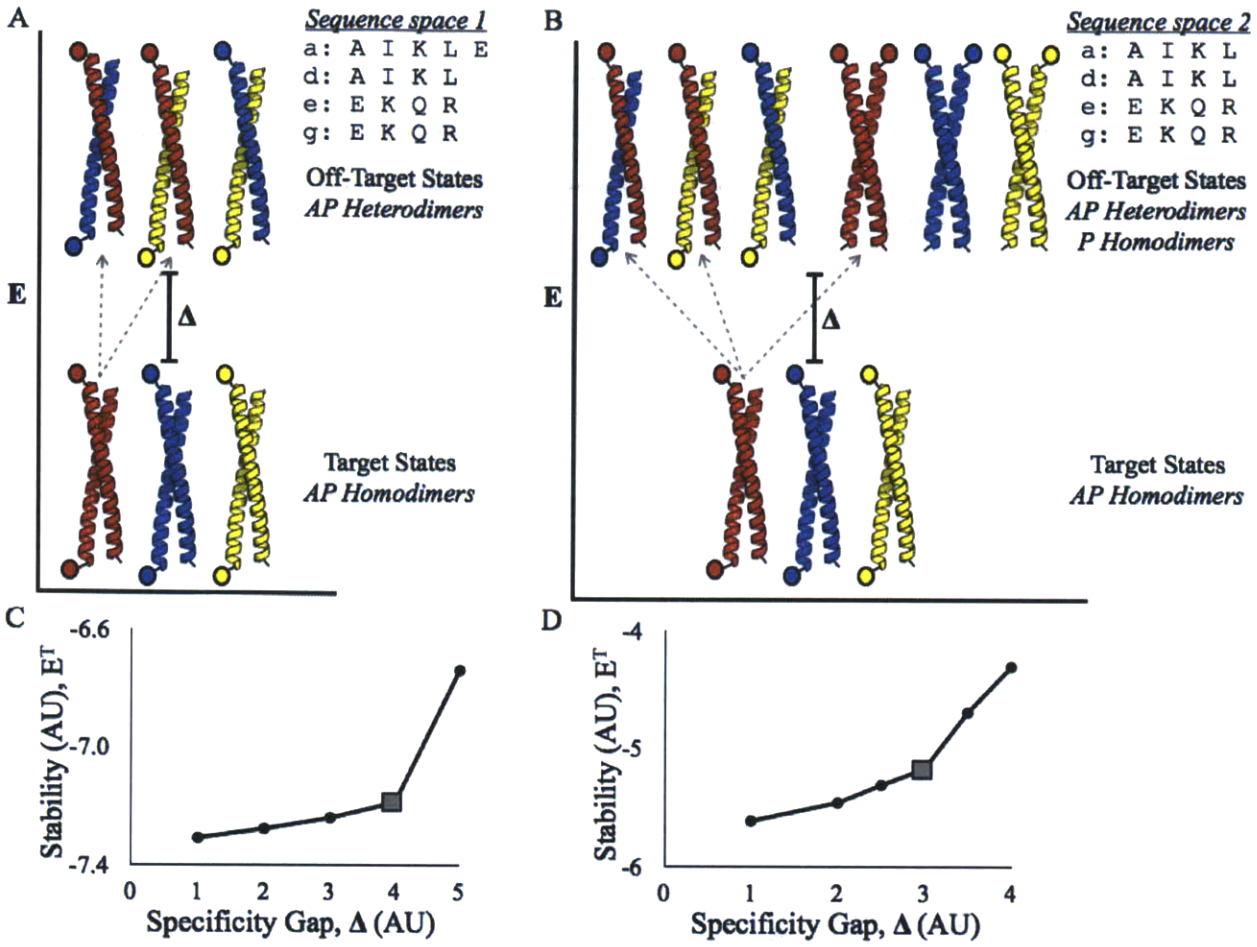


Figure 3-3. Computational design of orthogonal antiparallel homodimers. (A, B) All target and off-target states included in two design calculations. Colors represent distinct sequences, and colored circles indicate the N-terminus of each helix. An energetic constraint, Δ, was enforced between the energy of each target antiparallel homodimer state (E^1, E^2, E^3) and every off-target state that peptide could participate in (examples shown with gray dashed lines). The sequence space used for each design is indicated. Different numbers of off-target states were included for sequence space 1 (A) vs. sequence space 2 (B). (C, D) The total energy $E^T = E^1 + E^2 + E^3$ vs. Δ is plotted for sequence space 1 (C) and sequence space 2 (D). Each value of Δ led to a set of optimized sequences, and the grey squares mark the solutions chosen for experimental testing.

Table 3-5. Ratios of position-specific amino-acid frequencies in antiparallel vs. parallel coiled-coiled dimers.

	Heptad Positions			
	a	d	e	g
A	0.4	1.2	1.4	1.6
D	-	-	1.5	0.5
E	3.1	-	0.6	0.9
I	1.7	3.9	1.8	-
K	0.8	1.4	0.9	1.1
L	0.9	0.8	1.5	1.4
M	1.0	0.8	-	-
N	-	-	1.2	1.4
Q	5.2	1.0	0.8	0.9
R	1.0	-	0.8	0.7
S	1.3	0.8	1.2	1.0
T	-	1.2	0.9	1.2
V	0.6	1.2	2.1	1.2
Y	1.0	0.7	-	-

Antiparallel frequencies obtained from the orientation test set.
Parallel frequencies obtained from the NPS database (14).

CLASSY design was done iteratively, by progressively increasing the energy gap that was imposed between the target antiparallel homodimers and off-target antiparallel heterodimer states. As the gap to off-target states increased, the total predicted stability of the three antiparallel homodimers decreased (Figure 3-3C, 3-3D). This type of stability-specificity tradeoff has been observed previously in the case of parallel dimer design using CLASSY.[32] Two sets of solutions, one from each of the sequence spaces, were rationally chosen based on good stability-specificity tradeoffs. The designs in sequence space 1 are referred to as APH_{i}, APH_{ii}, and APH_{iii}. The designs in sequence space 2 are referred to as $APH_{iv'}$, $APH_{v'}$, $APH_{vi'}$. For each set of designed sequences, parallel and antiparallel homo- and heterodimer states were scored with the

original DFIRE* structure-based model to predict relative energies of target and off-target structures. For the antiparallel homodimers designed in sequence space 1, the predicted energies of all modeled off-target dimers were much higher than the predicted energies for the antiparallel homodimers. The smallest gap, of 0.77 AU, was between the antiparallel homodimer state of APH_{iii} and the parallel heterodimer APH_{iii} would form with APH_i (Figure 3-4A). However, APH_i gap to this state was 1.13 AU. At gaps of this magnitude, DFIRE* predicts the orientation preference of native sequences with an AUC = 1.0. Thus, no additional states were added to the optimization protocol for sequence space 1. For sequence space 2, we observed that one of the parallel homodimers was predicted to be lower in energy than the corresponding antiparallel homodimer (Figure 3-4B). Furthermore, other parallel homodimer states were closer in energy to the antiparallel homodimers than when design was done in sequence space 1.

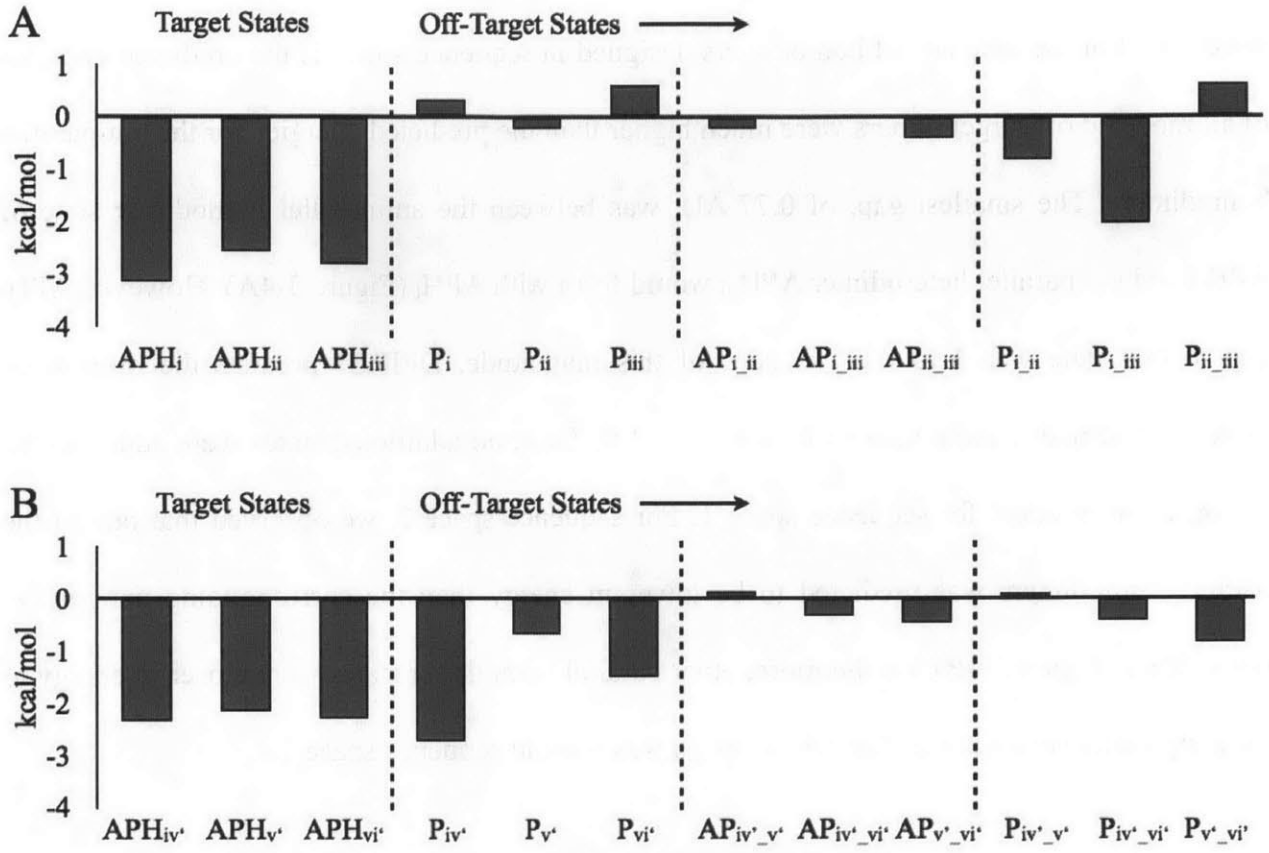


Figure 3-4. DFIRE* scores for design solutions obtained with constraints imposed only against antiparallel heterodimers. DFIRE* scores were calculated for design solutions chosen in sequence space 1 (A) or sequence space 2 (B); see text for details. Subscripts indicate the sequence (for a homodimer) or sequences (for a heterodimer) composing the coiled coil.

To address this, we added parallel homodimer states as off-target states in the optimization protocol used for sequence space 2, and chose a new set of solutions in that space. The final six designed sequences are shown in Table 3-6, with APH_i, APH_{ii} and APH_{iii} resulting from design in sequence space 1, and APH_{iv}, APH_v and APH_{vi} from design in sequence space 2. The two sets of designed sequences were also scored for cross-reactivity using DFIRE*. Predicted energies for all parallel and antiparallel heterodimers that could be formed between sets were significantly larger than predicted energies for the antiparallel homodimer states, with

the smallest energy gap of 0.61 AU between the antiparallel and parallel homodimer states of APH_{iv}.

Table 3-6. Sequences of APH and candidate antiparallel homodimers.

Design	fgabcde	fgabcde	fgabcde	fgabcde	fgabcde	fgabcde	fgabc[a]
APH	KQLE	KELKQLE	KELQAIE	KQLAQLQ	WKAQARK	KKLAQLK	KKLQA
APH_i	KEEKQIE	KELKQIE	KELQAIE	WRLAQLR	KRLQALR	KRKAQKR	E
APH_{ii} (APH2)	KRLKQLE	KRLKQLR	KRKQAKR	WEEAQIE	KELQAIE	KQLAQIR	E
APH_{iii} (APH3)	KRKKQKR	KRAKQLR	KRLQALE	WQLAQIR	KELQAAE	KEEAQIE	E
APH_{iv}	KEKKQLR	KELKQLE	KELQALR	WRLAQIE	KRLQAIR	KRLAQKE	E
APH_v	KRLKQKE	KRKKQLR	KRLQALR	WQLAQIE	KELQAAE	KEAAQLR	E
APH_{vi} (APH4)	KQLKQIE	KRLKQIE	KRLQAKE	WEKAQLR	KELQALR	KKLAQLR	E

[a] Indicates the heptad register.
[b] Some sequences have two names, as described in the text.

3.3.4 *Oligomerization states of designs*

The molecular weights of complexes formed by designed peptides APH_i – APH_{vi} were determined using sedimentation equilibrium analytical ultracentrifugation (see Methods). We anticipate that the APH coiled coils will be used as fusion proteins in many applications, so we did two sets of experiments: one in which the peptides were fused to maltose binding protein (MBP) and one in which they were not. The results are shown in Table 3-7. The data for two designed peptides, APH_{iii} and APH_{vi}, were consistent with these peptides forming homodimers. APH_i was determined to have a molecular weight greater than that expected for a dimer, and no further data were collected on this construct. Single-species fits to APH_{ii} and APH_{iv} gave

molecular weights less than and greater than what was expected for a dimer, respectively. APH_{ii} and APH_{iv} were re-tested at higher concentrations to stabilize higher-order states. At 20 μM, APH_{ii} formed a homodimer, whereas APH_{iv} formed a homotrimer. Further experiments were carried out only on designs APH_{ii}, APH_{iii} and APH_{vi}, which we re-named APH2, APH3 and APH4, respectively (see Table 3-7).

Table 3-7. Molecular weights determined by analytical ultracentrifugation.

Protein	Concentration (μ M)	MW (global fit)/MW (calc.)[a]
APH_i	4, 8, 12	1.7
APH_{ii} (APH2)	4, 7.4, 11	0.76
[b]APH_{ii} (APH2)	20, 40	0.99
APH_{iii} (APH3)	4.5, 9, 14	0.94
[b]APH_{iii} (APH3)	20, 40	1.16
APH_{iv}	7.7, 15.3	1.23
[b]APH_{iv}	20	1.58
APH_{vi} (APH4)	4, 7.4, 12	0.96
[b]APH_{vi} (APH4)	20, 40	1.08

[a] MW (calc.) is the expected dimer mass of each designed coiled coil.

[b] Data collected using interference optics, and a construct not fused to MBP.

3.3.5 *Orientation and orthogonality of designs*

To determine the helix orientation in complexes formed by APH2, APH3, and APH4, we performed disulfide-exchange experiments, and resolved the products of the reactions using HPLC (see Method). Key peaks are labeled in Figure 3-5, which shows changes in the chromatograms over time. For all three designs, starting with a combination of oxidized parallel species and/or reduced peptides, only one oxidized peak was detected at the end of five hours, corresponding to a disulfide-linked antiparallel homodimer. Based on the smallest detectable peak area, we estimate a minimum 10^5-fold preference for forming antiparallel complexes in preference to parallel complexes for all designs.

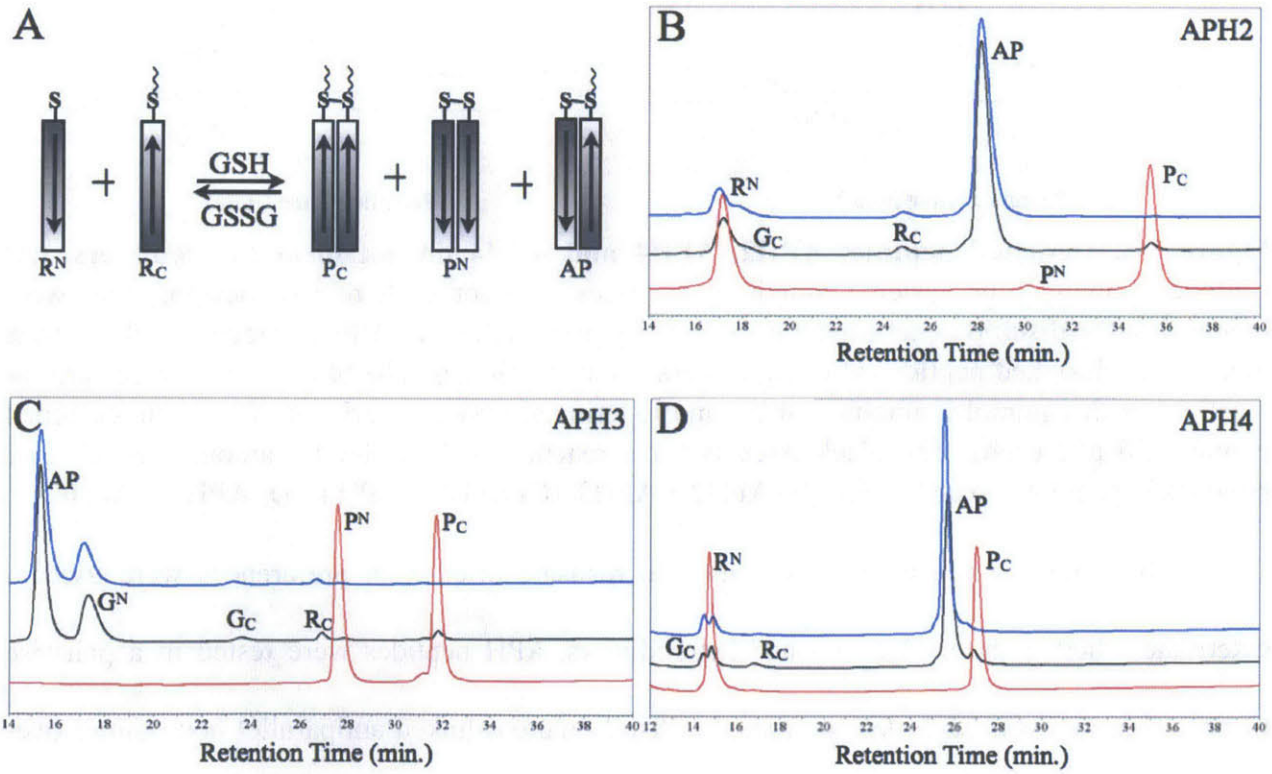


Figure 3-5. Designed peptides APH2, APH3, and APH4 adopt an antiparallel helix orientation. (A) Schematic view of the assay. Arrows indicate helix direction from N to C terminus. The wavy line indicates two amino acids added to the designed sequence to change peptide retention times (APH2 = YY, APH3 = QW, APH4 = YY). S represents the sulfur atom in

cysteine residue(s). (B, C, D) HPLC chromatograms show the results for the disulfide-exchange reactions upon mixing equimolar amounts of N-terminal and C-terminal cysteine variants of each design sequence (20 μM each). The reactions were quenched at 0 minutes (red), 15 minutes (black), or 5 hours (blue). Peaks are labeled according to the scheme shown in panel A, with G indicating a glutathione adduct.

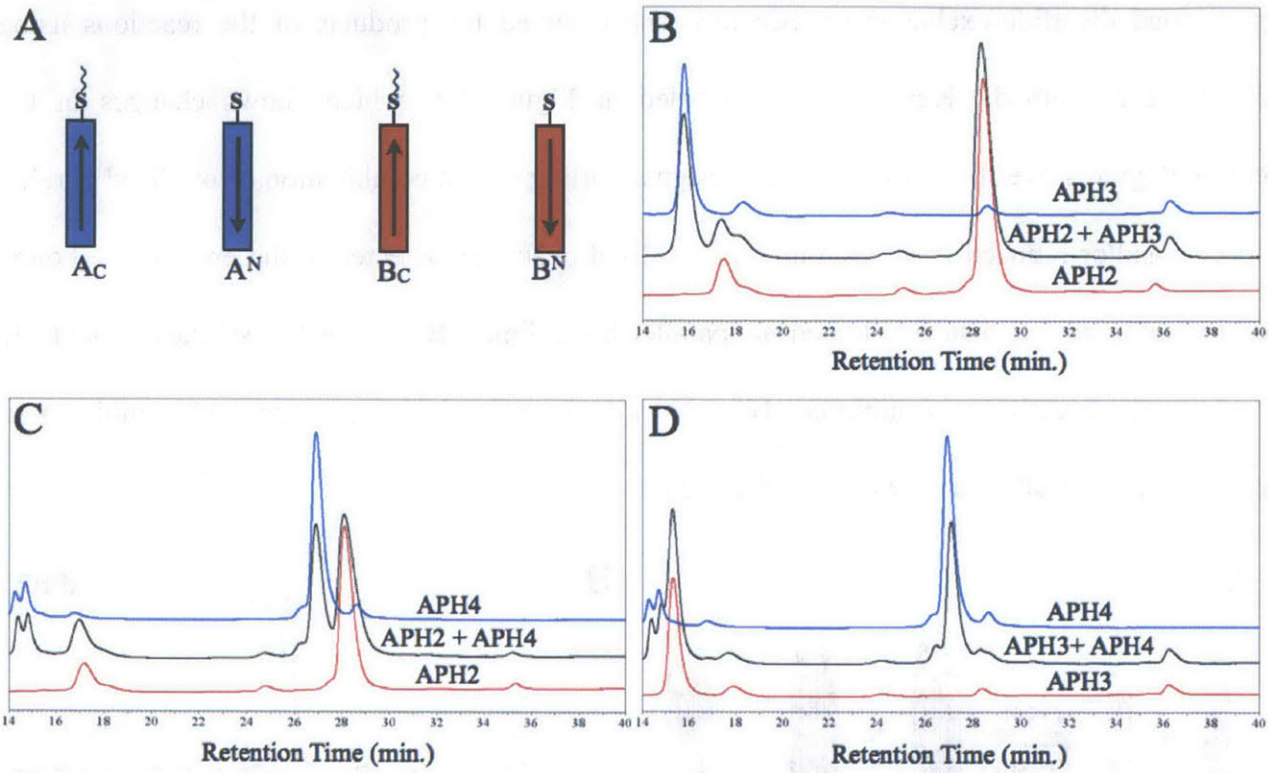


Figure 3-6. Designed peptides APH2, APH4 and APH4 do not form heterodimers. (A) Cartoon showing four cysteine-containing peptides, two for each of two designs, that were included in the disulfide-exchange cross-reactivity assay. (B, C, D) HPLC traces for all pairwise mixtures of designed peptides after equilibration for 15 minutes. The blue and red traces are for reactions with equimolar amounts of N- and C-terminal cysteine variants of a single designed peptide (20 μM each). The black trace is for a reaction with equimolar amounts of all four peptides in panel A (20 μM each). (B) APH2 + APH3, (C) APH2 + APH4, (D) APH3 + APH4.

The same constructs that were used to measure orientation preferences were used to determine whether the designs formed heterodimers. APH peptides were tested in a pairwise manner (Figure 3-6). Each design formed a disulfide cross-linked antiparallel homodimer over time, but we did not detect any disulfide bond formation between any pairs of designed peptides. Each design was additionally measured for cross reactivity with the antiparallel homodimer-

forming peptide APH, in a pairwise manner (Figure 3-7). No design showed any detectable cross-reactivity with APH, in either orientation, extending the number of orthogonal antiparallel homodimers from three to four.

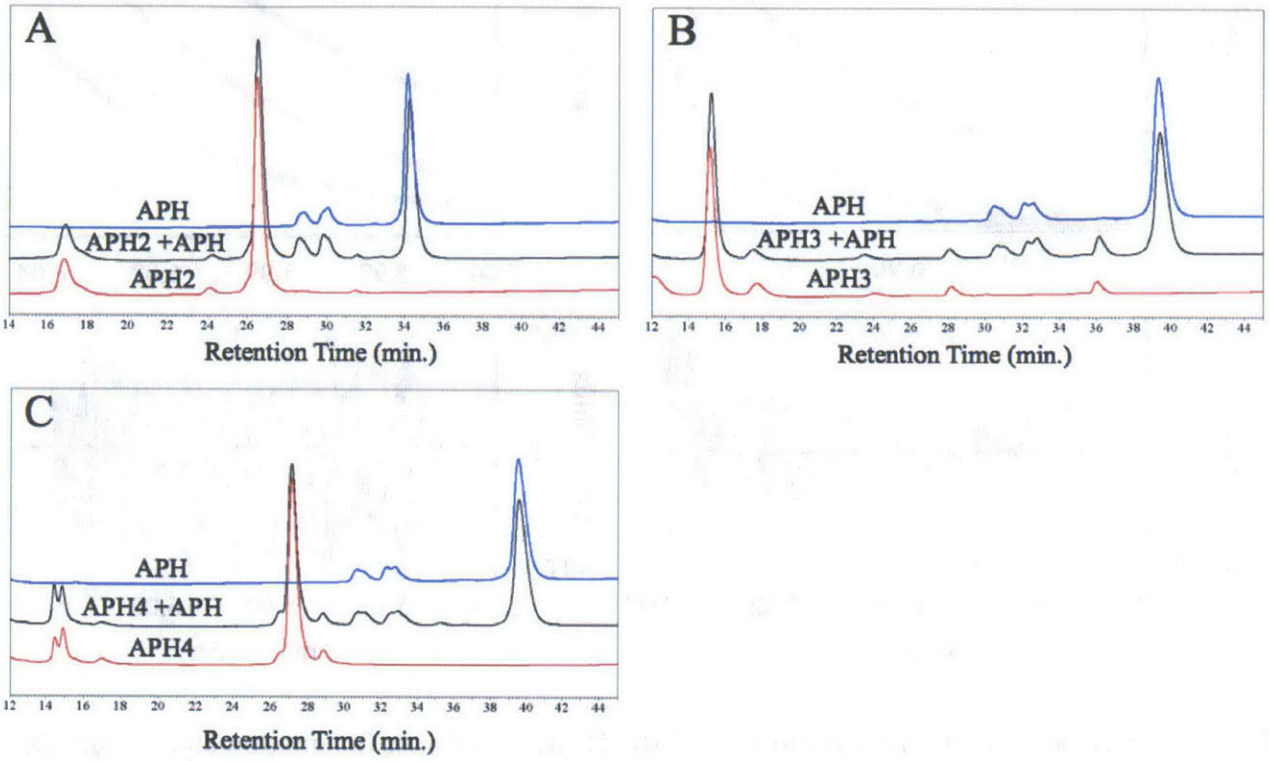


Figure 3-7. Designed peptides APH2, APH3, and APH4 do not heterodimerize with APH. (A, B, C) HPLC traces for all pairwise combinations of APH with designed sequences, with experimental conditions as for Figure 3-6. The blue and red traces are for equimolar mixtures of N- and C-terminal cysteine variants of APH (blue) or APH2, APH3 or APH4 (red) (20 μM each). The black trace is for a mixture of four peptides, APH and the indicated design, each modified at the N- or C-terminus with a cysteine residue (20 μM each). (A) APH + APH2, (B) APH + APH3, (C) APH + APH4.

To determine whether mixtures of more than two APH coiled coils formed complexes other than the expected dimers, MBP fusions of all four APH peptides were mixed at 20 or 40 μM of each APH design and analyzed by sedimentation equilibrium ultracentrifugation (as done for individual MBP fusion proteins, see Methods). The ratio of the fitted mass to the dimer mass was 0.91, with good fit quality (representative data in Figure 3-8), indicating that dimers formed

as expected and no higher-order species were present in a mixture of all four APH fusions.

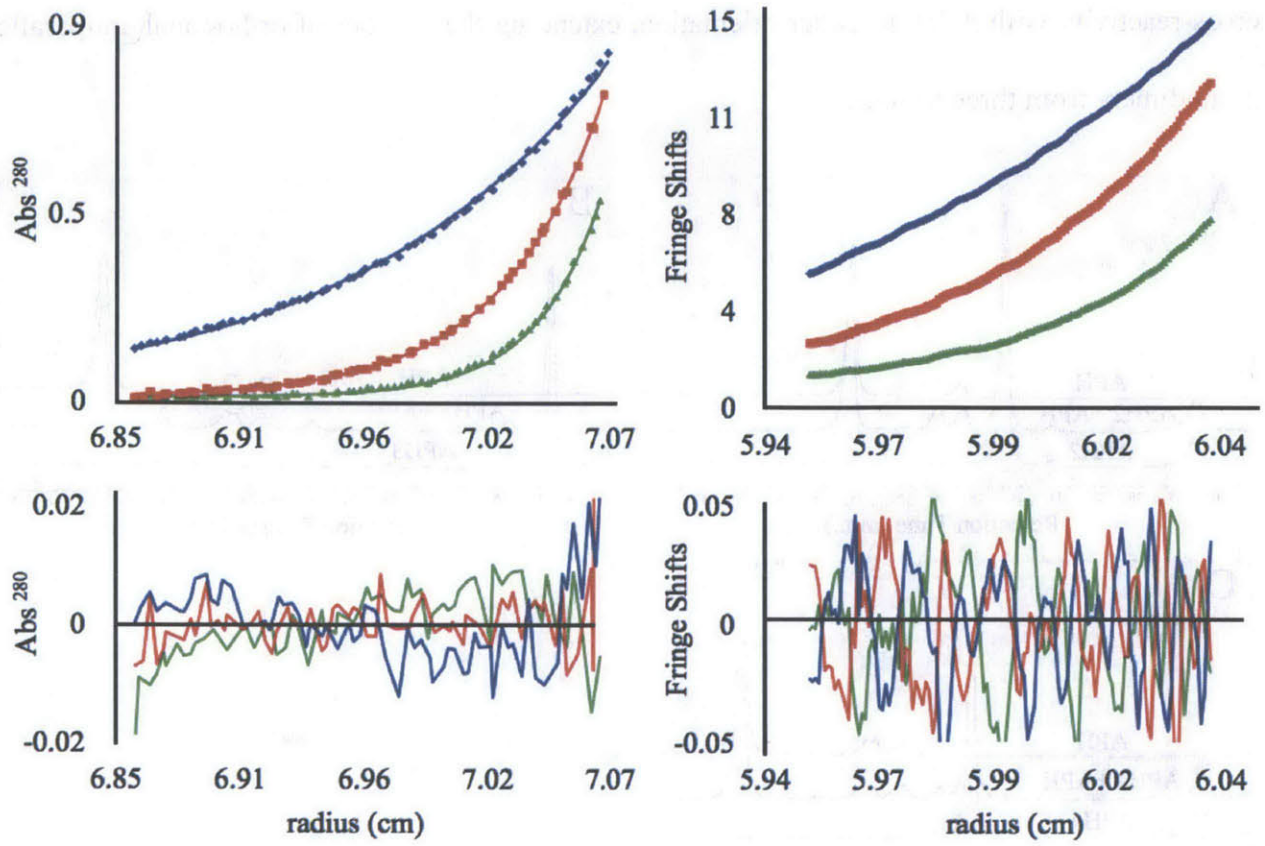


Figure 3-8. Analytical ultracentrifugation data. Representative sedimentation equilibrium data (points) and fits (lines) for absorbance (top-left) and interference optics (top-right) are shown for constructs fused to maltose binding protein (MBP). Plots of the residuals for the fits are shown at the bottom. The fits were obtained with data from three spin speeds, and at least two concentrations. The left plot shows data collected for APH3 fused to MBP. The right plot shows data collected for a mixture of APH, APH2, APH3, and APH4 at 20 μM each. In both plots, blue corresponds to 10,200 rpm, red corresponds to 16,300 rpm, and green corresponds to 20,400 rpm.

3.3.6 *Helicity and thermal stability*

We measured the circular dichroism (CD) spectra of the three designed peptides APH2, APH3, and APH4, using the N-terminal cysteine constructs in a reduced state. Each construct contained 65 residues, of which 43 correspond to the designed coiled-coil sequence (Table 3-3).

Our APH construct contained 66 residues, of which 44 correspond to the APH sequence. The CD spectra of all three designs were characteristic of coiled coils, with distinct minima at 208 and 222 nm (Figure 3-9A). The mean residue ellipticity (MRE) of the designed peptides was similar to that of APH, which is longer by one residue in the coiled-coil region. Thermal denaturation experiments established that all designs unfolded cooperatively, which is a characteristic property of coiled coils (Figure 3-9B). The thermal stabilities (T_m) of the designs at 20 μM ranged from 47.4 °C for APH2, to 59.3 °C for APH4 and 78.3 °C for APH3, with APH3 being slightly less stable then APH which had a T_m of 79.3 °C. All melts were reversible. Upon re-cooling, all peptides regained ≥95% of the original MRE and fits of re-folding curves gave melting temperatures within 1.5 °C of values obtained from the denaturing curves.

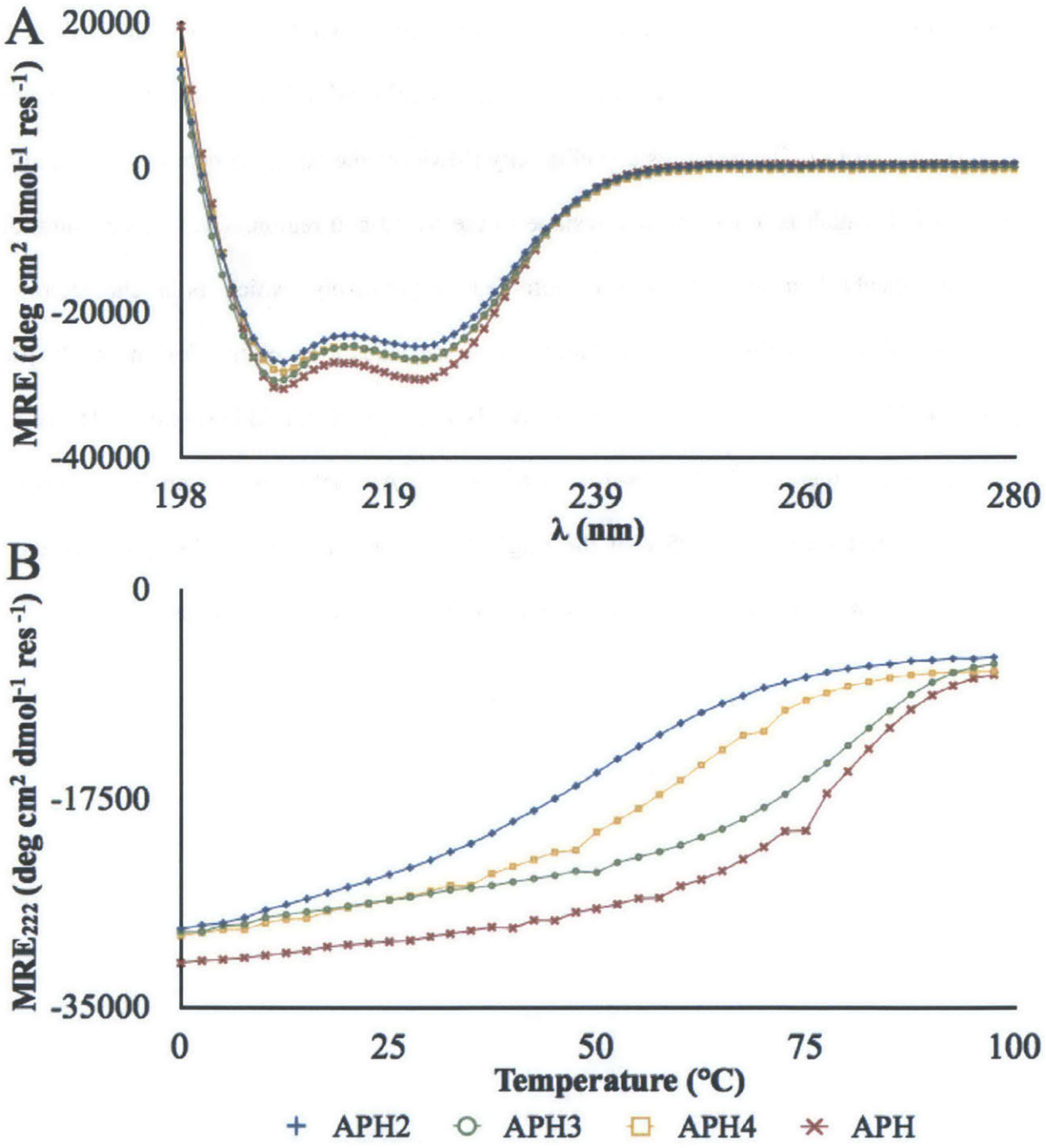


Figure 3-9. Circular dichroism spectra and thermal denaturation curves. (A) CD spectra and (B) thermal denaturation curves measured at 25 °C in PBS with 1 mM DTT. APH (red), APH2 (blue), APH3 (green) and APH4 (orange).

3.4 Discussion

An expanded toolkit of coiled-coil interaction parts would be of great utility in protein engineering. Many papers have reported the successful design of coiled-coil structures of diverse topologies, but apart from parallel dimers, the number of biochemically characterized complexes of any one type is limited.[3,6,55] Designing coiled coils *de novo* is complicated by the fact that different coiled-coil topologies have similar sequence requirements, and small sequence changes can alter coiled-coil structure. For these reasons it is often necessary to explicitly consider competing states in the design process.[25,28,32]

Treating off-target states in computational protein design can be costly, particularly when there are many such states that must be modeled. One strategy is to incorporate a design element known to strongly destabilize a set of off-target topologies, to reduce the number of off-target states that must be modeled. For instance, Thomas et al. observed that the *de novo* design of parallel heterodimeric coiled coils composed entirely of isoleucine and leucine cores did not reliably destabilize higher-order states.[55] But the same design strategy in the background of a single asparagine-asparagine interaction – which was known from prior work to favor parallel dimer states over higher-order states – consistently gave dimeric assemblies.[28] Unfortunately, incorporating simple design elements that reliably destabilize all off-target topologies, in all sequence contexts, is not feasible. Exceptions have been reported for even the most thoroughly studied coiled-coil structural specificity determinants,[3,55] and for many coiled-coil topologies, the sequence-structure relationship is not well understood. Of relevance for this work, our current understanding of interactions that favor antiparallel over parallel helical alignments is very

incomplete. Oakley et al. showed that in analogy to the role of asparagines favoring dimers over higher order states, paired asparagines can be introduced at opposing *a* and *d'* positions to favor an antiparallel helix alignment.[56] Additionally, McClain et al. demonstrated that charge-charge interactions at *e* and *g* positions across the interface can impart an antiparallel vs. parallel preference.[57] Lastly, isoleucine at a *d* heptad position paired with an alanine residue at an *a* heptad position is thought to contribute to a bias towards the antiparallel state of APH over its parallel homodimer state, in combination with charge-charge interactions at *e* and *g* heptad positions.[21] But these types of interactions do not adequately explain the orientations of native coiled coils.

Modeling off-target states explicitly and including them in the design process provides a broadly applicable mechanism for engineering specificity. In this work, we used explicit negative design to disfavor antiparallel heterodimer states by imposing energy gaps between antiparallel homo and heterodimers. Most of the sequence elements in our APH designs that disfavored antiparallel heterodimerization within a design set involved charged residues predicted to participate in repulsive interactions in heterodimer states. For example, all antiparallel heterodimer states contained *a*-to-*e*' and *d*-to-*g*' charge-charge repulsions between lysine or arginine residues. Designs from sequence space 1 additionally contained *a*-to-*e*' charge-charge repulsions between glutamate residues. These core-to-edge charge-charge repulsions are the most destabilizing weights available to the antiparallel CE DFIRE* model in the design sequence spaces chosen, with lysine at *d* to arginine at *g*' being the most destabilizing.

The design strategies that led to destabilization of parallel homodimers differed in sequences spaces 1 and 2. In sequence space 1, we allowed glutamate at *a* positions, and all

designed sequences included this element. In fact, we identified a motif consisting of two glutamate residues at *a* and *g*, and a lysine at *d'* with an arginine at *e'* on the opposing helix that was present in all of the sequence space 1 designs (Figure 3-10). Interactions between residues in this motif contain the first and fourth most favorable weights available in the CE DFIRE* model in sequence space 1, such that the motif is predicted to contribute strongly to antiparallel homodimer stability. Interestingly, in a parallel homodimer, the residues of this motif form unfavorable interactions sufficient to provide a large energy gap between parallel and antiparallel states. This can be demonstrated by modeling an artificial homodimer that includes the motif embedded in a poly-alanine sequence. Due to symmetry of the homodimer, this results in two copies of the motif in the structure. Scoring parallel and antiparallel homodimeric structures with this sequence using DFIRE* revealed a significant preference of 1.64 energy units for the antiparallel state (poly-alanine alone has a preference of 0.14 energy units for the antiparallel state using this model). Thus, in sequence space 1, charge networks that stabilize the antiparallel state lead to substantial destabilization of parallel homodimers, without explicit negative design. The situation was different in sequence space 2, which did not include glutamate residues at *a* positions. In this sequence space, designing antiparallel homodimers while disfavoring heterodimers did not automatically lead to large energy gaps to parallel homodimer states for all sequences (see Figure 3-2B); it was necessary to include parallel structures as off-target states in the optimization problem. Doing so led to sequences that placed more isoleucines at *d* heptad positions to favor antiparallel over parallel homodimers. For example, of the three sequences originally chosen in sequence space 2, two sequences had one isoleucine residue at a *d* position, while one sequence had no isoleucine residues at all. After placing constraints to the parallel

homodimer state, all design sequences contained at least one isoleucine residue at the *d* heptad position, with two sequences containing two isoleucine residues at the *d* heptad position. Each designed isoleucine at a *d* position leads to a *d-d'* isoleucine pairing across the coiled-coil interface in the parallel homodimer state. As previously mentioned, this interaction destabilizes parallel dimers. The effect is captured in our models: isoleucine at *d-d'* is the fourth most destabilizing weight for parallel dimers in sequence space 2.

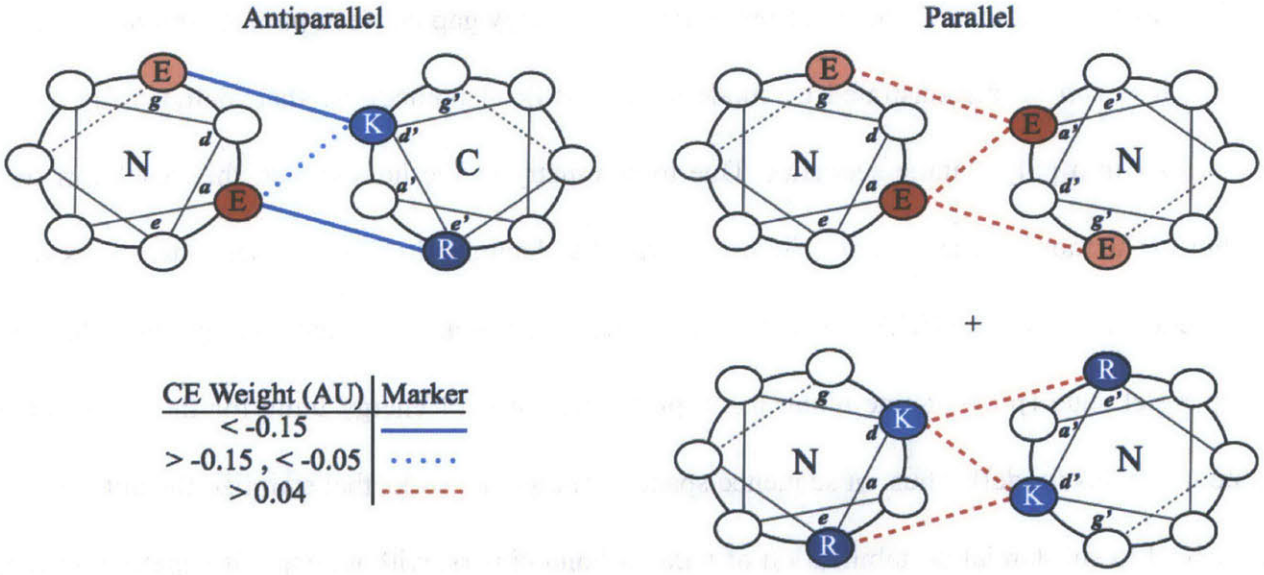


Figure 3-10. Sequence space 1 motif that favors AP helix orientation. Helical-wheel diagrams show a motif that was included in all three of the designed sequence space 1 AP homodimers. Positively and negatively charged amino acids are shown in blue and red, respectively. The motif residues make strongly favorable interactions according to the CE DFIRE* model in the AP state, but repulsive interactions in the parallel state, as indicated with colored and solid/dashed lines.

Explicit consideration of off-target states requires enumerating and modeling the relevant competing states. We successfully used this strategy to destabilize antiparallel heterodimer states in sequence spaces 1 and 2, and to destabilize parallel homodimers when designing in sequence space 2. But we did not explicitly model formation of higher-order assemblies, and as a result, oligomers larger than dimers were formed by designs APH_i and APH_{iv}. Modeling higher-order coiled coils is challenging due to the many different topologies that are possible. Each helix pair

can be antiparallel or parallel, heteroassemblies can form with different stoichiometries, and the geometry of helix associations can vary in subtle ways.[58,59] It is therefore difficult to include a comprehensive set of competing states and, even if such a set could be generated, the computational modeling costs for considering all possibilities explicitly would be high.

One approach to disfavoring higher-order states could be to include just a small number of trimer and tetramer topologies in the calculations. Adding representative off-target structures would minimally alter the computational complexity of the current design framework, yet might lead to broader destabilization of additional higher-order states. Indeed, our study provided such an example where specificity was obtained against states that were not explicitly modeled, possibly due to constraints on specificity against related states. The design solutions from sequence space 1 were predicted to not form heterodimers with design solutions from sequence space 2, despite not being explicitly constrained during optimization. We hypothesize that this occurred because the consideration of many off-target dimer states gave rise to interfaces with charge patterns low in symmetry, as well as hydrophobic cores with unique geometries due to the placements of beta-branched residues in the core. As a result, the probability of cross-reacting with another sequence to form dimers was low.

Considering just a few higher-order states may also have the effect of reducing or removing design features known to favor higher-order states generally. For example, isoleucines at *d* heptad positions are known to favor parallel trimer and tetramer states in preference to parallel dimer states.[3,27] Yet isoleucines at *d* heptad positions also favor antiparallel dimers over parallel dimers, and were included in many of our designs for this reason, as discussed above (also see Table 3-6). Interestingly, in native coiled coils isoleucines are approximately four-fold

more common in AP dimers than in P dimers (Table 3-5). Isoleucines at *d* heptad positions that were included in the design to favor antiparallel dimers might have inadvertently favored higher-order assembly, which was not treated in the model. A constraint to disfavor just a few trimers or tetramers might be sufficient to limit the use of this sequence element, or to drive inclusion of compensating elements that are poorly accommodated in higher-order assemblies.

A significant obstacle to including even a few higher-order states in design is the small amount of structural data available for coiled-coil trimers, tetramers, and higher-order assemblies of a specific toplogy.[37] Benchmarking the predictive power of models using experimental data is important for determining the limitations of any model, and is useful for setting meaningful energy cutoffs in design calculations. As the number of solved structures for higher-order states increases over time, our ability to rigorously assess and validate models will improve, as seen by the development of models like LOGICOIL.[9] It should be noted that LOGICOIL does predict that APH, APH3, and APH4 will form dimers, though it predicted that these would form parallel dimers. LOGICOIL does not score inter-chain terms, which were designed to be the main determinants of orientation preference in the APH sequences and may be the reason why LOGICOIL assigns the wrong orientation to these sequences.

The new APH designs have many desirable properties for synthetic biology and materials science. The peptides use well-known sequence features to establish orientation bias and orthogonality that should aid in manipulating them. For instance, it appears most target states are stabilized by salt-bridge interactions across the interface, and aliphatic residues at core *a* and *d* heptad positions (Figure 3-11). Off-target states appear to be destabilized by charge repulsions along the interface, and by steric clashes between beta-branched residues at core heptad

positions. The surface residues of all APH designs were engineered to be passive, and may provide useful positions for adding novel functions or modulating stability.[52,60,61] The designed structures also provide users with a range of thermal stabilities. Finally, the designs are orthogonal to each other when used in pairwise or higher-order combinations. Proteins with this property have been highly sought for many applications in synthetic biology and are thought to be one of the limiting reagents slowing progress in this field.[19,62] It should also be noted that these designs could be used as off-target states in future design studies using the CLASSY framework, allowing for the extension of this set. In conclusion, the antiparallel homodimer sequences represent a significant expansion to the coiled-coil toolkit, which is currently dominated by parallel dimers, and thus may find application in many molecular engineering projects.[2,3]

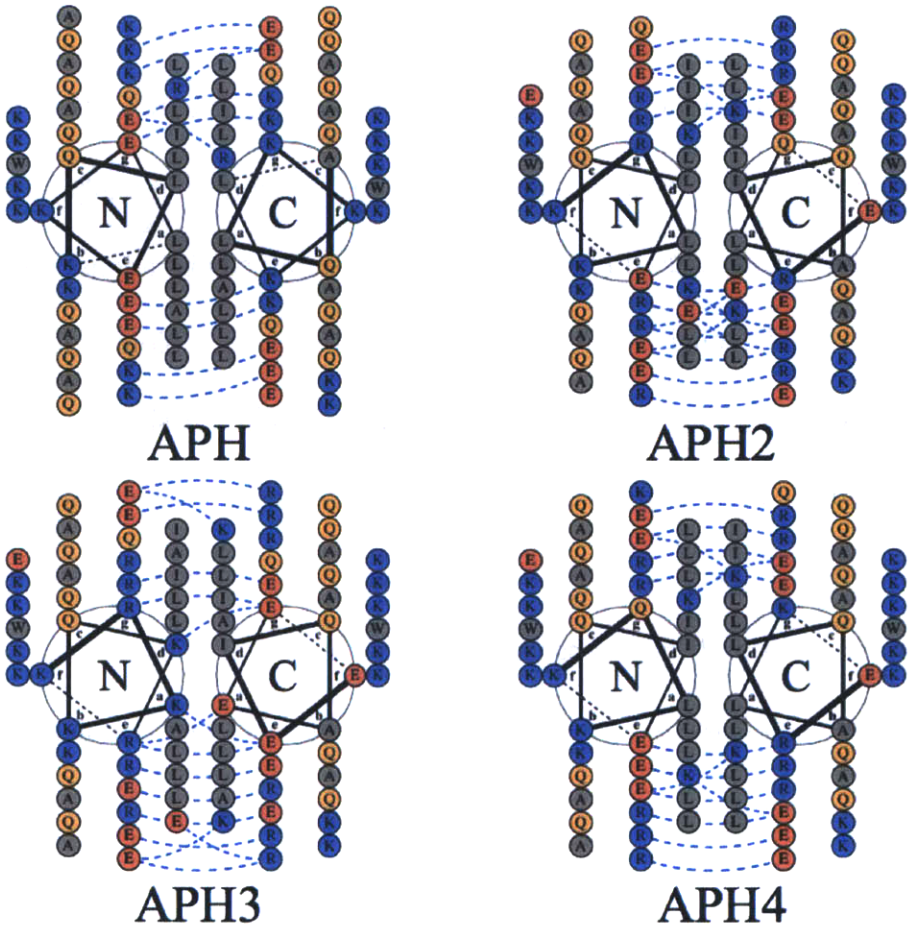


Figure 3-11. Helical-wheel diagrams of APH, APH2, APH3, and APH4 as antiparallel homodimers. Positively and negatively charged amino acids are shown in blue and red, respectively, with non-charged polar residues in orange and hydrophobic residues in grey. Potentially attractive salt bridges are shown as dashed lines. Sequences start at an *f* position and end at an *e* position. Diagrams were generated using DrawCoil 1.0, http://www.grigoryanlab.org/drawcoil.

3.5 Acknowledgements

We thank members of the Keating lab, especially R. Rezaei Araghi for performing mass spectrometry analysis of peptides, as well as J. B. Kaplan and K. Hauschild for experimental advice. This work used the MIT Bioinstrumentation Facility and we are grateful to D. Pheasant for analytical ultracentrifugation support. Lastly, we thank C. M. Kougentakis and V. Xue for

careful reading of the manuscript. Funding was from a National Science Foundation Graduate Research Fellowship awarded to C. N., and from NSF award MCB-0950233 (supporting experimental applications) and NIH award GM67681 (supporting computational design method development) to A. K. We used computer resources provided by National Science Foundation award DBI-0821391.

3.6 References

1.Purnick, P. E.; Weiss, R. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 410–422.

2.Thompson, K. E.; Bashor, C. J.; Lim, W. A.; Keating, A. E. *ACS Synth. Biol.* **2012**, *1*, 118-129.

3.Fletcher, J. M.; Boyle, A. L.; Bruning, M.; Bartlett, G. J.; Vincent, T. L.; Zacci, N. R.; Armstrong, C. T.; Bromley, E. H. C.; Booth, P. J.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N. *ACS Synth. Biol.* **2012**, *1*, 240–250.

4.Wolf, E.; Kim, P. S.; Berger; B.; *Protein Sci.* **1997**, *6*, 1179-1189.

5.Rackham, O. J.; Madera, M.; Armstrong, C. T.; Vincent, T. L.; Woolfson, D. N.; Gough, J. *J. Mol. Biol.* **2010**, *403*, 480-493.

6.Reinke, A. W.; Grant, R. A.; Keating, A. E. *J.Am.Chem.Soc.* **2010**, *132*, 6025-6031.

7.Lupas; A. *Trends Biochem. Sci.* **1996**, *21*, 375-382.

8.Woolfson, D. N. *Adv. Protein Chem.* **2005**, *70*, 79-112.

9.Vincent, T. L.; Green, P. J.; Woolfson, D. N. *Bioinformatics* **2013**, *29*, 69-76.

10.Woolfson, D.N.; Alber, T. *Protein Sci.* **1995**, *4*, 1596–1607.

11.Wolf, E.; Kim, P. S.; Berger; B.; *Protein Sci.* **1997**, *6*, 1179-1189.

12.Armstrong, C.T.; Vincent, T. L.; Green, P. J.; Woolfson, D. N. *Bioinformatics* **2011**, *27*,

1908–1914.

*13.*Mahrenholz, C. C.; Abfalter, I. G.; Bodenhofer, U.; Volkmer, R.; Hochreiter, S. *Mol. Cell. Proteomics.* **2011**, *10.*

14.Trigg, J.; Gutwin, K.; Keating, A. E.; Berger, B. *PLoS ONE* **2012**, *6*, e23519.

15.Wolfe, S. A.; Grant, R. A.; Pabo, C. O. *Biochemistry* **2003**, *42*, 13401–13409.

16.Bashor, C. J.; Helman, N. C.; Yan, S.; Lim, W. A. *Science* **2008**, *319*, 1539-1543.

17.Lanci, C. J.; MacDermaid, C. M.; Kang, S.; Acharya, R.; North, B.; Yang, X.; Qiu, X. J.; DeGrado, W. F.; Saven, J.; G. *Proc. Natl. Acad. Sci.* **2012**, *109*, 7304–7309.

18.Shlizerman, C.; Atanassov, A.; Berkovich, I.; Ashkenasy, G.; Ashkenasy, N. *J.Am.Chem.Soc.* **2010**, *132*, 5070–5076.

19.Gradišar, H.; Božič, S.; Doles, T.; Vengust, D.; Hafner-Bratkovič, I.; Mertelj, A.; Webb, B.; Šali, A.; Klavžar, S.; Jerala, R. *Nat. Chem. Biol.* **2013**, *9*, 362-366.

20.Lumb, K. J.; Carr, C. M.; Kim, P. S.; *Biochemistry* **1994**, *33*, 7361-7367.

21.Gurnon, D. G.; Whitaker, J. A.; Oakley, M. G. *J.Am.Chem.Soc.* **2003**, *125*, 7518-7519.

22.Taylor, C. M.; Keating, A. E.; *Biochemistry* **2005**, *44*, 16246-16256.

23.Gradišar, H.; Jerala, R. *J. Pept. Sci.* **2010**, *17,* 100-106.

24.Reinke, A.W.; Baek, J.; Ashenberg, O.; Keating, A.E. *Science* **2013**, *340*, 730-734.

25.Havranek, J. J; Harbury, P. B. *Nat. Struct. Biol.* **2003**, *10*, 45-52.

26.O'Shea, E. K.; Lumb, K. J.; Kim, P.S. *Curr. Biol.* **1993**, *3*, 658-667.

27.Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T. *Science* **1993**, *262,* 1401-1407.

28.Lumb, K. J.; Kim, P. S. *Biochemistry* **1995**, *34*, 8642-8648.

29.McClain, D. L.; Gurnon, D. G.; Oakley, M. G. *J. Mol. Biol.* **2002**, *324*, 257-270.

30.London, N.; Ambroggio, X. *J. Struct. Biol.* **2013**, *185*, 136-146.

31.Davey, J. A.; Chica, R. A.; *Protein Sci.* **2012**, *21*, 1241-1252.

32.Grigoryan, G.; Reinke, A. W.; Keating, A. E. *Nature* **2009**, *458*, 859-864.

33.Negron, C.; Keating, A. E. *Methods Enzymol.* **2013**, *523*, 171-190.

34.Reinke, A. W.; Grigoryan, G. Keating, A. E. *Biochemistry* **2010**, *49*, 1985–1997.

35.Chen, T. S.; Reinke, A. W., Keating, A. E. *J. Mol. Biol.* **2011**, *408*, 304-320.

36.Yang, Y.; Zhou, Y. *Proteins* **2008**, *72*, 793-803.

37.Testa, O. D.; Moutevelis, E.; Woolfson, D. N. *Nucleic Acids Res.* **2009**, *37*, 315-322.

38.Crick, F. H. C. *Acta Crystallogr.*, **1953** *6*, 685-689.

39.Grigoryan, G.; DeGrado, W. F. *J. Mol. Biol.* **2011**, *405*, 1079-1100.

40.Kuhlman, B.; Baker, D. *Proc. Natl. Acad. Sci.* **2000**, *97*, 10383–10388.

41.Grigoryan, G.; Zhou, F.; Lustig, S. R.; Ceder, G.; Morgan, D.; Keating, A. E. *PLoS Computational Biology* **2006**, *2*, 551-563.

42.Hahn, S.; Ashenberg, O.; Grigoryan, G.; Keating, A. E. *J. Comput. Chem.* **2010**, *31*, 2900-2914.

43.IBM Corp. IBM ILOG CPLEX Optimization Studio **2012**. Version 12.5.

44.Hoover, D. M.; Lubkowski, J. *Nucleic Acids Res.* **2002**, *30*, e43.

45.Edelhoch, H. *Biochemistry* **1967**, *6*, 1948-1967.

46.Schuck, P. *Anal. Biochem.* **2003**, *320*, 104-124.

47.Vistica, J.; Dam, J.; Balbo, A.; Yikilmaz, E.; Mariuzza, R. A.; Roualt, T. A.; Schuck, P. *Anal. Biochem.* **2004**, *326*, 234-256.

48.Laue, T. M.; Shah, B. D.; Ridgeway, T. M.; Pelletier, S. L. In *Analytical ultracentrifugation in*

biochemistry and polymer science (eds. S.E. Harding et al.), pp. 90–125. The Royal Society of Chemistry, Cambridge, UK.

49.John, D. M.; Weeks, K. M. *Protein Sci.* **2000**, 9, 1416-1419.

50.Apgar, J. R.; Gutwin, K. N.; Keating, A. E. *Proteins* **2008**, *72*, 1048-1065.

51.Zhou, F.; Grigoryan, G.; Lustig, S. R.; Keating, A. E.; Ceder, G.; Morgan, D. *Phys. Rev. Lett.* **2005**, 95, 148103.

52.Li, Y.; Kaur, H.; Oakley, M.G. *Biochemistry* **2008**, *47*, 13564-13572.

53.Mason, J.M.; Hagemann, U. B.; Arndt, K.M. *Biochemistry* **2009**, *48*, 10380-10388.

54.Straussman, R.; Ben-Ya'acov, A.; Woolfson, D. N.; Ravid, S. *J. Mol. Biol.* **2007**, *366*, 1232-1242.

55.Thomas, F.; Boyle, A. L.; Burton, A. J.; Woolfson, D. N. *J.Am.Chem.Soc.* **2013**, *135*, 5161−5166.

56.Oakley, M. G.; Kim, P.S *Biochemistry* **1998** 37, 12603-12610.

57.McClain, D. L.; Woods, H. L.; Oakley, M. G. *J.Am.Chem.Soc.* **2001** *123,* 3151-3152.

58.Deng, Y.; Zheng, Q.; Liu, J.; Cheng, C. S.; Kallenbach, N. R.; Lu, M. *Protein Sci.* **2007**, *16*, 323-328.

59.Liu, J.; Zheng, Q.; Deng, Y.; Li, Q.; Kallenbach, N. R.; Lu, M. *Biochemistry* **2007**, *46*, 14951-14959.

60.Dahiyat, B. I.; Gordon, B.; Mayo, S. L. *Protein Sci.* **1997**, *6*, 1333–1337.

61.Kaplan, J. B.; Reinke, A. W.; Keating, A. E. *Protein Sci.* **2014**, 23, 940-953.

62.Kapp, G. T.; Liu, S.; Stein, A.; Wong, D. T.; Reményic, A.; Yeh, B. J.; Fraser, J. S.; Taunton, J.; Lim, W. A.; Kortemme, T. *Proc. Natl. Acad. Sci.* **2012**, *109*, 5277–5282.

Chapter 4

Conclusions and Future Directions

Computational protein design has successfully produced several PPIs that are orthogonal to native PPIs (Kortemme et al., 2004; Sammond et al., 2010; Kapp et al., 2012). However, using computational protein design in *de novo* design of orthogonal PPI pairs has never been attempted. In this thesis, I described a proof-of-concept application of computational protein design that produced a set of *de novo* orthogonal antiparallel homodimeric coiled coils. In this chapter, I will discuss how our computational design framework CLASSY is useful for the design of orthogonal PPIs. I will then describe strategies for improving our computational models, and present a strategy to more rapidly screen large sets of candidate orthogonal coiled coils experimentally. Lastly, I will speculate on possible applications that could benefit from large sets of orthogonal coiled coils.

4.1 Designing orthogonal sets with CLASSY

There are several advantages to designing orthogonal PPIs with CLASSY. For example, CLASSY can be used with a diverse set of energy functions (Grigoryan et al., 2006). This allows

it to be coupled with an energy function that is empirically observed to have good performance on the system of interest. In the CLASSY framework, it is possible to design multiple orthogonal pairs at once, which is a problem for other orthogonal design strategies (Kortemme et al., 2004). CLASSY can be used to perform rapid searches through sequence space, and if a solution is found, this is guaranteed to be the minimum energy solution for the energy function used (Kingsford et al., 2005). This facilitates the evaluation of energy functions, because any failed design sequences can be attributed directly to the energy function, and not the search algorithm. This is not true of other multi-state design strategies (Havranek & Harbury, 2003). Additionally, users can systematically evaluate the tradeoff between stability of the target state and orthogonality with respect to off-target states. Lastly, it has been shown that incorporating backbone flexibility in molecular models is crucial for recapitulating certain properties of proteins (Ollikainen et al., 2013). Yet, most protein design algorithms are built to perform side-chain optimization on a single fixed backbone (Gordon et al., 1999). This is often due to the computational time added in exploring a larger structure space. Cluster expansion provides a way to reduce this computational time by deriving energies that approximate scoring on multiple backbones (Apgar et al., 2009).

As a proof a principle, we experimentally characterized a set of six orthogonal antiparallel homodimeric coiled coils designed using CLASSY. These designs represent the first computationally designed antiparallel coiled coils, and are the only antiparallel coiled coils designed and experimentally tested for orthogonality. Despite the success of three of these designs, two of the design sequences did not adopt their target structure. However, several strategies exist for improving the models used in this work.

4.2 Improving models for the design of antiparallel coiled coils

4.2.1 *Incorporating coupling energies*

An accurate energy function is a critical component in any computational protein design problem. In this work, a cluster-expanded model of a modified version of the statistical potential DFIRE was used. This model's performance on predicting the orientation preference of known coiled-coil dimers was suboptimal, with an area under the curve of 0.84. An important factor when predicting orientation preference is the accurate scoring of interactions among residues at the core *a* and *d* heptad positions (Apgar et al., 2008). As a result, interactions between core residues are a potential issue for the cluster-expanded energy functions used in CLASSY, and may be an avenue for improvement.

Previously, our group showed that simple structure-based models used in the design of parallel dimeric coiled coils poorly capture the energetic contributions of core residues (Grigoryan et al., 2006). In a later study, Grigoryan et al. were able to overcome this by replacing the inter-chain energies of core residues with experimentally measured coupling energies (Grigoryan et al., 2009). Coupling energies ($\Delta\Delta\Delta G$) are measured by performing a double mutant thermodynamic cycle between a pair of amino acids in a structure (Serrano et al., 1990). Many coupling energies have been measured for core residues in parallel coiled-coil dimers (Acharya et al., 2006). Hadley and Gellman measured coupling energies for antiparallel coiled coils (Hadley & Gellman, 2006). These can in theory be used to replace the weights for interactions between core residues in the cluster expansion model, as done by Grigroyan et al. for

parallel coiled-coil dimers (Grigoryan et al., 2009). However, there are far fewer measurements for antiparallel coiled coils than for parallel coiled coils, 25 vs. 100, respectively (Hadley & Gellman, 2006; Acharya et al., 2006). This is problematic when predicting properties of native coiled coils, because native coiled coils have many core residue interactions for which there is no measured coupling energy. In the *de novo* design of coiled coils, however, it is possible to restrict the sequence space that is used. With such a restriction, the calculation can be set up such that evaluated sequences are highly enriched in interactions for which coupling energy measurements are available. With such an arrangement, the coupling energies can potentially make a significant impact improving design performance, despite their sparse description of the possible core residue interactions.

4.2.2 *Allowing greater backbone flexibility*

In deriving cluster expansion models in this study, ensembles of backbones were used to represent the antiparallel and parallel states. However, these backbones ensembles were chosen to have minimal structural diversity (backbone R.M.S.D. < 1 Å). This was due to the concern that a single cluster expansion model would not be able to accurately approximate a scoring procedure that used structurally diverse backbones. However, a recent study by Murphy et al. on the four-helix bundle CheA showed that backbone motions between 1-2 Å were needed to design a sequence with a fully mutated core within the Rosetta molecular modeling software (Murphy et al., 2012). Perhaps the backbones used in modeling antiparallel and parallel dimers in this study lacked sufficient structural diversity to accurately capture the properties of core residues in

dimers. If true, using a more structurally diverse set of backbones could improve prediction of orientation preference.

The cluster expansion models used to approximate energies on a more structurally diverse set of backbones would need to be tested, to determine how accurately they capture this new scoring protocol. It should be noted that the cluster expansion models in this study captured energies derived from backbone ensembles that differ by up to 1 Å RMSD very accurately, $R^2 \sim 0.90$. Thus, obtaining a cluster expansion that accurately captures backbone motions of 1-2 Å seems feasible. Additionally, one could compare the cluster-expansion weights of core residue interactions to their corresponding coupling energies, and systematically check how incorporating more structurally diverse backbones affects the correlation between these values. As mentioned earlier, coupling energies are experimentally derived weights for residue-residue interactions, and thus provide a way to evaluate predicted residue-residue interactions that arise when fitting the pairwise terms in the cluster expansion models.

4.2.3 *Using a standard set of terminal heptads*

A strategy used by Havranek and Harbury to design parallel coiled-coil homo and hetero dimers was to mutate only the central heptad of a variant of GCN4 (Havranek & Harbury, 2003). GCN4 is a native parallel homodimer, and using it as a scaffold provides two advantages. First, the flanking sequence around the central heptad contains sequence determinants that favor the dimeric state of a coiled coil in preference to higher-order assemblies. As mentioned earlier, formation of higher-order complexes rather than antiparallel dimers was the most common way

that my designs failed. Additionally, terminal ends are difficult to model since these are often very flexible regions in the absence of capping motifs (Harper & Rose, 1993). Moreover, the cluster expansion models were derived to approximate energies within the central heptad of a coiled-coil structure. As a result, the cluster expansion models are not well equipped to score residue interactions at the N and C terminal ends of a coiled coil. Having standard terminal heptads would prevent the need to model these terminal ends.

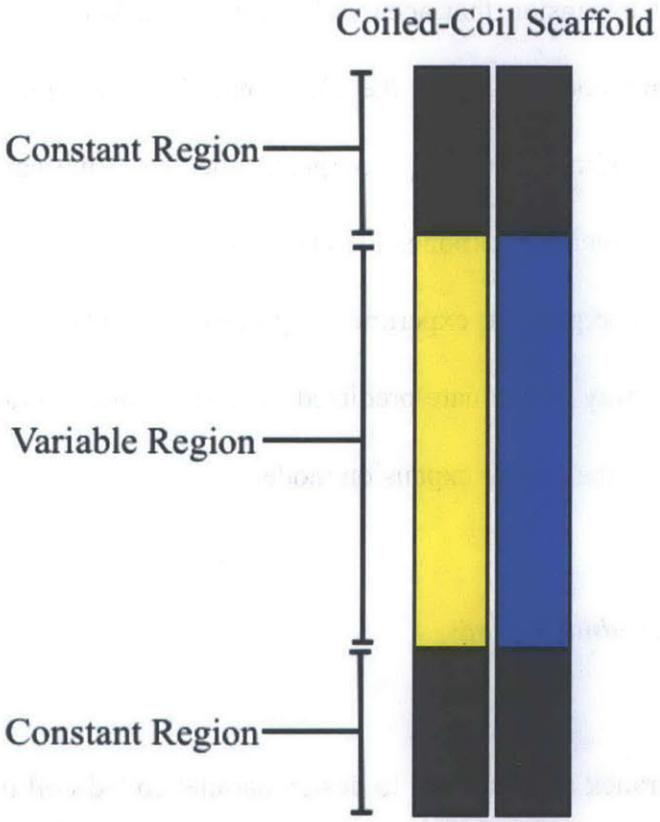


Figure 4-1. Using a standard set of terminal heptads. A cartoon of a dimeric coiled coil is shown as two rectangles. The rectangles are composed of two regions. The constant regions are located at the terminal ends of the coiled coil and are colored dark grey. The sequences in these constant regions are unchanged. These sequences can be taken from wild-type or synthetic coiled-coil sequences. The second region is the variable region, colored in yellow and blue. These represent the design positions, and can be mutated. The differential coloring indicates that these sequences can be independent of each other, indicative of a heterodimer, or they can be equivalent sequences, indicative of a homodimer.

GCN4 would not be an ideal scaffold for the terminal heptads in designed antiparallel homodimers because of its preference for the parallel state. However, the antiparallel homodimer domain of Bcr, from the Bcr-Abl oncoprotein, represents one of the most well studied antiparallel homodimers; it could be an ideal scaffold (Taylor & Keating, 2005). For instance, this domain has been used as a reagent in polyhedron design, establishing its use as a molecular reagent (Gradišar et al., 2013). Another scaffold to base the design of new antiparallel coiled coils is the rationally designed antiparallel homodimer APH. This could be a particularly useful scaffold due to its high thermostability (Gurnon et al., 2003). There is one challenge that arises from using either of these scaffolds, however: terminal heptads derived from homodimers will promote heterodimerization of different designs. However, this may be a minor problem as long as adequate negative design elements are included in the designed part of the sequence. For example, the antiparallel homodimers designed in this work using CLASSY were six heptads long. If only the two terminal heptads were replaced with scaffold sequence to promote the formation of antiparallel dimers, many specificity determinants could still be placed in the central four heptads to destabilize any cross talk between the designed PPIs. It is also important to note that studies on coiled coils suggest that the central heptads play a larger role in stabilizing a complex than terminal heptads (Zhou et al., 1992). As a result, it is likely that the sequence features in the central heptads can override the unwanted stabilization of antiparallel heterodimers from the use of known terminal ends.

4.2.4 *Including higher-order off-target states*

Lastly, one strategy to improve the design of antiparallel homodimers is to include models that capture higher-order states. In my design of antiparallel homodimers, two designs appeared to form higher-order states. Analytical ultracentrifugation data were consistent with one being a trimer and the other appearing to be in monomer-tetramer equilibrium. Including higher-order states in the design methodology would be challenging, but could be attempted. Trimers and tetramers can adopt a wide range of topologies because the helices in these oligomerization states can have a variety of orientations with respect to each other, as well as form different types of heteroassemblies. However, one approach for dealing with the multitude of higher-order off-target states may be to place energetic constraints to just a single trimer and tetramer state. This may have the effect of destabilizing several higher-order states that are not explicitly constrained. For instance, in chapter 3 two sets of three antiparallel homodimers were designed with constraints to various dimeric off-target states, but no explicit constraints were placed on interactions between the two sets. However, computational modeling predicted that the sequences in these two sets would not cross react. For the subset of these sequences that folded into antiparallel homodimers, two of these predictions (APH2 and APH4; APH3 and APH4) were experimentally validated to not cross react. This demonstrates that under certain conditions, negative design can be obtained without explicit modeling, as long as several closely related off-target states are considered.

4.3 Screening libraries of orthogonal PPIs

As computational resources continue to increase in speed, and our algorithms become ever faster to evaluate, as well as more accurate, it may be possible to design large sets of dozens to hundreds of orthogonal PPIs. This could be done by computationally designing small sets of orthogonal interactions and then computationally testing for cross-reactivity, as done in this thesis, or by simultaneously designing a dozen or more orthogonal PPIs at once. As larger orthogonal PPI sets are designed, the number of PPI to experimentally test will increase rapidly. For instance, designing ten homodimers would require testing 55 PPIs. The overwhelming majority of these PPIs, 45/55 in this example, would be predicted to be weak. Ways to quickly discriminate non-interacting proteins from those that do associate could dramatically reduce the time needed to experimentally evaluate designs.

Several high-throughput strategies exist for measuring whether two proteins interact (Fields & Song, 1989; Newman & Keating, 2003; Remy & Michnick, 2006). An assay developed by Magliery et al. to test for protein interactions is particularly noteworthy (Magliery et al., 2005). To detect binding in this assay, a split version of the green fluorescent protein (GFP) is made, which can no longer fold on its own. The split version is composed of an N-terminal half and a C-terminal half. The N-terminal half of GFP is genetically fused to one member of the PPI, and the C-terminal half of GFP is fused to the other PPI partner (Figure 4-2).

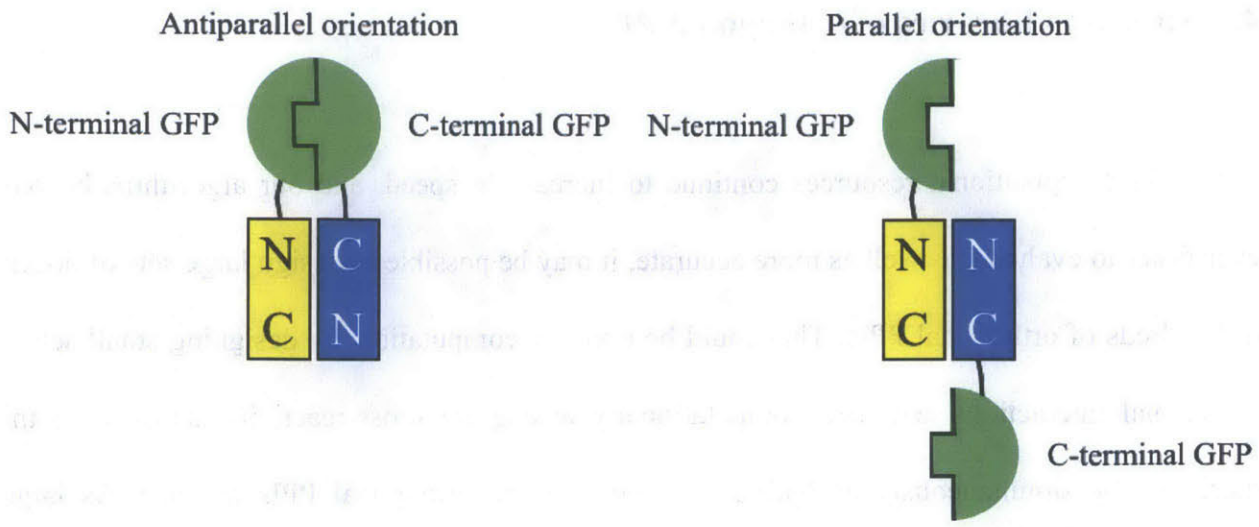


Figure 4-2. Split GFP Assay. Green semi-circles represent the two green fluorescent protein (GFP) fragments. Yellow and blue rectangles represent the subunits of a coiled coil, with N indicating the N-terminal end and C indicating the C-terminal end. In the antiparallel orientation, GFP assembly can occur because both fragments are presented on the same side of the coiled coil. However, in the parallel orientation, the C-terminal fragments are spatially secluded from the N-terminal fragment, preventing assembly.

The constructs are then transformed into bacteria. In the absence of binding the GFP will not fold, but upon binding and proper colocalizing of GFP fragments, the GFP will fold and fluoresce in the cell. This assay could be used to quickly assess whether ten sequences designed to be homodimers interact in a manner that allows reconstitution of GFP. Additionally, the same assay could be used to assess whether any of the 45 possible heterotypic interactions compete with the 10 homotypic interactions, thus testing orthogonality. The benefit of this strategy is that it does not require purification of any protein components. Additionally, this assay determines whether the PPI being evaluated functions inside cells, a property useful for synthetic biology. Magliery et al. applied this assay to evaluate binding for 256 antiparallel heterodimeric coiled coils, suggesting the technique is amendable to the antiparallel homodimer design problem mentioned in this thesis.

Several caveats with respect to the GFP reconstitution assay should be considered when applying this to the screening of orthogonal antiparallel homodimers. The folding of GFP is irreversible, and as a result even if the equilibrium of a binding reaction heavily favors homodimers, GFP assembly may trap a transient interaction between heterodimers. Additionally, even if a pair of sequences are detected to be orthogonal antiparallel homodimers, the orientation preference of these sequences are not known, since these sequences may be trapped in the antiparallel state due to the GFP, but actually prefer the parallel state. Thus orientation preference would need to be further tested using more time consuming methods. Lastly, in the context of testing antiparallel homodimeric coiled coils, if heterodimers form in the parallel orientation, they would evade detection. This would occur because the GFP components would be on opposite termini and could not assemble (Figure 4-2). One solution to this issue is to make an additional construct for every sequence in the design set that has the C-terminal GFP component at the N-terminus of the design sequence, allowing testing of both antiparallel and parallel interactions. This would also have the added benefit of reporting on topological features of the interactions. As ever larger sets of orthogonal pairs are designed, the experimental speed up provided by this split-GFP assay would likely outweigh the cost of setting up the assay and developing additional reporters.

4.4 Application of orthogonal coiled coils

It is interesting to speculate on the possible applications of a large set of orthogonal coiled coils. I will describe two possible applications. Gradišar et al. developed a technique to

design polyhedra of arbitrary shapes using orthogonal parallel and antiparallel coiled coils, as mentioned in previous chapters (Gradišar et al., 2013). This technique requires an orthogonal coiled coil for every edge in a polyhedron. Given a large set of orthogonal coiled coils, one could attempt to design icosahedrons. Icosahedrons have 30 edges, and are the typical geometry used by viral capsids (Zandi et al., 2004). Icosahedrons are attractive since they generate the maximal enclosed volume for shells composed of a single subunit (Caspar & Klug, 1962), and therefore may be ideal molecular cages for the transport of large molecules, or a large collection of molecules.

Beyond polyhedra design, a large set of orthogonal coiled coils may be useful for modifying the activity of nonribosomal peptide-synthetase (NRPS) pathways. NRPS pathways synthesize peptides independent of mRNA. The peptide products of NRPS pathways have many important medical applications. For example, they are used as antibiotics, antitumor agents, and immunosuppressants (Strieker et al., 2010). NRPS pathways are made of multiple enzymatic modules, with each enzymatic module performing a unique catalytic function. Given a large set of orthogonal heterodimeric coiled coils, it would be possible to genetically fuse a member of each heterodimer to a terminal end of an NRPS module to form assembly lines (Figure 4-3A). Orthogonal homodimeric coiled coils can additionally be used to create clusters of assembly lines, which may increase the efficiency of these pathways due to their higher density (Figure 4-3B) (Tsai et al., 2013). Lastly, some groups have begun reengineering the enzymatic substrate specificities of the enzymes in the NRPS modules (Chen et al., 2009). Combing these achievements with orthogonal coiled coils will allow for the design of completely synthetic NRPS pathways, which may revolutionize how molecules are synthesized.

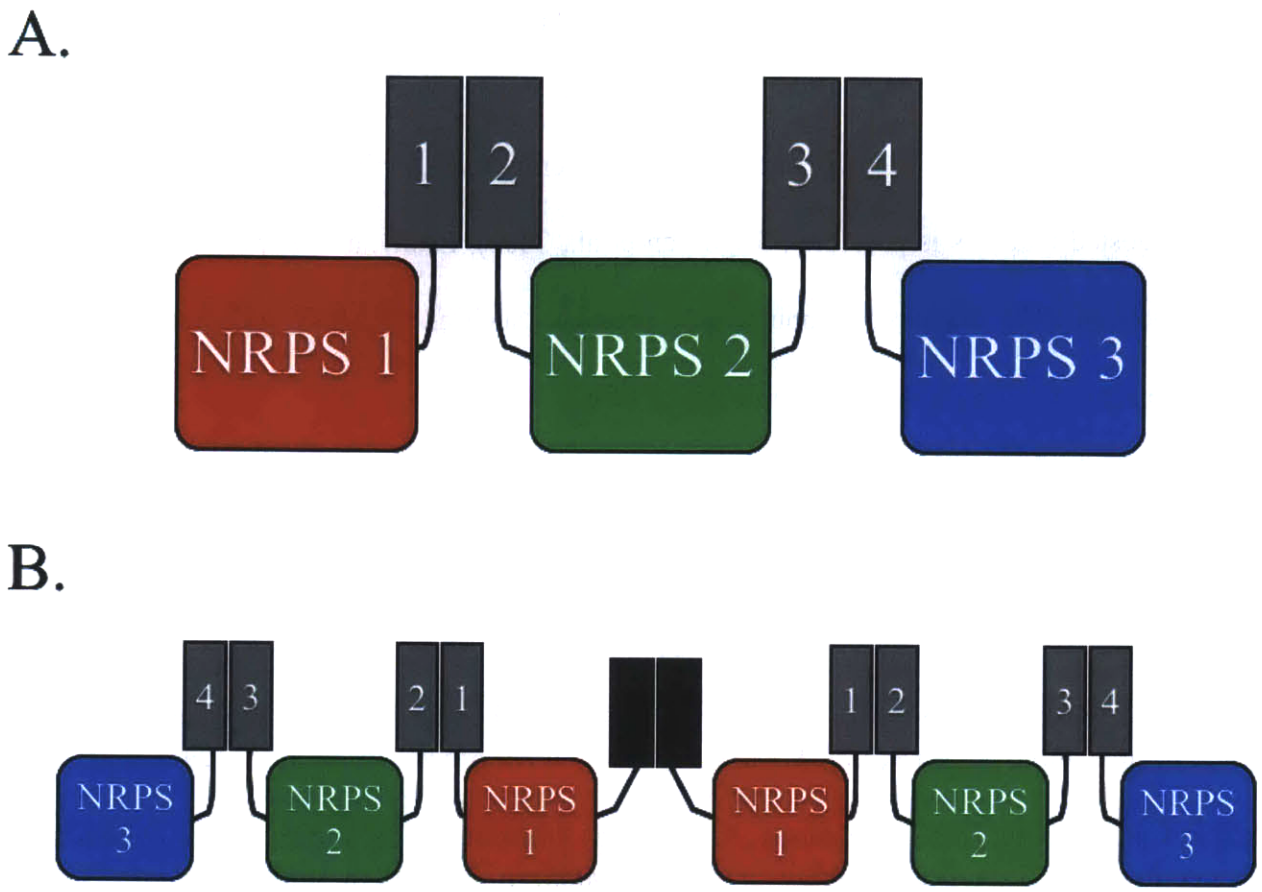


Figure 4-3. A schematic of coiled coils manipulating NRPS pathways. Coiled coils are shown as elongated rectangles, and NRPS modules are shown as rounded rectangles. (A) Shows orthogonal coiled coils as tools for directing assembly line formation among a set of NRPS modules. (B) A homodimer, that is orthogonal to the other coiled coil heterodimers, is used to promote the formation of clusters.

4.5 Summary

Orthogonal PPIs are of high value to many fields that involve molecular engineering. CLASSY provides a computational strategy for designing multiple orthogonal PPIs at once, and has now been experimentally validated as an approach for orthogonal design. It should also be noted that CLASSY was previously combined with an energy function that was developed specifically for parallel dimers (Grigoryan et al., 2009). In this work, CLASSY was combined

with an energy function broadly developed for modeling protein properties. As a result, it may be possible to extend the strategy used in this work to many other PPI systems. The ability of CLASSY to model non coiled-coil interactions has only been minimally explored, but results have suggested that this is possible (Grigoryan et al., 2006). Lastly, as computational resources and models of PPIs improve, computational strategies such as CLASSY will become a more attractive approach for the design of orthogonal PPIs.

4.6 References

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.

Apgar, J. R., Gutwin, K. N., & Keating, A. E. (2008) Predicting helix orientation for coiled-coil dimers. *Proteins: Structure, Function, and Bioinformatics*, *72* (3), 1048-1065.

Apgar, J. R., Hahn, S., Grigoryan, G., & Keating, A. E. (2009) Cluster expansion models for flexible-backbone protein energetics. *Journal of Computational Chemistry*, *30* (15), 2402-2413.

Caspar, D. L. D., & Klug, A. (1962) Physical principles in the construction of regular viruses. *Cold Spring Harbor Symposia on Quantitative Biology*, *27*, 1-24.

Chen, C. Y., Georgiev, I., Anderson, A. C., & Donald, B. R. (2009) Computational structure-based redesign of enzyme activity. *Proceedings of the National Academy of Science*, *106*, 3764–3769.

Fields, S., & Song, O. K. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, *340*, 245-246.

Gordon, D. B., Marshall, S. A., & Mayo, S. L. (1999). Energy functions for protein design. *Current Opinions in Structural Biology*, *9*, 509–513.

Gradišar, H., Božič, S., Doles, T., Vengust, D., Hafner-Bratkovič, I., Mertelj, A., Webb, B., Sali, A., Klavžar, S., & Jerala, R., (2013) Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nature Chemical Biology*, *9* (6), 362-366.

Grigoryan, G., Reinke, A. W., & Keating, A. E. (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature*, *458* (7240), 859-864.

Grigoryan, G., Zhou, F., Lustig, S. R., Ceder, G., Morgan, D., & Keating, A. E. (2006) Ultra-fast evaluation of protein energies directly from sequence. *PLoS Computational Biology*, *2* (6), 551-563.

Gurnon, D. G., Whitaker, J. A., & Oakley, M. G. (2003) Design and characterization of a homodimeric antiparallel coiled coil. *Journal of the American Chemical Society*, *125*, 7518-7519.

Hadley, E. B., & Gellman, S. H. (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.

Kapp, G. T., Liu, S., Stein, A., Wong, D. T., Reményic, A., Yeh, B. J., Fraser, J. S., Taunton, J., Lim, W. A., & Kortemme, T. (2012) Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. *Proceedings of the National Academy of Science*, *109* (14), 5277–5282.

Kingsford, C. L., Chazelle, B., & Singh, M. (2005) Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, *21* (7), 1028-1036.

Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L., & Baker, D. (2004) Computational redesign of protein-protein interaction specificity. *Nature Structural Biology and Molecular Biology*, *11* (4), 371-379.

Harper, E. T., & Rose, G. D. (1993) Helix stop signals in proteins and peptides: The capping box. *Biochemistry*, *32* (30), 7605-7609.

Havranek, J. J, & Harbury, P. B. (2003) Automated design of specificity in molecular recognition. *Nature Structural Biology* , *10* (1), 45-52.

Magliery, T. J., Wilson, C. G. M., Pan, W., Mishler, D., Ghosh, I., Hamilton, A. D., & Regan, L. (2005) Detecting protein–protein interactions with a green fluorescent protein fragment reassembly trap: Scope and mechanism. *Journal of the American Chemical Society*, *127* (1), 146-157.

Murphy, G. S., Mills, J. L., Miley, M. J., Machius, M., Szyperski, T., & Kuhlman, B. (2012) Increasing sequence diversity with flexible backbone protein design: The complete redesign of a protein hydrophobic core. *Structure*, *20* (6), 1086-1096.

Newman, J. R. S., & Keating, A. E. (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science*, *300* (5628), 2097-2101.

Ollikainen, N., Smith, C. A., Fraser, J. S., & Kortemme, T. (2013) Flexible backbone sampling methods to model and design protein alternative conformations. *Methods in Enzymology*, *523*, 61-85.

Remy, I., & Michnick, S. W. (2007) Application of protein-fragment complementation assays in cell biology. *Biotechniques*, *42* (2), 137-145.

Sammond, D. W., Eletr, Z. M., Purbeck, C., & Kuhlman, B. (2010) Computational design of second-site suppressor mutations at protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, *78*, 1055-1065.

Serrano, L., Horovitz, A., Avron, B., Bycroft, M., & Fersht, A. R. (1990) Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry*, *29* (40), 9343-9352.

Strieker, M., Tanovic, A., & Marahiel, M. A. (2010) Nonribosomal peptide synthetases: structures and dynamics. *Current Opinion in Structural Biology*, *20* (2), 234-240.

Taylor, C. M. & Keating, A. E. (2005) Orientation and oligomerization specificity of the Bcr coiled-coil oligomerization domain. *Biochemistry*, *44*, 16246-16256.

Tsai, S. L., DaSilva, N. A., & Chen, W. (2013) Functional display of complex cellulosomes on the yeast surface via adaptive assembly. *ACS Synthetic Biology*, *2*, 14–21.

Zandi, R., Reguera, D., Bruinsma, R. F., Gelbart, W. M., & Rudnick, J. (2004) Origin of icosahedral symmetry in viruses. *Proceedings of the National Academy of Science*, *101* (44), 15556–15560.

Zhou, N. E., Kay, C. M., & Hodges, R. S. (1992) Synthetic model proteins. *Journal of Biological Chemistry*, *267* (4), 2664-2670.

Appendix A

Experimental data on antiparallel homodimers from the literature

Below are the "Specification Sheets" for antiparallel homodimers. The "Specification sheets" summarize the data for antiparallel homodimers that have been biophysical characterized in the literature. The known, or hypothesized, sequence alignments are shown. Additionally, the helical-wheel diagrams for those alignments, generated using DrawCoil 1.0 (http:// www.grigoryanlab.org/drawcoil/), are provided. At the bottom of each sheet is a table that lists the values from experimental measurements performed on the antiparallel homodimer in question. Names and references involving the antiparallel homodimer are listed on the top. SEC refers to size-exclusion chromatography. AUC refers to analytical ultracentrifugation, and AP stands for antiparallel.

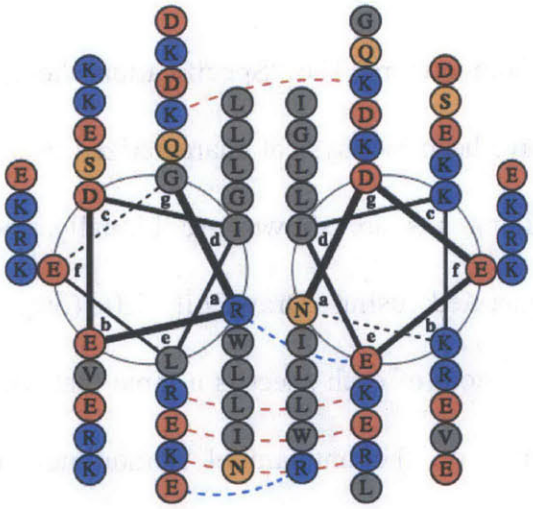
Name: Oligomerization domain of hepatitis delta antigen

Paper Reference:
Zuccola et al. Structural basis of the oligomerization of hepatitis delta antigen. *Structure* (1998) vol. 6 (7) pp. 821-830.

Alignment:

```
heptad position
.........gabcdefgabcdefgabcdefgabcdefgabcdefga
.........GREDILEQWVSGRKKLEELERDLRKLKKKIKKLEEDNPWLGNIKGIIGKY
............agfedcbagfedcbagfedcbagfedcbagfedcbag
YKGIIGKINGLWPNDEELKKIKKKLKRLDRELEELKKRGSVWQELIDERG
- from crystal structure, PDB ID: 1A92
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
-	-	-	X-ray	-

Additional Comments:
Structure has a unique property that towards the bottom of the coiled coil the helix kinks out allowing the dimer to form octamers. Removing the helices that kink out results in structures that are significantly less helical according to the authors.

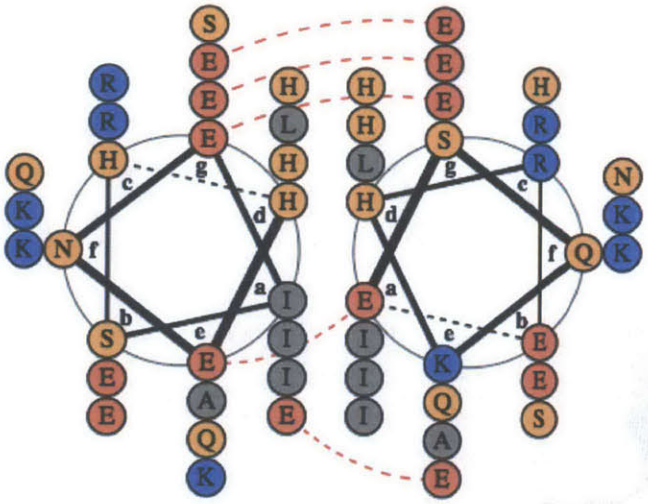
Name: Bovine IF1

Paper Reference: Gordon-Smith et al. Solution structure of a C-terminal coiled-coil domain from Bovine IF1: The inhibitor protein of F1 ATPase. *Journal of Molecular biology* (2001) vol. 308 pp. 325-339.

Alignment:

```
heptad position
.....defgabcdefgabcdefgabcdefga
ALKKHHENEISHHAKEIERLQKEIERHKQSEDDD
.....agfedcbagfedcbagfedcbagfed
..DDDESQKHREIEKQLREIEKAHHSIENEHHKKLA
- from NMR structure, PDB ID: 1HF9
- Color = CC+ defined coiled coil region
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
-	-	-	NMR	-

Additional Comments:

At a pH value below 6.5 it forms an active dimer. At higher pH values, two dimers associate to form an inactive tetramer.

Name: Coiled-coil domain of Osmosensory Transporter ProP in *E. Coli*

Paper References:

Zoetewey et al. Solution structure of the C-terminal antiparallel coiled-coil domain from Escherichia coli osmosensor ProP. *Journal of Molecular biology*. (2003) vol. 334 pp. 1063-1076.

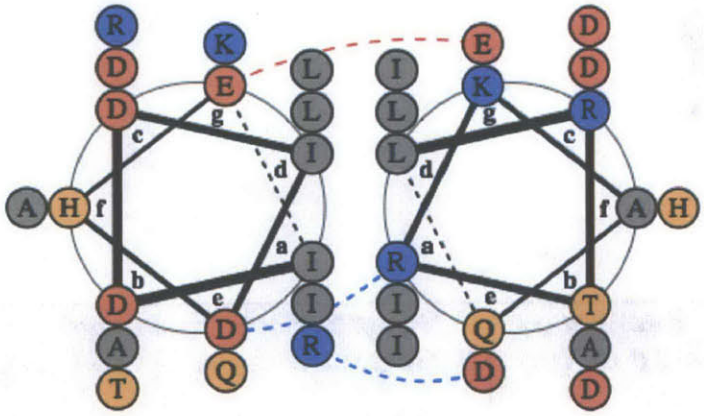
Hillar, et al. Detection of R-helical coiled-coil dimer formation by spin-labeled synthetic peptides: A model parallel coiled-coil peptide and the antiparallel coiled-coil formed by a replica of the ProP C-terminus, *Biochemistry.* (2003) 42, 15170-15178.

Hillar et al. Formation of an antiparallel, intermolecular coiled-coil is associated with in ViVo dimerization of osmosensor and osmoprotectant transporter ProP in Escherichia coli, *Biochemistry.* (2005) 44, 10170-10180.

Alignment:

```
heptad position
.........abcdefgabcdefgabcd
CGGDNIEQKIDDIDHEIADLQAKRTRLVQQHPR
.........dcbagfedcbagfedcba
...RPHQQVLRTRKAQLDAIEHDIDDIKQEINDGGC
- from crystal structure, PDB ID: 1R48
- Color = CC+ defined coiled coil region
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
-	-	28 °C @ 95μM	NMR	Measured transporters uptake rate
				Cross-linking study to check dimerization occurred in vivo in an AP orientation.

Additional Comments:
The construct is four heptads long. Authors noted the existence of a five-heptad construct. A Pfam model, Osmo_CC, exists, and contains about 544 sequences from 540 species. Also important to note, the structure has a unique bend in the coiled coil.

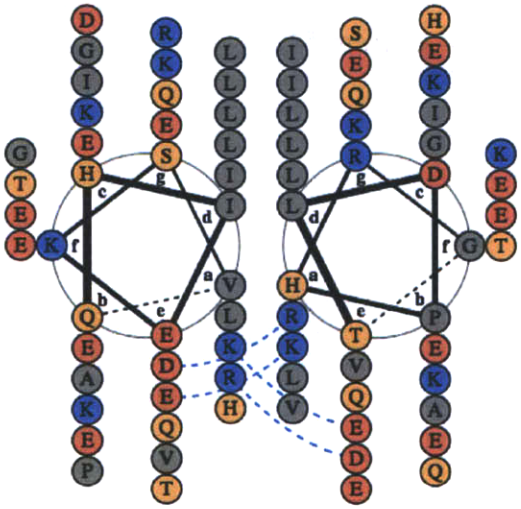
Name: Coiled-coil domain of Osmosensory Transporter ProP in *A. tumefaciens*

Paper Reference:
Tsatskis et al. Core Residue Replacements Cause Coiled-Coil Orientation Switching in Vitro and in ViVo: Structure-Function Correlations for Osmosensory Transporter ProP. *Biochemistry* (2008) vol. 47 pp. 60-72.

Alignment:

```
heptad position
..bcdefgabcdefgabcdefgabcdefgabcdefgabcde
..QHIEKSVEEIDEELAKLEEQKKILQTKREGLVGRHPDLT
edcbagfedcbagfedcbagfedcbagfedcbagfedcb
TLDPHRGVLGERKTQLIKKQEELKALEEDIEEVSKEIHQ
- alignment based on a crystal structure, PDB ID: 1R48 *Crystal structure is
of a homolog found in E. Coli.
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
AP pref	-	45.5 °C @ 50 μM	-	Measured transporters uptake rate
				Cross-linking study to check dimerization occurred in vivo in an AP orientation.

Additional Comments:
Authors also tested a mutation of this sequence, K498I, that showed increased thermal stability, (63.5°C). It also showed antiparallel preference in a disulphide competition, and similar kinetics in measured transporters uptake rate. Important to note that this paper may be the only paper to show a change in biological function when orientation is switched.

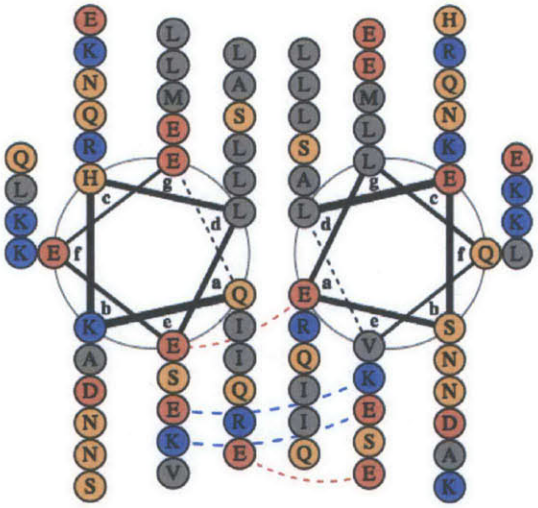
Name: Mitofusin domain HR2 V686M/I708M mutant (Mfn1 HR2)

Paper Reference: Koshiba et al. Structural Basis of Mitochondrial Tethering by Mitofusin Complexes. *Science* (2004) vol. 305 (5685) pp. 858-862.

Alignment:

```
heptad position
.....................................abcdefgabcdefgabcdefgabcdefgabcdefgabcd
LVPRGSHMFTSANCSHQVQQEMATTFARLCQQVDMTQKHLEEEIARLSKEIDQLEKMQNNSKLLRNKAVQLESELEN
FSKQFLH
                                          dcbagfedcbagfedcbagfedcbagfedcbagfedcba
.........................HLFQKSFNELESELQVAKNRLLKSNNQMKELQDIEKSLRAIEEELHKQTM
DVQQCLRAFTTAMEQQVQHSCNASTFMHSGRPVL
- from crystal structure, PDB ID: 1T3J
- Color = CC+ defined coiled coil region
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
-	-	78 °C @ 50 µM	X-ray	Tested for mitochondrial fusion activity

Additional Comments:

Authors also checked for proteolysis resistance. The proteolysis product corresponded well with the predicted heptad repeat. In addition, the construct showed the ability to self-associate in an immunoprecipitation assay.

Name: Bcr Coiled-Coil Oligomerization Domain

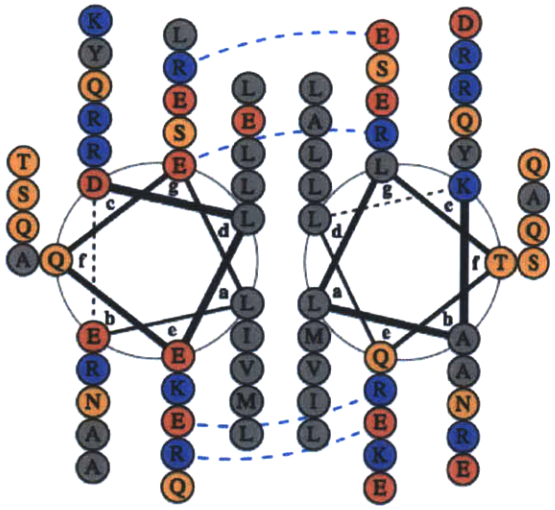
Paper Reference: Taylor and Keating. Orientation and oligomerization specificity of the Bcr coiled-coil oligomerization domain. *Biochemistry* (2005) vol. 44 pp. 16246-16256.

Alignment:

```
heptad position
....cdefgabcdefgabcdefgabcdefgabcdefgabc
....DLEQELERLKASIRRLEQEVNQERSRMAYLQTLLAKGGC
...cbagfedcbagfedcbagfedcbagfedcbagfedc
CGGKALLTQLYAMRSRLQNVEQELRRISAKARELEQELD
```

- based on a crystal structure, PDB ID: 1K1F *Crystal structure is of a tetramer formed from the homodimer (really a dimer of two Helix-loop-Helix constructs). It also includes an N-terminal extension sequence. The authors also made three mutations to the sequence in the crystal structure.

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
AP preference	monomer*	53 °C* @ 25 μM	-	-
	*This is for the disulphide linked monomer	*This is for the disulphide linked monomer		

Additional Comments:

The construct tested is technically not a homodimer due to two point mutations.

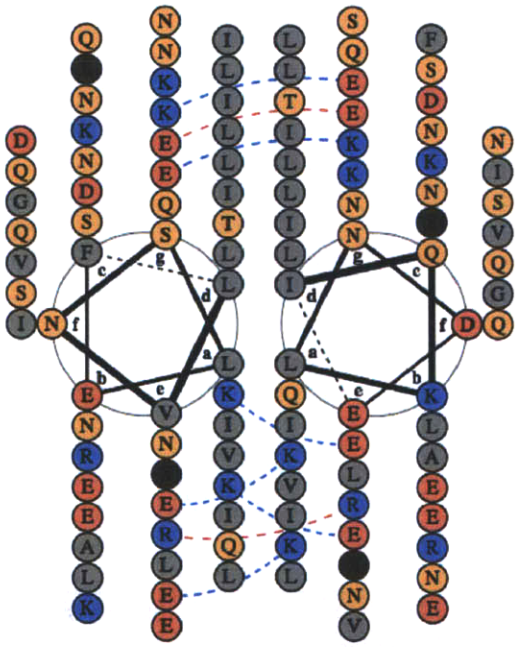
Name: Coiled-coil domain of Mdv1

Paper Reference: Koirala et al. Molecular architecture of a dynamin adaptor: implications for assembly of mitochondrial fission complexes. *The Journal of Cell Biology* (2010) vol. 191 (6) pp. 1127-1139.

Alignment:

```
heptad position
..............defgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcd
.........GPQRTLVNSLEFLNIQKNSTXSEIRDIEVEVENLRQKKEKLLGKIANIEQNQLXLEDNLKQIDDRLDF
LEEYG
...........dcbagfedcbagfedcbagfedcbagfedcbagfedcbagfedcbagfedcbagfed
GYEELFDLRDDIQKLNDELXLQNQEINAIKGLLKEKKQRLNEVEVEIDRIESXTSNKQINLFELSNVLTRQPG
- from a crystal structure, PDB ID: 2XU6
- Color = CC+ defined coiled coil region
- X = Selenomethionine
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
-	dimer	-	X-ray	A set of KO with replacement followed by various functional assays.

Additional Comments:

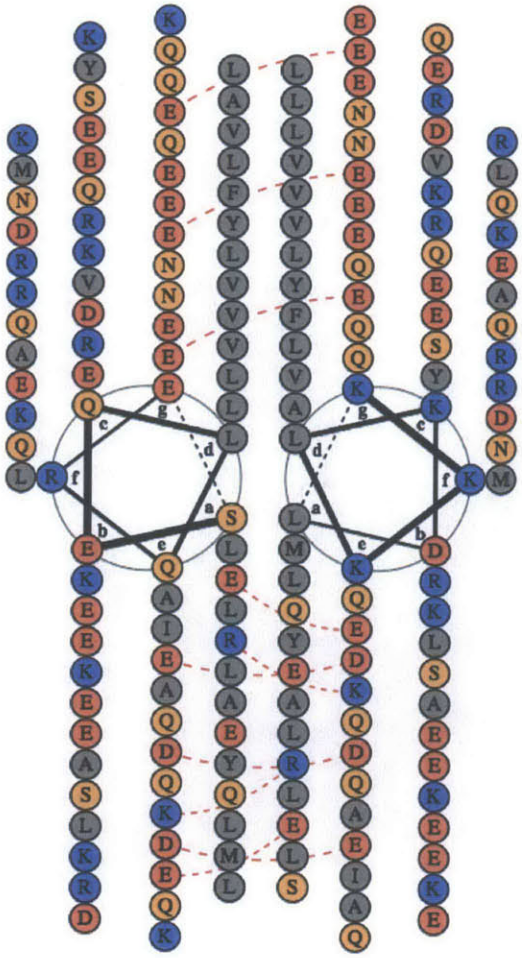
Name: Coiled-coil domain of Beclin1

Paper Reference: Li et al. Imperfect interface of Beclin1 coiled-coil domain regulates homodimer and heterodimer formation with Atg14L and UVRAG. Nature Communications (2012) vol. 3 pp. 662-11

Alignment:

```
heptad position
.....abcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgabcdefgab
GPGSDSEQLQRELKELALEEERLIQELEDVEKNRKVVAENLEKVQAEAERLDQEEAQYQREYSEFKRQQLELDDELK
cdefgabcdefgabcde
SVENQMRYAQMQLDKLKKK
....edcbagfedcbagfedcbagfedcbagfedcbagfedcbagfedcbagfedcbagfedcbagfedc
..KKKLKDLQMQAYRMQNEVSKLEDDLELQQRKFESYERQYQAEEQDLREAEAQVKELNEAVVKRNKEVDELEQILR
bagfedcbagfedcba
EEELALEKLERQLQESDSGPG
- from a crystal structure, PDB ID: 3Q8T
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
-	dimer	40 °C @ 89 μM	X-ray	Characterize the self association of Beclin 1 in vivo

Additional Comments:
Authors made mutational variants of Beclin1, referred to as MutStab, with improved stability. Predominately, the mutants have substitution of the E in the core with L. Mutants showed dimeric behavior by AUC, and increased thermal stability, with the most dramatic Tm increase to 60°C.

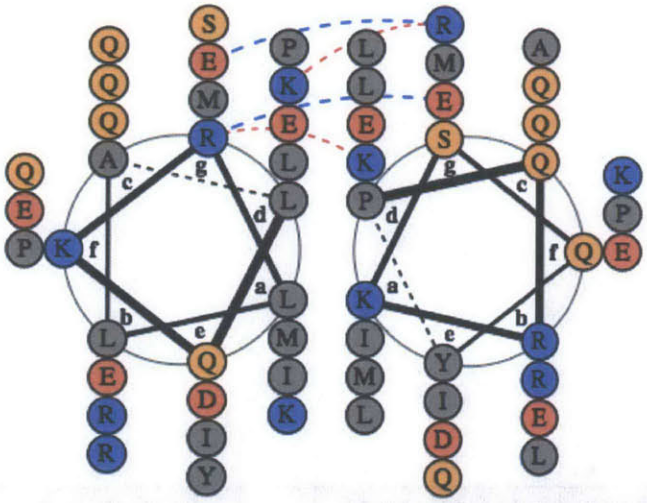
Name: SARAH domain.

Paper Reference: Aruxandei et al. Dimerization-Induced Folding of MST1 SARAH and the Influence of the Intrinsically Unstructured Inhibitory Domain: Low Thermodynamic Stability of Monomer. *Biochemistry* (2011) vol. 50 pp. 10990−11000.

Alignment:

```
heptad position
.............defgabcdefgabcdefgabcdefgabcddefga
GSDYEFLKSWTVEDLQKRLLALDPMMEQEIEEIRQKYQSKRQPILDAIEAK
.............agfeddcbagfedcbagfedcbagfedcbagfed
...........KAEIADLIPQRKSQYKQRIEEIEQEMMPDLALLRKQLDEVTWSKLFEYDSG
- from a crystal structure, PDB ID: 2JO8
- Color = CC+ defined coiled coil region
```

Helical Wheel:



* Only the longest canonical region of this coiled coil is shown.

Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
-	dimer	1.07 μM	NMR	-

Additional Comments:

This is a non-canonical coiled coil. A *d* position follows another *d* position. In addition, a unique 3_{10}-helix at the N-terminus makes an important set of interactions.

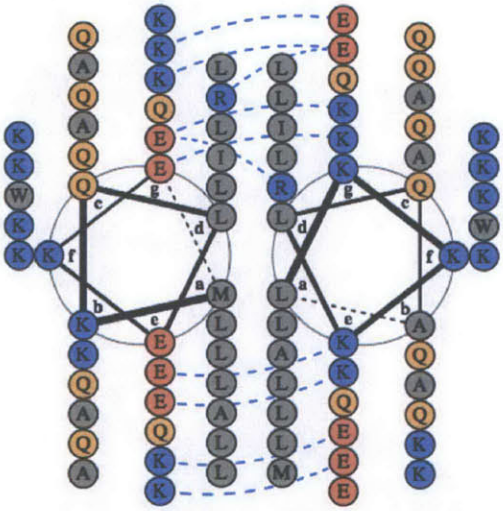
Name: Oakley's AP homodimer

Paper Reference: Gurnon et al. Design and characterization of a homodimeric antiparallel coiled coil. *Journal of the american chemical society* (2003) vol. 125 pp. 7518-7519

Alignment:

```
heptad position
abcdefgabcdefgabcdefgabcdefgabcdefgabcdefga
MKQLEKELKQLEKELQAIEKQLAQLQWKAQARKKKLAQLKKKL
...agfedcbagfedcbagfedcbagfedcbagfedcbagfedcba
...LKKKLQALKKKRAQAKWQLQALQKEIAQLEKELQKLEKELQKM
- based on intended design
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
AP pref	dimer	~2 nM	-	-

Additional Comments:

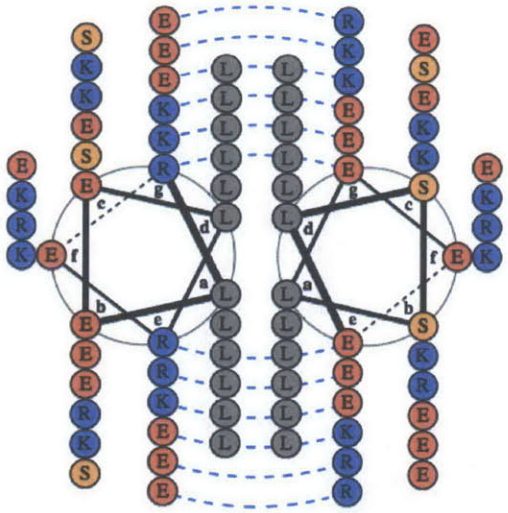
Name: Koksch's AP homodimer

Paper Reference: Pagel et al. Advanced approaches for the characterization of a de novo designed antiparallel coiled coil peptide. Organic & Biomolecular Chemistry (2005) vol. 3 (7) pp. 1189.

Alignment:

```
heptad position
gabcdefgabcdefgabcdefgabcdefgabcdefgabcde
RLEELREKLESLRKKLEELKRELRKLEKELKKLEEELSSLE
edcbagfedcbagfedcbagfedcbagfedcbagfedcbag
ELSSLEEELKKLEKELKRLERKLEELKKRLSELKERLEELR
- based on intended design
```

Helical Wheel:



Experimental Characterization:

Disulphide Competition	Oligomerization by (SEC or AUC)	Stability by (Kd or Tm)	Source of Structure	*in vivo*
-	dimer*	> 100°C @ 50 μM	-	-
	*Oligomerization was tested using ESI-FTICR-MS		* Authors measure a 1H,15N HSQC for a subset of side chains confirming an AP dimeric state in solution.	

Additional Comments:

Authors additionally conducted a FRET study that was consistent with the presence of an antiparallel coiled coil in solution. In addition, the NMR experiments mentioned above, had an L to F mutation at the second L in the *a* position.

References:

Testa, O. D., Moutevelis, E., & Woolfson, D. N. (2009) CC+: a relational database of coiled-coil structures. *Nucleic Acids Research, 37*, 315-322.