Computational Design and Experimental Characterization of Protein Oligomers

Thesis by

Po-Ssu Huang

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2004

(Defended May 27, 2004)

Abstract

Previous efforts in designing protein binding interfaces have focused on altering binding specificities. These methods fall short, however, when applied to the design of novel binding sites due to difficulties in accurately modeling protein backbones. The goal of this project is to create dimers from monomeric proteins. We developed a special docking algorithm that positions the member protein subunits to a plausible configuration with respect to each other using parameters determined from known complex structures. The docking procedure treats the proteins as rigid bodies and uses Fourier correlation theorem and fast Fourier transform to efficiently search for dimers with the highest interfacial surface complementarities. Using the docked structures as scaffolds for design and employing hydrophobic surface residues to drive dimer formation, we have demonstrated two successful designs, one heterodimer and one homodimer, using protein G and engrailed homeodomain respectively as the starting monomeric proteins. The designed dimers were characterized using circular dichroism, nuclear magnetic resonance, analytical ultracentrifugation, and X-ray crystallography methods. This is the first report of computationally designed *de novo* protein homodimers generated using a combination of protein docking and protein design tools. These results suggest that this strategy can be used to address the protein recognition problem, and is generally applicable to creating novel binding sites with compatible binding partners.

Acknowledgements

Getting to this point of writing the acknowledgements section is really a great relief from the intensive month-and-a-half of thesis writing. I can finally take a break and think about the things and people that I have encountered over my entire graduate school years. Looking back at the six long years I have spent here at Caltech, I am very appreciative of everything I have received. It has been a bumpy ride with mostly discouraging research results, but fortunately, there have been many friends offering support throughout the years. I cannot thank enough those who have always believed in me, even during the worst period. Part of my memory has faded and I cannot mention all the names I should, but I am certainly grateful to them.

I would like to thank my thesis adviser, Stephen Mayo, for his guidance. Steve never gave up on me even when my progress was obviously behind my peers. He gave me a lot of freedom and resources to develop my own ideas and always kept me in check when I got distracted. I have always tried to talk Steve into letting me do some of the most unrealistic projects, and I thank him for bringing me back to reality. Also, I thank him for trusting me with the project on protein-protein interactions. Being probably Steve's first graduate student to go beyond five years, I thank Steve for his meticulous scientific guidance, patience, advice, and support, both on scientific and personal levels throughout my entire graduate career.

I would like to thank Pamela Bjorkman, Doug Rees, and Rich Roberts for serving as my graduate advisory committee. They have always been generous in offering their scientific expertise and advice, and I have learned a lot from each of them.

Several people have helped directly with my thesis project. Thanks to John Love for getting me started with computer graphics and programming and being a great friend. I owe him for laying much of the foundation for the projects. Andy Herr and I have become good friends through getting me started on the analytical ultracentrifugation experiments. Andy offered a lot of good advice and has always been very patient in explaining things to me. I would also like to thank Rich Olson, a postdoc in Pamela Bjorkman's lab, for helping me with some of the most difficult ultracentrifugation problems. Peter Schuck at NIH spent a great deal of time exchanging long emails with me in helping me understand some of the sedimentation velocity concepts, and I am grateful to him. I would like to thank Len Thomas, who recently started training me on X-ray crystallography and has been helping me process diffraction data. I appreciate all the useful tips on X-ray crystallography given to my by Takumi Koshiba. Thanks also to Pingwei Li for collecting my diffraction data at the SSRL synchrotron.

Thanks to Cynthia Carlson for taking care of all the administrative mess I created over the years, and for making the lab run smoothly. She is always enthusiastic and has had a positive influence over my gloomy days, and I am very lucky to have her around.

Rhonda Digiusto is my favorite lab member and a great friend. She has helped in every aspect of my graduate career, from dealing with my lab wastes to career advice. I would like to thank her for her friendship and her support.

This thesis would probably never be completed if it weren't for the help of Marie Ary. She taught me a lot about technical writing and has helped improve my English. Cynthia, Rhonda and Marie formed a consortium to help me get through graduate school, and I thank them very much for that.

I have benefited from a number of very smart people in the Mayo lab. When I started, Ben Gordon and Arthur Street introduced me to protein design. It would have been a lot harder if it were not for Ben and Arthur. Scott Ross taught me everything I needed to know about NMR. Niles Pierce helped me with script writing, programming, and career advice. J. J. Plecs showed me the basics of X-ray crystallography. Eric Zollars and Ben Allen helped me edit parts of the thesis and offered interesting research ideas. Thanks to Christina Vizcarra and Deepshikha Datta for sharing their benches with me and for being such nice benchmates. I would like to thank Chris Voigt, Premal Shah, John Love, and Geoff Hom especially for reminding me of the ill effects of gambling. Not mentioning the specifics, I would like to thank the rest of the Mayo (and Pierce) lab for creating a wonderful working environment. Thanks to Oscar Alvizo, Dan Bolon, Monica Breckow, Eun Jung Choi, Karin Crowhurst, Achilleas Frangakis, Sarah Hamilton, Jeniffer Keeffe, Kyle Lassila, Julie Mayo, Shannon Marshall, Jessica Mao, Chantal Morgan, Peter Oelschlaeger, Heidi Privett, Cathy Sarisky, Julia Shifman, Pavel Strop, Tom Treynor, Justin Bois, and Robert Dirks.

Most of my lunch hours were spent with people from Rich Roberts' lab. I would like to thank Terry Takahashi for his friendship, BBQ, and encouragements. Terry has been very kind to me since I first met him at the Biology retreat during our first year in graduate school. I have benefited from his wealth of knowledge over the years. I would also like to thank other members from the Roberts lab: Ryan Austin, Chris Balmaseda, Jeff Barrick, Adam Frankel, Harry Green, Bill Ja, Shuwei Li, Shelly Starck, Todd Thorsen, Christine Ueda, and Tianbing Xia, also Chris Otey from Francis Arnold's lab.

Thanks to William Agassounoun, Dan Lieberman, Anne-laure Lannie, Paul Wiggins and George Ouyang for being fun roommates. A special thanks goes to Torsten Bistritschan for helping me survive the stressful first year by introducing me to several of his friends and parties.

During my last year at Caltech, I ran into my high school classmate from Taiwan, Yulung Tang who is studying in the electrical engineering division of Caltech. He has helped me with a lot of mathematical concepts in signal processing.

Before coming to Caltech, several people carefully guided my career development and did their best to ensure my future success, and for that, I would like to thank Randy Schekman, my undergraduate research advisor at U.C. Berkeley; Gerry McDermott and Thomas Earnest, my mentors at Lawrence Berkeley National Laboratory; Tatsuo Kurihara, who taught me everything I know about molecular biology; and Al Hermann, my chemistry professor at Fullerton College, who introduced me to biochemistry research.

I would like to thank my parents, Ting-hsiung and Li-mei Huang and for their sacrifice over the years to bring me to the United States and for their endless support. I would also like to thank my bother, Zen-yu Huang, and sister, Ching-Ying Huang. Lastly, I would like to thank my beloved wife Shao-Yun Wei for her love and for giving up her comfortable life and career in Taiwan just to be with me and marry me.

Table of Contents

List of Tables

List of Figures

Abbreviations

DEE	dead end elimination
MD	molecular dynamics
MC	Monte Carlo
GRA	geometric recognition algorithm
FCT	Fourier correlation theorem
HSQC	heteronuclear single quantum coherence
PDB	Protein Data Bank
DFT	discrete Fourier transform
FFT	fast Fourier transform
CD	circular dichroism
T_m	melting temperature
UV	ultraviolet
Kd	dissociation constant
Ka	association constant
AUC	analytical ultracentrifugation
S	Svedberg (the unit of sedimentation coefficient)
NMR	nuclear magnetic resonance
E. coli	*Escherichia coli*

Chapter 1

Introduction

Computational Protein Design

Rational protein design refers in general to methods that follow the "inverse folding" approach[1]. Computational methods have proven to be particularly valuable and have stimulated renewed interest in the field. In this approach, sequences are threaded onto an experimentally determined backbone, and the energies for each of the sequences are computed according to a mechanical force field description of the system[2]. The design process starts with choosing a protein backbone scaffold, followed by renovating the amino acid positions with discrete side-chain rotamers occupying statistically significant conformational space. Optimization algorithms such as dead-end elimination (DEE) or Monte Carlo (MC) methods are then used to find a sequence or sequences that will adopt the same fold when produced experimentally.

Protein design methods following this general scheme have been applied to several systems to improve stability; however, these methods are now also being used to create proteins with catalytic activity[3,4] and altered binding specificities[5-9]. Applications have expanded to several areas, including: (1) assisting directed evolution studies by suggesting focused libraries of reduced complexity, (2) improving native constructs by enhancing stability, enzyme activity, or binding affinity/specificities, (3) exploring unknown protein characteristics through extensive comparison between the wild-type and designed models, (4) creating molecules with novel properties, and (5) providing a "controlled" environment for testing the validity of bio-physical principles, theories, or energy functions. Previous successes in designing stable proteins have validated several aspects of the computational protein design approach, such as the use of fixed backbones, discrete rotamers, classical mechanics force field terms, and rigorously defined polar surface and hydrophobic core compositions, to name a few. The development of protein design algorithms is ongoing and far from perfect, but with the already-sophisticated tools currently available, we can start investigating some of the most intriguing problems in protein biochemistry.

Can we create self-associating protein oligomers from monomeric proteins using protein design tools? To address this question, we started with the simple model of treating the binding interface as a hydrophobic core and performing designs on heterodimer (Chapter 3) and homodimer (Chapter 4) models created by virtually docking the coordinates of the monomers. Through such efforts, we can explore the protein recognition problem by creating dimers purely from computational calculations.

Computational Approaches to The Protein Recognition Problem

Protein recognition is one of the most intriguing problems in biology. Proteins bind to molecules of all sizes and shapes and have binding interfaces fine-tuned for their substrates. A fundamental understanding of the recognition process is, therefore, crucial but extremely difficult to achieve. However, it is essential that we have reasonable models to describe the interactions.

Due to the importance of small molecules in drug discovery, various computational methods have been developed to predict small molecule binding affinities and ligand-receptor orientations[10]. One of the most sophisticated docking programs, DOCK, created by Kuntz *et al.*[11] and improved over the years, uses spheres to describe ligands and receptor cavities; by matching the sphere sets, ligand conformations and orientations can be scored. Although the shape matching features of methods like this can be scaled for protein-protein interactions, they are usually limited to local matches. Most protein-small molecule (or protein-drug) docking algorithms focus on simulating small ligands in the receptor environment. Protein design methods complement small molecule docking strategies by studying the recognition problem from a different perspective. Possible sequences for a receptor's binding pocket can be computed for a particular ligand based on known principles. In other words, a protein receptor can be "molded" around its ligand. By focusing on the steric contacts and optimizing hydrogen bonds, Looger *et al.* redesigned proteins with altered binding specificities, replacing their wild-type ligands with small molecules of interest[9]. As illustrated by this example, protein design can be used to identify and test mechanisms that drive complex formation and possibly to generate molecules with novel applications.

Protein-protein interactions, however, are less well characterized, despite continuing efforts to thoroughly analyze protein-protein interfaces[12-15]. The major obstacle lies in the seemingly endless ways Nature utilizes amino-acid side-chains. Furthermore, the problem is aggravated when backbone conformational changes are involved. Despite being a more complicated problem, the parallelism between methods to study protein-protein and protein-small molecule interactions is clear.

Protein design methods can similarly complement docking approaches in studying protein-protein interactions. Computational design has been successful in both altering binding specificities[6] and creating novel molecules[16] for a well characterized model system like coiled-coils. Several other systems have also been studied, and have recently been reviewed by Kortemme and Baker[17]. To make a protein complex design successful, a few key points must be addressed. First, high quality models must be available. Since current protein design algorithms require high-resolution structures, the best starting point is to use the backbones of experimentally determined crystal structures. This is the strategy used by several groups[6,18], including ourselves[19]. It is also possible to use backbone models that closely resemble crystal structures as in the work of Chevalier *et al.*[8].

Second, negative design can be important. Most protein design methods explicitly consider only the folded state of a molecule, and implicitly account for the unfolded state through energy function descriptions. While this "single state" design strategy simplifies the design problem for a monomer, it may not be adequate to describe a system where multiple states are involved. A more realistic description requires optimization of the target state while taking into account other competing states. Havranek and Harbury developed a strategy that takes into account the unfolded state, the folded homodimer conformation, the folded heterodimer conformation and the aggregated state of a dimeric coiled-coil model system; this strategy resulted in a designed specificity that was confirmed by experimental results[6]. A thorough design effort should adopt a similar approach, as illustrated by Bolon *et al.*[7], who found that sequences favoring heterodimer formation were generated by optimizing the energy of the heterodimer relative to that of the native homodimer sequence. However, these studies are usually too computationally demanding if more than a few residue positions are to be optimized.

Third, the energy functions should be reasonably accurate. This is where most of the variability lies when comparing methods used by different research groups. The ability to computationally design protein side-chains is based on a subtle balance between the force field terms, and there are many different ways to describe the same phenomenon. This is especially true for the solvation models used in computational design. Various methods with different levels of accuracy are used depending on the application, whether the consideration be speed, pair-wise decomposability, historical success, etc. For protein-protein interactions, hydrogen bonding and electrostatics terms are found to be important[20,21] in addition to solvation terms, which are considered most critical. Sometimes it is necessary to adjust the relative weights of the terms to achieve desired results, as illustrated in Shifman and Mayo's calmodulin designs; they found that tweaking the electrostatic terms and applying bias towards inter-molecular interactions yielded more successful results[20]. Because the chemical moieties that line protein interfaces vary greatly, extra care should be taken to address force field issues.

Last, since no current protein design algorithms can explicitly model water-mediated hydrogen bonds, they should be avoided if possible until the required methods become available.

One of the major drawbacks of the current computational approaches to the protein recognition problem, for both docking algorithms and design algorithms alike, is the inability to accurately predict or simulate backbone conformational changes. For docking algorithms, the most difficult predictions are for "unbound" situations. Docking algorithms are usually applied to two different kinds of problems, "bound" and "unbound" cases. The bound cases refer to the situations where the structures for both the ligand and the receptor are known in the bound conformation, most likely from a co-crystal of the complex. In such cases, each member in the complex is separated artificially for the subsequent docking tests. The purpose of such an exercise is to reconstruct the crystal conformation, usually as a test of the validity of a docking algorithm. The unbound cases refer to situations where each member of a complex is crystallized separately. It is conceivable that for a docking algorithm to perform well in unbound tests, conformational changes either on the side-chain or backbone level have to be accounted for.

Three different kinds of conformational changes have been proposed[22]. The first comprises the fast, small-scale thermal fluctuations observed in proteins, such as an ensemble of NMR structures. The second involves the slow, hinge bending type of domain movements. The third results from disorder that cannot be observed in either X-ray crystallography or NMR experiments; only upon complex formation, the intrinsically disordered part of the protein shifts its equilibrium to achieve a population time sufficient for detection by X-ray or NMR experiments. Although rigorous perturbation methods combined with molecular dynamics (MD) simulations can be used to model conformational changes reasonably accurately[23], they are not efficient for docking studies.

Most docking methods available today do not handle conformational changes that involve domain movements or induced fits. Instead, proteins are treated as rigid bodies and flexibility is implicitly included by using “soft” scoring potentials. This will be discussed in a later chapter. In order to completely cover all possible binding sites on the protein surface, all six degrees of rotational and translational freedom must be considered. At this stage, docking algorithms are required to make trade offs between accuracy and speed. With their limited ability to model protein flexibility, most docking algorithms perform poorly in unbound cases.

Design algorithms, on the other hand, have proven valuable in creating molecules with altered interfaces, but fall short in creating novel binding sites. Due to the requirement of high quality backbones, design algorithms are largely limited by the structures available in the Protein Data Bank (PDB[24]). In an ideal scenario, design algorithms would be able to evaluate the interface between any two molecules in contact with each other and generate sequences that accommodate the conformation, given the conformation is reasonable. In reality, design algorithms can only be applied to known complex structures. Depending on the nature of the interface, there is the possibility that a better answer could be obtained if some part the backbone were allowed to move, but the tools required to explore this possibility are not yet available.

Generation of *De Novo* Protein Dimers

The body of this thesis presents a brand new approach to the protein recognition problem. We are interested in combining our powerful design tools with a docking protocol to examine the possibility of creating novel dimers from scratch. To this end, we need to develop a docking algorithm that serves a different purpose. Most docking algorithms developed to date are for predicting complexes, with the idea of creating a high throughput approach to identify possible complexes from unbound monomeric structures. Since the number of protein-protein complexes deposited in the PDB is relatively low compared to that of single proteins, and experimental identification of complexes are tedious and time consuming, it is necessary to use this type of computer based algorithm to facilitate the discovery of associating molecules. But for our purpose, the docking algorithm is used to position the individual monomers to a plausible configuration while taking no information from surface side-chains. The factors involved in this docking process include: (1) a surface representation of the molecule, (2) a fast search algorithm to screen dimer conformations, and (3) an evaluation procedure to identify plausible targets. These factors are discussed in detail in Chapter 2.

By combining both the docking and design strategies in creating self-associating dimers, we examine the possibility of using a highly complementary hydrophobic surface patch to achieve oligomer assembly. Two successful design cases are presented, one heterodimer and one homodimer, in Chapter 3 and Chapter 4, respectively. Based on the knowledge gained from these studies, we believe that our strategy, which combines docking and design algorithms to address the protein recognition problem, is generally applicable to creating novel binding sites with compatible binding partners. The results also suggest that using hydrophobic amino acids to drive dimer formation is physically plausible.

However, designing protein oligomers with affinities comparable to naturally forming complexes remains an astounding challenge. Currently, we are also far from being able to incorporate allostery – proposed to be one of the major reasons for the existence of protein quaternary assembly[25] – in our design efforts. Over the years, tremendous efforts have been focused on learning, rationalizing, and systematically dissecting the binding interfaces found in crystal structures; studies include detailed surveys, modeling, calculations, etc. These efforts have gathered a vast amount of knowledge about protein-protein recognition. Attempts to apply this knowledge in building artificially designed functional proteins, however, are just beginning. With our small contribution to this mammoth problem, we hope to lay the first stone in building the foundation for future studies.

References

1. Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82-7 (1997).

2. Gordon, D. B., Marshall, S. A. & Mayo, S. L. Energy functions for protein design. *Curr Opin Struct Biol* **9**, 509-13 (1999).

3. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 14274-14279 (2001).

4. Bolon, D. N., Voigt, C. A. & Mayo, S. L. De novo design of biocatalysts. *Current Opinion in Chemical Biology* **6**, 125-129 (2002).

5. Kortemme, T. et al. Computational redesign of protein-protein interaction specificity. *Nature Structural & Molecular Biology* **11**, 371-379 (2004).

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

7. Bolon, D. N., Wah, D. A., Hersch, G. L., Baker, T. A. & Sauer, R. T. Bivalent tethering of SspB to ClpXP is required for efficient substrate delivery: A protein-design study. *Molecular Cell* **13**, 443-449 (2004).

8. Chevalier, B. S. et al. Design, activity, and structure of a highly specific artificial endonuclease. *Molecular Cell* **10**, 895-905 (2002).

9. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. Computational design of receptor and sensor proteins with novel functions. *Nature* **423**, 185-190 (2003).

10. Taylor, R. D., Jewsbury, P. J. & Essex, J. W. A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design* **16**, 151-166 (2002).

11. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *Journal of Molecular Biology* **161**, 269-288 (1982).

12. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. Dissecting subunit interfaces in homodimeric proteins. *Proteins-Structure Function and Genetics* **53**, 708-719 (2003).

13. Janin, J. & Seraphin, B. Genome-wide studies of protein-protein interaction. *Current Opinion in Structural Biology* **13**, 383-388 (2003).

14. Janin, J. & Wodak, S. J. in *Protein Modules and Protein-Protein Interactions* 1-8 (2003).

15. Wodak, S. J. & Janin, J. in *Protein Modules and Protein-Protein Interactions* 9-73 (2003).

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

17. Kortemme, T. & Baker, D. Computational design of protein-protein interactions. *Current Opinion in Chemical Biology* **8**, 91-97 (2004).

18. Reina, J. et al. Computer-aided design of a PDZ domain to recognize new target sequences. *Nature Structural Biology* **9**, 621-627 (2002).

19. Shifman, J. M. & Mayo, S. L. Modulating calmodulin binding specificity through computational protein design. *Journal of Molecular Biology* **323**, 417-423 (2002).

20. Shifman, J. M. & Mayo, S. L. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 13274-13279 (2003).

21. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14116-14121 (2002).

22. Halperin, I., Ma, B. Y., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins-Structure Function and Genetics* **47**, 409-443 (2002).

23. Lamb, M. L. & Jorgensen, W. L. Computational approaches to molecular recognition. *Current Opinion in Chemical Biology* **1**, 449-457 (1997).

24. Bernstein, F. C. et al. Protein Data Bank - Computer-Based Archival File for Macromolecular Structures. *Journal of Molecular Biology* **112**, 535-542 (1977).

25. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure* **29**, 105-153 (2000).

Chapter 2

Adapting a Fast Fourier Transform-Based Docking Algorithm for Protein Design

Abstract

Designing proteins with novel binding properties can be achieved by combining the powerful tools originally developed independently for protein docking and protein design. For the docking component, we use 3D grids to represent the protein molecules in space, and a fast Fourier transform-based algorithm to efficiently cover all translational dimensions and search through all six degrees of freedom. Proteins are modeled with a reduced representation that approximates the side-chains with spheres defined by experimentally determined atomic radii. C2 symmetry related homodimers are used to parameterize the docking algorithm, since their interfaces are more protein core-like and can be modeled well by our protein design algorithms. Imposing C2 symmetry also reduces the search space and significantly improves computational efficiency. The fitness of the docked structures are evaluated based on their surface shape complementarities. The resulting docking algorithmn successfully predicted the wild-type dimer conformations in 65 out of the 121 dimer test cases, with most of the predictions exhibiting less than 1 Å RMSD compared to the wild-type. The reduced protein representation therefore appears to be a reasonable estimate and can be used by our docking algorithm to position protein backbones in a plausible configuration for dimer design.

Surface Complementarity Between Two Molecules

The first step in our effort to generate *de novo* protein dimers involves the creation of a docking algorithm. The protein docking process described below is its most basic form.

Input to the docking algorithm consists of the structures of two molecules, a ligand and a receptor. Their initial spatial orientation with respect to one another can be random or assigned. The two molecules are brought to intimate proximity, hence "docked" to each other. Specifically, one molecule is moved and the other is held stationary. The molecule being moved is first translated in steps of a specified distance along the x, y, and z dimensions, until the steps cover all of the combined translations. For each translational step taken, the algorithm evaluates the docked conformation based on a scoring criterion. After exhausting all the translational steps, the molecules are returned to their initial positions, rotated a fixed increment with respect to each other, and the entire translational search is repeated. Assuming 150 steps are required for each of the translational axes and one degree increments are used for each of the rotational angles, as many as $360 \times 360 \times 180 \times 150 \times 150 \times 150$ (7.8732×10^{13}) steps may be required to exhaustively search all possible rotational and translational degrees of freedom. Using 150 steps in each translational dimension is reasonable if the two molecules are about 50 Å along the longest axis and if the step size is 1 Å.

One of the most important factors in this process is the method used to evaluate the fitness of the docked molecules. The key to this question depends on how the molecules are represented in space. Because of the large number of search steps involved, the molecular representations used for docking algorithms are usually not complicated. Although a complete molecular representation with all atom descriptions and side-chain rotamers is the most accurate, it is impractical for docking searches. The simplest and most commonly used method describes the molecules as rigid bodies and uses surface shape complementarity to score the docked structures. Energy terms can also be used, but these are usually limited to simple electrostatic and distance-restricted hydrogen bond terms. Molecular surface is a good way to describe rigid bodies for this purpose. A widely used method to calculate molecular surface was introduced by Connolly in 1983. This approach rolls water probes on the surface of proteins and places dots along the path to form a solvent-accessible surface[1]. Other representations that are also good approximations include treating protein surfaces as abstract concave and convex points; an example of this is the surface sparse critical points method introduced by Lin *et al*. in 1994[2]. The use of 3D grids is also very common among docking algorithms, especially for the methods that use fast Fourier transform (FFT) for translational searches. In one case, proteins are represented as spherical harmonics[3,4].

Each of these molecular representation methods has its advantages and disadvantages, and in general, there is a trade-off between speed and accuracy. Connolly's method produces very accurate molecular surfaces, but the number of dots generated for protein-protein docking searches is too large to be handled efficiently. The surface critical points method offers great docking speed, but does not capture interactions in atomic detail. Even though the 3D grid approach may require a large number of grids to represent the molecules accurately, it can still be very fast when FFT is used in the translational searches. The 3D grid representation thus offers a good balance between speed and accuracy.

The molecular representation method chosen can also significantly affect the physical model used to evaluate the docked dimers. For example, electrostatic properties of the protein surface can be embedded in the 3D grids and evaluated alongside surface complementarities, thus facilitating the identification of favorable dimer conformations by ruling out those with disfavorable electrostatic interactions[5].

In our dimer design, the side-chains are explicitly designed in the subsequent steps using the ORBIT protein design software. The docking algorithm must therefore generate a list of plausible dimer orientations based on a molecular representation that builds estimated side-chains on a poly-alanine backbone. A crystal structure is used as the scaffold for each of the monomers, with the side-chains beyond the C_β atom deleted. To maintain the overall shape of the surface, the volume originally occupied by a side-chain cannot be left empty. Therefore, the most important criterion for our choice of a molecular representation is its ability to estimate this void space relatively accurately and easily. In theory, it is possible to use the original side-chains in the docking process, then subsequently replace them during design. If we use this full side-chain representation, however, we will need a very "soft" scoring function to allow surfaces to overlap, which can lead to backbone clashing. Since we are not allowing backbone flexibility, this kind of clashing is strictly prohibited. Moreover, because the side-chains on the surface of a protein are usually longer than those found in the core, the use of full surface side-chains in the docking process would make the creation of a hydrophobic interface difficult. If the two halves of a dimer are positioned implausibly far away from each other, the design algorithm cannot make good sequence predictions for the interface. The molecular representation most suitable for our application is therefore one that allows easy “padding” in the side-chain voids while maintaining the topology of a potential binding surface. These requirements can be met using 3D grids. Since our poly-alanine representation of the backbone does not consider the chemical properties of the side-chains, our docking algorithms uses a simple scoring function that only evaluates surface complementarities.

3D Grid Discretization

We selected the 3D grid method to represent our protein molecules in space for the purpose of evaluating surface shape complementarities. In addition to being the best molecular representation for the requirements mentioned above, this approach has the distinct advantage of allowing the use of FFT for translational searches, which significantly decreases the compute time. Details on the use of FFT will be covered later in this chapter. Representing protein molecules with 3D grids requires a discretization process where grid points are assigned as either part of the protein or part of the empty space surrounding it. The 3D grids in our implementation correspond to the members of a one dimensional computer data storage array with each member representing a cubic box in space, called a voxel (Figure 2-1).

Before voxel assignments can begin, the cubic 3D space must be defined. This includes specifying: (1) the size of the entire arena, (2) the resolution to be used for the docking exercise, and (3) the number of array elements to be used. The size of the arena depends on the dimensions of the protein molecules to be discretized. Because calculations will be performed on every single one of the voxels, the arena size should be as small as possible while leaving enough room for the proteins to move around each other along all three translational axes. A good rule of thumb is to make each of the dimensions of the arena three times the length of the protein's long axis. The length across the arena must be large enough to simultaneously hold two copies of the mobile protein (usually the smaller molecule) and one copy of the stationary protein, as this is the minimum requirement for one molecule to see both sides of the other molecule by translation. Because one of the molecules is usually held stationary, if the size of the arena fits only one copy of each molecule, some parts of the stationary molecule will never be checked. While a smaller arena is sometimes used when docking to a focused region, it should never be allowed in a global docking search.

The resolution of a docking calculation is determined by the relationship

$$\text{resolution} = \frac{\text{size of the arena}}{\text{number of equivalent grid points}}$$

and is not usually explicitly defined. A smaller numerical value means a higher resolution and a better approximation of the proteins by the 3D grids. Since the size of the arena is set by the size of the molecules, the resolution is usually limited by the number of equivalent grid points, which is in turn dependent on the amount of physical memory available on the computer. The discretization process involves checking every atom on the proteins against the grid boundaries. Each grid boundary is evaluated in turn, and if any part of its sides or edges falls within the van der Waals radii of an atom, a value is assigned to the grid voxel (Figure 2-2). It is convenient to use the 3D grid representation for our purpose because we can easily make a projection of the atomic radii and have grids assigned to the volumes that are originally taken up by the side-chains (Figure 2-3).

Structural Correlations

In order to evaluate the surface complementarities between the two molecules represented by 3D grids, the two molecules must be discretized separately into two storage arrays that cover the same arena. For every round of rotational and translational search, the array that contains information about the stationary molecule is not changed, but the array that holds the mobile molecule is regenerated according to the molecule's new orientation and spatial position. After being discretized, the proteins can be treated as discrete functions (or signals), and the correlation between the two discrete functions can be obtained by summing the products of the two signal amplitudes at each sampling interval (Figure 2-4); the sampling interval can be time delay, position steps, or any property that describes the two functions. The correlations between two functions A and B can also become a function of the same variable (Figure 2-4). In our case, the correlation between the two protein molecules discretized into 3D grids can be calculated using the same principle. Since our interest is in finding good correlations between the surfaces of the two molecules, we can shape the individual discrete functions to achieve this.

We adopted a scoring scheme similar to the one used by Katchalski-Katzir et al., where all voxels for the mobile molecule are assigned the value "1", those for the stationary molecule are assigned different values according to their locations, and voxels that are not part of the protein are assigned "0"[6]. The voxels associated with the stationary protein are categorized as either core or surface, with the core covering the space around the atoms defined by some radii, and the surface covering the space between the core and 1.5 Å beyond the core. For the stationary molecule, grid points corresponding to the surface are favored and given positive values (usually “1”) while the ones in the core are penalized with negative values (usually “-15”) (Figure 2-5). By setting up the grid points this way, we can evaluate surface complementarity by counting the number of overlapping surface voxels between the molecules. Assuming “-15” is used for the voxels in the core, fifteen overlapping surface voxels are required to offset every voxel from the mobile molecule that penetrates the core of the stationary molecule. When docking is carried out at a high resolution, where the interface consists of hundreds of voxels, the use of a small penalty value such as “-15” allows slight penetration to the core while maintaining a high level of correlation on the surface. Therefore, the scoring function is intrinsically “soft” when a small penalty is used.

The scoring function described above, however, is not the most appropriate one for protein design purposes, since it provides no distinction between side-chain and backbone penetrations. Since a reduced representation of the surface side-chains is used in our docking protocol, some penetration on the side-chain level is considered favorable, as this will create more surface overlap and possibly make the designed interface more viable. Backbone penetrations, on the other hand, must be prohibited. To account for this, a third category of voxel scores was created. In addition to having values of “0” (for vacuum), “1” (for favorable surface) and “-15” (for unfavorable but allowed penetration), a voxel can also be assigned the value of “-1000” if it falls within 1 Å of an atom center. Although rarely needed, this third “hard shell” ensures no backbone clashes during docking.

If N is the number of grids used in each dimension, evaluating the correlation between two molecules at a particular conformation using the method described above requires $N^3+(N^3-1)$ calculations (N^3 multiplications and (N^3-1) additions). To complete an entire correlation map (correlation as a function of translations), the total number of computational steps involved is on the order of N^6. Calculations of this size still take a long time to finish, even on modern computers. Fortunately, by using fast Fourier transform, the problem can be reduced to $N^3 \mathrm{Log}_2 N^3$.

Fourier Transform and Fast Fourier Transform

This section partly summarizes the material used in our program that is adapted from "Numerical Recipes in C," Cambridge University Press, Second Edition, Chapters 12 and 13, pages 496 - 546. [7]

Fourier transform, named after Jean Baptiste Joseph Fourier (1768 – 1830), is described as a continuous generalization of the complex Fourier series to none periodic functions (when periodic boundary $\rightarrow\infty$). It is widely used in the scientific and engineering community as an efficient computational tool for data processing and has several interesting properties. Fourier transform converts the same function between the time domain and the frequency domain,

$$h(t) = \int_{-\infty}^{\infty} H(f) e^{-2\pi i f t} df \qquad \text{Eq. 2.1}$$

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{2\pi i f t} dt \qquad \text{Eq. 2.2}$$

where $h(t)$ is the function that describes a physical process of some value h as a function of time t, and $H(f)$ is the function that describes the same physical process with its amplitude H as a function of frequency f. Here,

$$h(t) = F_f[H(f)](t) = \int_{-\infty}^{\infty} H(f) e^{-2\pi i f t} df \qquad \text{Eq. 2.3}$$

is the “inverse” (-*i*) Fourier transform, and

$$H(f) = F_t^{-1}[h(t)](f) = \int_{-\infty}^{\infty} h(t)e^{2\pi ift}dt \qquad \text{Eq. 2.4}$$

is the “forward” (+*i*) Fourier transform. The “forward” and “inverse” transforms can be used to convert the same function back and forth between the two different representations, and *h*(*t*) and *H*(*f*) are described as a “transform pair,” denoted

$$h(t) \Leftrightarrow H(f).$$

In essence, Fourier transform decomposes a function into sinusoids of different frequencies whose sum returns the original function. Several interesting properties can be obtained when operating the function or functions in the frequency domain, such as convolution and correlation of two functions. For protein docking purposes, we are interested in the correlation theorem.

The correlation between two real functions $h(t)$ and $g(t)$ (with their Fourier transform represented as *H*(*f*) and *G*(*f*), respectively) can be expressed as

$$C(t) = \int_{-\infty}^{\infty} g(\tau + t)h(\tau)d\tau \qquad \text{Eq. 2.5}$$

where C(*t*) is the correlation of *h*(*t*) and *g*(*t*) as a function of *t*. The expression in Eq. 2.5 corresponds to this transform pair:

$$C(t) = \int_{-\infty}^{\infty} g(\tau + t)h(\tau)d\tau \Leftrightarrow G(f)H(-f) \qquad \text{Eq. 2.6}$$

Since by definition, for real functions

$$H(-f) = H^*(f)$$

where $H^*(f)$ is the complex conjugate of *H*(*f*), the expression in Eq. 2.6 can be written as

$$C(t) \Leftrightarrow G(f)H^*(f)$$

The correlation of two real functions can be obtained simply by multiplying the Fourier transform of one function with the complex conjugate of the Fourier transform of the other.

In real applications, function *h*(*t*) is presented as data sampled at evenly spaced intervals, and the approximation of the integral in Eq. 2.2 for this type of function is the discrete Fourier transform (DFT):

$$H(f_n) = \int_{-\infty}^{\infty} h(t) e^{2\pi i f_n t} dt \approx \sum_{k=0}^{N-1} h_k e^{2\pi i f_n t_k} \Delta = \Delta \sum_{k=0}^{N-1} h_k e^{2\pi i k n / N} \qquad \text{Eq. 2.7}$$

$$f_n \equiv \frac{n}{N\Delta}, \qquad n = -\frac{N}{2}, \ldots, \frac{N}{2}$$

where *n* defines the values at which each frequency will be evaluated, *N* is the total number of sampling points, and Δ is the sampling interval. Note that t_k is also defined as

$$t_k \equiv k\Delta,$$

and as *k* is incremented in the summation, since

$$h_k \equiv h(t_k),$$

Eq. 2.7 integrates over the entire data. The discrete Fourier transform maps N complex numbers into N complex numbers through N × N complex additions and multiplications. The last term in this equation,

$$H_n \equiv \sum_{k=0}^{N-1} h_k e^{2\pi i k n / N} \qquad \text{Eq. 2.8}$$

is referred to as the discrete Fourier transform of the N points h_k.

The inverse transform has a strikingly similar form:

$$h_k = \frac{1}{N} \sum_{n=0}^{N-1} H_n e^{-2\pi i k n / N} \qquad \text{Eq. 2.9}$$

The only differences between the "forward" and the "inverse" transforms are the additional factor of 1/N and the change of the sign in the exponential term.

Discrete Fourier transform traditionally was computed by defining

$$W \equiv e^{2\pi i/_N} \qquad \text{Eq. 2.10}$$

and thus Eq. 2.8 can be written as:

$$H_n \equiv \sum_{k=0}^{N-1} W^{nk} h_k \qquad \text{Eq. 2.11}$$

The Fourier transform of the discrete data h_k can then be obtained by multiplying the data with a matrix whose elements are W's to the power of the product of their indices. Because W is actually periodic with period N, when N = 4, for example, the (2, 3) element, $W^{2\times 3} = W^6 = W^2$. An 8-point DFT can be written as:

$$\begin{bmatrix} H_0 \\ H_1 \\ H_2 \\ H_3 \\ H_4 \\ H_5 \\ H_6 \\ H_7 \end{bmatrix} = \begin{bmatrix} W^0 & W^0 & W^0 & W^0 & W^0 & W^0 & W^0 & W^0 \\ W^0 & W^1 & W^2 & W^3 & W^4 & W^5 & W^6 & W^7 \\ W^0 & W^2 & W^4 & W^6 & W^0 & W^2 & W^4 & W^6 \\ W^0 & W^3 & W^6 & W^1 & W^4 & W^7 & W^2 & W^5 \\ W^0 & W^4 & W^0 & W^4 & W^0 & W^4 & W^0 & W^4 \\ W^0 & W^5 & W^2 & W^7 & W^4 & W^1 & W^6 & W^3 \\ W^0 & W^6 & W^4 & W^2 & W^0 & W^6 & W^4 & W^2 \\ W^0 & W^7 & W^6 & W^5 & W^4 & W^3 & W^2 & W^1 \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \end{bmatrix} \qquad \text{Eq. 2.12}$$

The resulting vector H_n with N elements can be obtained from the input vector h_k with N elements[8].

The fast Fourier transform (FFT) relies on the clever trick proposed by Danielson and Lanczos in 1942. They demonstrated that a discrete Fourier transform of length N can be rewritten as the sum of two transforms, one formed by the transform of even-numbered points of the original data (with length N/2) and the other by the transform of odd-numbered points (also with length N/2):

$$F_k = F_k^{even} + W^k F_k^{odd} \qquad \text{Eq. 2.13}$$

This is called the *Danielson-Lanczos Lemma*. Instead of solving one complicated transform, it reduces the problem to two simpler transforms. Furthermore, this reduction can be applied recursively, splitting up the even and odd transforms to the transforms of their N/4 even-numbered data and N/4 odd-numbered data. This can go all the way until the data is subdivided down to transforms of length 1. Each element of the resulting vector H_n becomes a summation of the input array h_k with some combinations of W's as the coefficients for each element in the h_k array.

Because of this splitting of even and odd-numbered data, FFT is most efficient when the data length N is an integer power of 2. The bookkeeping on which combination of W's would go with which transform is dealt with by "bit-reversal", which is a process that involves swapping memory space in computers and can be done efficiently (Table 2-1). By using *Danielson-Lanczos Lemma* and bit-reversal, FFT has the advantage of reducing the N^2 calculations needed for obtaining a DFT to $N\log_2 N$ calculations. The improvement in speed is immense. The matrix representation of the same problem in Eq. 2.12 is now shown bit-reversed and factored:

$$\begin{bmatrix} H_0 \\ H_1 \\ H_2 \\ H_3 \\ H_4 \\ H_5 \\ H_6 \\ H_7 \end{bmatrix} = \begin{bmatrix} W^0 & W^0 & W^0 & W^0 & W^0 & W^0 & W^0 & W^0 \\ W^0 & W^4 & W^2 & W^6 & W^1 & W^5 & W^3 & W^7 \\ W^0 & W^0 & W^4 & W^4 & W^2 & W^2 & W^6 & W^6 \\ W^0 & W^4 & W^6 & W^2 & W^3 & W^7 & W^1 & W^5 \\ W^0 & W^0 & W^0 & W^0 & W^4 & W^4 & W^4 & W^4 \\ W^0 & W^4 & W^2 & W^6 & W^5 & W^1 & W^7 & W^3 \\ W^0 & W^0 & W^4 & W^4 & W^6 & W^6 & W^2 & W^2 \\ W^0 & W^4 & W^6 & W^2 & W^7 & W^3 & W^5 & W^1 \end{bmatrix} \begin{bmatrix} h_0 \\ h_4 \\ h_2 \\ h_6 \\ h_1 \\ h_5 \\ h_3 \\ h_7 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & . & . & . & W^0 & . & . & . \\ . & 1 & . & . & . & W^1 & . & . \\ . & . & 1 & . & . & . & W^2 & . \\ . & . & . & 1 & . & . & . & W^3 \\ 1 & . & . & . & W^4 & . & . & . \\ . & 1 & . & . & . & W^5 & . & . \\ . & . & 1 & . & . & . & W^6 & . \\ . & . & . & 1 & . & . & . & W^7 \end{bmatrix} \begin{bmatrix} 1 & . & W^0 & . & . & . & . & . \\ . & 1 & . & W^2 & . & . & . & . \\ 1 & . & W^4 & . & . & . & . & . \\ . & 1 & . & W^6 & . & . & . & . \\ . & . & . & . & 1 & . & W^0 & . \\ . & . & . & . & . & 1 & . & W^2 \\ . & . & . & . & 1 & . & W^4 & . \\ . & . & . & . & . & 1 & . & W^6 \end{bmatrix} \begin{bmatrix} 1 & W^0 & . & . & . & . & . & . \\ 1 & W^4 & . & . & . & . & . & . \\ . & . & 1 & W^0 & . & . & . & . \\ . & . & 1 & W^4 & . & . & . & . \\ . & . & . & . & 1 & W^0 & . & . \\ . & . & . & . & 1 & W^4 & . & . \\ . & . & . & . & . & . & 1 & W^0 \\ . & . & . & . & . & . & 1 & W^4 \end{bmatrix} \begin{bmatrix} h_0 \\ h_4 \\ h_2 \\ h_6 \\ h_1 \\ h_5 \\ h_3 \\ h_7 \end{bmatrix}$$

Eq. 2.13

where "." represents zero[8]. By comparing the last set of matrix multiplication in Eq. 2.13 to those in Eq. 2.12, it should be straightforward to see the advantage FFT provides, since it is easier to multiply a vector by the three sparse matrices in Eq. 2.13 than the one dense matrix in Eq. 2.12. FFT reduces a problem of $O(N^2)$ complexity to $O(N\log_2 N)$. In our docking calculations where there are N grid points in each dimension, the data length is N^3 (for x, y, z dimensions), and therefore the problem is reduced from N^6 to $N^3\log_2 N^3$. Using the FFT correlation theorem to study protein docking was first introduced by Katchalski-Katzir et al. in 1992[6]. The correlation map between two molecules can easily be obtained by first applying FFT to both molecules, followed by multiplying the Fourier transform of one molecule by the complex conjugate of the Fourier transform of the other, and finally inversely transforming the data. The resulting correlation map contains information about every translational step the mobile molecule can take and the corresponding correlation for each of these moves. Effectively, the FFT correlation map covers all translational searches without having to physically move the mobile molecule across the arena. Instead of searching through all six rotational and translational degrees of freedom by brute force, the docking algorithm must only explicitly define rotational steps.

C2 Symmetry Related Dimers

Homodimers are dimers composed of identical subunits[10]. They are by far the most common type of protein assembly[11] and are well represented in the Protein Data Bank (PDB)[9]; about 30% of *E. coli* proteins are homodimers. Although there are exceptions such as hexokinase, most homodimers can be described by a single two-fold axis of rotational symmetry, characterized by the symmetry group C2. Due to the cyclic nature of this type of assembly, each subunit contributes equally to form the binding interface. While C2 symmetry related homodimers are the simplest type of protein assembly, their interfaces are still very diverse. Many are obligatory dimers, as they are permanent assemblies that self-associate upon folding. Their highly coupled folding and binding processes are largely the result of having protein core-like interfaces. On average, homodimer interfaces are more hydrophobic and bury twice as much surface area as heterodimer interfaces[10]. Their hydrophobic interfaces and symmetry make homodimers especially suitable for design with ORBIT. ORBIT uses a force field that has been experimentally validated through the correlation of the calculated energies and stabilities of the designed proteins[12]. A well-packed core often contributes significantly to protein stability, and ORBIT models interactions in the core relatively accurately. Homodimers are therefore plausible targets for (1) parameterizing our docking algorithm so that the computationally docked molecules are likely to have interfaces that accept hydrophobic residues and (2) redesigning the interfaces since we can treat them as protein cores.

The use of symmetry offers several important advantages. Most significantly, C2 symmetry can greatly reduce the search space required for the docking process, and by imposing rotameric symmetry on the subunits, the complexity of the design can be simplified. Using FFT in conjunction with C2 symmetry in the matching stage provides an additional reduction in computational steps. All of these features were incorporated into our docking algorithm.

Some of the reduction in search space results from redundancy associated with C2 symmetry. This can be explained using a coordinate system composed of two symmetry related coordinate systems as follows. According to Euler's rotation theorem, any rotation can be described by a set of three angles called Euler angles. There are many conventions for the Euler angles; we can depict the concept simply by using the ZXZ convention. In this convention, the three Euler angles, ϕ, θ, ψ, are defined as follows: ϕ is the first rotation ranging from 0 to 2π about the Z-axis, θ is the second rotation ranging from 0 to π about the x'-axis, and ψ is the third rotation ranging from 0 to 2π about the z'-axis (Figure 2-6). These three rotations are not commutative, and therefore must be applied in this specific order. The three angles are depicted in Figure 2-6 with a modified coordinate system to illustrate the search space reduction associated with C2 symmetry.

In a Cartesian coordinate system, to thoroughly explore the rotations of a rigid body with respect to the coordinate system (or with respect to another rigid body in the case of docking searches), the rotational space that must be covered is $2\pi \times \pi \times 2\pi$, as defined by the full ranges of the three Euler angles. In Figure 2-6, the coordinate system shown can be described as a combination of two separate coordinate systems, XYZ and $\xi\eta Z$, related to each other by a two-fold (C2) symmetry about the Z-axis. The quadrants covered by the $\xi\eta Z$ coordinate system are shaded. By the definition of C2 symmetry, any point in the XYZ coordinate system mirrors to a point in the $\xi\eta Z$ system by a 180° rotation. Due to the 180° rotational symmetry with respect to the Z-axis, the ranges of ϕ and ψ in this XYZ, $\xi\eta Z$ combined coordinate system are both reduced by half to $0 \leq \phi \leq \pi$, $0 \leq \psi \leq \pi$, while the range of θ remains the same. Rotations beyond the range of 0 to π are redundant since the resulting positions can always be folded back to positions within the range of 0 to π, as illustrated by vectors x' and x''. Vector x' is obtained by rotating vector x^o about the Z-axis by ϕ, and by C2 symmetry it mirrors to the vector x'', which can also be obtained by rotating vector x^o by $\phi + \pi$. Due to this redundancy, the range of ϕ can be reduced from 2π to π, and this is also true for ψ.

An additional reduction of rotational search space can be achieved when translational searches are performed. This concept is illustrated in Figure 2-7. In order to maintain the C2 symmetry, rotations performed on the subunits must be synchronized – the same rotational operation must be performed on both molecules. For clarity, the two molecules in Figure 2-7 are set at a fixed distance from each other when they are rotated, and the rotations are performed at their respective geometric centers about axes that are parallel to the symmetry axis. One of the properties of cyclic symmetry groups such as

C2 is that the subunits are related by rotation about a symmetry axis and they are always on a plane perpendicular to the symmetry axis. As illustrated in Figure 2-7, when each of the subunits of the C2 symmetry related dimer is rotated to a new C2 symmetry related conformation, the rotational steps required to achieve this new conformation can always be replaced by translations on this plane. Therefore, when translational steps are included in a docking search, rotations around the symmetry axis or any axis parallel to the symmetry axis (defined by ϕ) can be eliminated from the searches. FFT, incidentally, can replace explicit translational searches with an efficient computational process that generates the translational correlation map. As a result, to thoroughly search all possible C2 related dimer conformations, we only need to cover $\pi \times \pi$ (the ranges of θ and ψ) when FFT is used; the search space is reduced by a factor of 4π.

The computer memory required for discretizing the molecules can also be reduced when docking C2 symmetry related dimers. As described previously, the cyclic symmetry requires the subunits of a C2 related dimer to be on the same 2D plane. This requirement eliminates the need to explore any translations parallel to the symmetry axis. If the Z-axis is used as the symmetry axis for a pair of dimers, only translations along the X- and Y-axis are required to produce dimers with preserved C2 symmetry. The dimension of the arena that is parallel to the symmetry axis can therefore be reduced to the length of the long axis of the molecule instead of three times this length. The FFT implementation used in our docking algorithm, however, requires the number of grid points along each dimension to be a power of 2, and thus it is convenient to simply reduce the dimension parallel to the symmetry axis by a factor of 2. The number of grid points required for the search becomes half (Figure 2-8). Assuming 1° is used as the rotational increment, docking a C2 related dimer is 1440 times faster than docking a dimer with no symmetry (4π rotational search space reduction (Figure 2-6) times two-fold grid point reduction (Figure 2-8) equals a total reduction of 8π; for 1° rotational increments, this equals 8 × 180 or 1440-fold reduction).

Determination of Practical Atomic Radii

Because the side-chains are truncated in our protein models, we approximate the side-chain volumes with spheres centered at the C_β atoms (Figure 2-3); the projection distance from the C_β atoms (atomic radii) should be chosen to ensure that the resulting dimer interface retains enough space for side-chain placement. We determined an appropriate C_β atomic radius by calculating the surface complementarity scores for several high-resolution crystal structures of complexes in the PDB. The molecules were discretized in their native crystal conformations with the side-chains truncated, and surface complementarities were calculated. Initially, we tried several radii keeping a uniform radius for all the atoms. However, complexes that had backbone-to-backbone hydrogen bonds always clashed at these points. It was obvious that a uniform radius was inadequate, so we decided to parameterize the radii for each of the five atom types in our poly-alanine model. We tested a range of values for nitrogen, C_α, C_β, carboxyl carbon, and oxygen. The crystal structures of two high-resolution complexes were used: PDB entries 1a7w and 1c9o (1.55 Å and 1.17 Å resolution, respectively). All possible combinations of atomic radii within the following ranges were tested: nitrogen between 1.4 Å and 1.6 Å, oxygen between 1.3 Å and 1.5 Å, and all carbons (carboxyl carbon, C_α and C_β) between 1.7 Å and 2.4 Å. A 0.05 Å increment was used to step across each of the ranges. The resulting scores were sorted and the radii combinations that gave top ranking scores were analyzed; about 50 to 100 radii combinations were collected.

This cutoff is somewhat arbitrary, since the correlation score can be very sensitive to changes in radii for some atom types but not for others (Figure 2-9). The level of sensitivity is reflected in the "good radii" ranges shown in Table 2-2; a wider "good radii" range indicates that the correlation score is less sensitive. The values used for subsequent docking tests were determined as follows. If the radii ranges obtained from both structures were about equal, as for atom types N and O, the smallest range value was used. If the radii ranges from the two structures matched poorly, as for atom types C, C_α and C_β, the mean value of the smaller range was used.

The final atomic radii values obtained from the parameterizations (Table 2-2, last column) were used to test the docking algorithm. Two structures (PDB entries 1ecz and 1rfb) were chosen as test cases. Both are interlocking homodimers, with 1ecz making backbone-to-backbone contacts and 1rfb interacting mostly via side-chains. The dimer conformations predicted using the docking algorithm showed good agreement with the crystal structures (Figure 2-10). The result from docking 1ecz was particularly encouraging, exhibiting a C_α root mean square deviation (RMSD) of 0.45 Å. The two identical subunits in the 1ecz structure bind to each other by making backbone-to-backbone hydrogen bonds through their carboxyl terminal tails. Our successful prediction for the 1ecz dimer suggests that the radii used for atom types N, O, C, and C_α are plausible. However, the task of parameterizing the C_β atoms remains a challenge. The dimer conformation predicted for the 1rfb structure showed a less promising C_α RMSD of 2.38 Å when compared to the wild type crystal structure. Closer examination revealed that the offset is likely due to the phenylalanine and tyrosine residues in the interface. Aromatic residues have relatively long side-chains and cannot be modeled well by spheres.

The accuracy of three different sphere approximations is illustrated in Figure 2-11. A peptide chain with full atom van der Waals spheres is shown as our reference on the left (a). Poly-alanine approximations of the peptide chain using spheres of different radii are shown in b, c, and d; hydrogen atoms are not included in these models. When standard van der Waals radii are used (b), the volume of the peptide is underestimated (compared to a). Better approximations are produced using a uniform all-atom radius of 2.15 Å (c) or our experimentally determined radii for each of the atom types (d). We concluded that modeling a peptide with spheres centered on the C_β atoms results in fairly good approximations for short side-chains, but falls short in estimating aromatic and other long side-chains.

To test the general applicability of this conclusion, we parameterized the atomic radii again using an expanded set of protein structures. The results for atom types N, O, C, and C_α remained roughly the same as our initial results (in Table 2-2), but the "good radii" for C_β expanded to cover a broader range (1.8 ~ 2.55 Å vs. initial range of 2.15 ~ 2.25 Å); we also noted that the optimal C_β value for a given structure depends on the types of amino acids in the interface (data not shown). The mean value of this range, 2.15 Å, appears to be the right size for shorter non-polar side-chains, and is coincidentally the same as the final value obtained in the first round of testing. Another interesting observation that can be made from this expanded test set relates to the radii obtained for nitrogen and oxygen atoms. For proteins with backbone-to-backbone contacts, the sum of the optimal N and O radii is about 2.7 Å, which is the distance for a hydrogen bond. 1ecz, one of our initial test cases, forms extensive hydrogen bonds across the dimer interface; the N and O radii used in docking this dimer satisfy this distance requirement and may explain the low RMSD observed. The results from the expanded test set generally confirmed our initial values, so we used the final atomic radii (in Table 2-2) for all subsequent docking calculations.

Testing the Docking Algorithm

Even though our docking algorithm was developed to design novel dimers, it contains all the basic components found in docking algorithms that use FFT as a search tool and can be used as a stand-alone algorithm for predicting dimer conformations. However, because of differences in the nature of the problems to be addressed, our algorithm should not be compared to other general docking algorithms. Protein docking is an area of active research, and much progress has been made in developing algorithms suited to this purpose[13-25]. The trend is to model proteins with greater accuracy either through implicit energy terms or explicit simulations; this includes the incorporation of desolvation terms, electrostatic charges, side-chain flexibility, Monte Carlo simulations, etc. Our algorithm, on the other hand, relies largely on approximations, some of which have no physical basis. The biggest difference is in the handling of side-chains. While side-chains are very important in protein-protein interactions, our algorithm uses a simplified poly-alanine model for the side-chains and doesn't incorporate their chemical properties in the calculations. These approximations greatly reduce our chances of predicting the native conformation, especially if the driving forces are specific side-chain-to-side-chain interactions instead of surface geometric complementarities.

To minimize these problems, we tested the appropriateness of our docking algorithm by choosing test cases that utilize the same type of binding mechanism as core residues, since our protein design algorithm models protein cores fairly well. We selected homodimers, since their interfaces are on average more hydrophobic and protein core-like. We used the 122 homodimers chosen by Bahadur et al.[10], excluding one structure (1alo), which would not be appropriate for protein design due to its exceptionally large size.

The structures of the 121 dimers were downloaded in their biological unit coordinates from the PDB[26]. The protein coordinates were processed to clean up any naming and numbering discrepancies, and the subunits were separated into individual files. For each docking calculation, we loaded the coordinates of just one of the subunits (usually the subunit designated as chain A), and created the other subunit by duplicating and rotating the loaded coordinates by 180° about the x-axis. Except for orientation, the two subunits were thus identical to each other. The rotational search space was sampled with 1° increments over 180° for both the y- and the z-axes. Depending on subunit size, the number of grids used for the arena was either 128 or 256 in the y and z dimensions, and half this number in the x dimension. All tests were carried out at 1 Å docking resolution.

Docking searches and surface complementarities were calculated for each of the 121 dimers; coordinates for the 50 best-correlated docked conformations were generated for each dimer and compared to the coordinates of the wild type. RMSDs were evaluated between all the C_α atoms from both subunits of the docked conformations and their corresponding C_α atoms on the wild-type structures. The results are shown in Table 2-3; RMSDs are only listed if less than 3 Å. From the top 50 docked structures for each dimer, the rank with the lowest RMSD (best match) is listed along with the first rank with a RMSD < 3 Å. The buried interface surface area contributed by one of the subunits (reported as B/2) and the ratio of this surface area to the number of residues in the subunit (Area/Res) are also reported. The 121 dimers tested are sorted according to Area/Res in Table 2-4 along with indications of hits (shown with "+") and misses (shown with "."). A dimer is considered to be a hit if there is at least one docked dimer in its top 50 conformations with an RMSD < 3 Å.

We achieved 65 successful predictions (hits) out of 121 test cases, slightly above 50%. While there are no docking benchmarks focusing exclusively on homodimers, our results are comparable to the few homodimer docking cases reported previously, in which three of five test cases were within the top 50 ranked structures[5]. The ranks produced by our simple surface complementarity scoring scheme, however, do not always correlate with the RMSDs of the models. In several cases, the closest match to the wild-type conformation does not receive the highest correlation score. The best match is predicted as the top rank in only four cases (1ajs, 1tox, 1trk and 1vfr). This is not surprising since we are using a reduced model of the proteins. Since all four cases use backbones extensively to achieve binding specificity, it appears that only dimers with these properties are correctly ranked by our scoring scheme. The inability to rank models correctly is a problem that plagues all docking algorithms, and most research groups develop sophisticated scoring functions that include the properties of the side-chains to ameliorate this problem.

It should be noted that for our purpose the ranking does not significantly affect our results. We are interested in identifying all distinct dimer configurations for *de novo* protein design, and as long as the interface is plausible it is not necessary to recover the wild-type binding sites. When testing our docking algorithm, however, it is important to recover the wild-type binding sites (regardless of their reported ranks) and to obtain good matches when they are identified because these two factors directly check the validity of our atomic radii. As shown in Table 2-3, 45 of the 65 successfully docked dimers have RMSDs less than 1 Å, indicating that our reduced protein models are reasonable approximations. We found that we can reliably reproduce most of the wild-type dimers that have backbone-to-backbone interactions across the interface, namely the hydrogen bond pairing between two intermolecular β-strands (Figure 2-12). Examples of dimers making β-sheets across the dimer interface are shown at the top (A), and similar to the 1ecz test case described earlier, our practical atomic radii capture these hydrogen bonds well. However, for the helical proteins shown at the bottom (B), the overall dimer conformations are recovered, but the docked molecules are slightly offset from the wild-type configurations. The distances between the backbones of the docked 1rop dimer are too close, again indicating that the C_β radius used for this particular type of dimer is not large enough. Furthermore, because the docking algorithm searches for the highest complementarity between the dimers, inter-digitation between the spheres representing side-chains on the surface of a helix is preferred over stacking the spheres head-on.

Ironically, this feature is why we see hydrogen bonds positioned correctly even without the use of an explicit hydrogen bond definition.

The interface area-to-residue ratios (reported as Area/Res in Tables 2-3 and 2-4) illustrate another interesting point. These ratios are used as a rough measure of the relative size of the interface in the context of the entire subunit. Sorting the 121 dimers according to this ratio reveals that our docking success rate is much higher for dimers that bury relatively large surface areas compared to those that do not. For example, our success rate is 70% for dimers with Area/Res > 7.5 (60 of the 121 dimers listed) vs. 50% for the entire set. This may be explained by the fact that larger proteins have more competing sites on their surfaces that could provide good docking correlation scores. By reporting only the top 50 docked dimers in our tests, the rank listings may not be deep enough to include the wild-type conformations; these larger proteins are therefore more likely to be "misses." For example, the 1aor dimer crystal structure shows a highly complementary interface, but because this protein is relatively large (605 residues), our docking algorithm does not pick up the wild type conformation in the top 50. Other docking algorithms severely penalize the competing sites ("false positives") by incorporating biochemical data or electrostatic terms in the scoring function, features we can't include given that our protein model does not include side-chains. It would be interesting to see how these other docking algorithms perform on our data set.

In summary, our docking algorithm performs reasonably well in docking the 121 homodimers. Most of the matches have less than 1 Å RMSD from wild-type, indicating that the atomic radii used for the backbones and the reduced side-chains on the C_β atoms are reasonable. There is about a 50% success rate in finding the wild-type conformations in the entire docking set, but the success rate is significantly higher for dimers that form relatively large interfaces with respect to their amino-acid chain length.

Conclusions

In this chapter, we describe the development of a docking algorithm specialized in generating dimer configurations for protein design purposes. To position the backbones correctly for protein design and to avoid bias toward the wild-type sequences, the wild-type side-chains are not considered in the docking process. The strategies employed include the use of 3D grids to represent the protein molecules in space, and spheres to approximate the side-chains. The spheres are defined by atomic radii, which are determined experimentally using known high-resolution dimer structures. Established FFT correlation methods are employed to efficiently cover all translational dimensions and search through all six degrees of freedom, and surface shape complementarities are used to score the fitness of the docked structures. Because C2 symmetry related homodimers tend to bury more surface area and use more hydrophobic amino acids in the interface, their interfaces are more protein core-like and can be modeled well by our protein design algorithms. We therefore parameterized our docking algorithm using C2 symmetry related homodimers as test cases. Imposing C2 symmetry also allowed us to make modifications that significantly improve computational efficiency. The resulting docking algorithm performed reasonably well in the 121 test cases used to validate the experimentally determined atomic radii. These results suggest that the reduced protein side-chain representation employed by our algorithm is a reasonable estimate, and the shapes defined by this representation can be used to position the protein backbones to a plausible dimer configuration for dimer design.

References

1. Connolly, M. L. Solvent-Accessible Surfaces of Proteins and Nucleic-Acids. *Science* **221**, 709-713 (1983).

2. Lin, S. L., Nussinov, R., Fischer, D. & Wolfson, H. J. Molecular-Surface Representations by Sparse Critical-Points. *Proteins-Structure Function and Genetics* **18**, 94-101 (1994).

3. Ritchie, D. W. & Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins-Structure Function and Genetics* **39**, 178-194 (2000).

4. Ritchie, D. W. Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins-Structure Function and Genetics* **52**, 98-106 (2003).

5. Chen, R. & Weng, Z. P. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins-Structure Function and Genetics* **47**, 281-294 (2002).

6. Katchalskikatzir, E. et al. Molecular-Surface Recognition - Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 2195-2199 (1992).

7. Press, W. H., Teukolsky, S. A., Vettering, W. T. & Flannery, B. P. in *Numerical Recipes in C* 496-546 (Cambridge University Press, New York, 1992).

8. Heckbert, P. in *Fourier Transform and the Fast Fourier Transform (FFT) Algorithm. Lecture notes given at Carnegie Mellon Univeristy*. (1995).

9. Bernstein, F. C. et al. Protein Data Bank - Computer-Based Archival File for Macromolecular Structures. *Journal of Molecular Biology* **112**, 535-542 (1977).

10. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. Dissecting subunit interfaces in homodimeric proteins. *Proteins-Structure Function and Genetics* **53**, 708-719 (2003).

11. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure* **29**, 105-153 (2000).

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

13. Fernandez-Recio, J., Totrov, M. & Abagyan, R. Identification of protein-protein interaction sites from docking energy landscapes. *Journal of Molecular Biology* **335**, 843-865 (2004).

14. Zacharias, M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science* **12**, 1271-1282 (2003).

15. Murphy, J., Gatchell, D. W., Prasad, J. C. & Vajda, S. Combination of scoring functions improves discrimination in protein-protein docking. *Proteins-Structure Function and Genetics* **53**, 840-854 (2003).

16. Mendez, R., Leplae, R., De Maria, L. & Wodak, S. J. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins-Structure Function and Genetics* **52**, 51-67 (2003).

17. Li, L., Chen, R. & Weng, Z. P. RDOCK: Refinement of rigid-body protein docking predictions. *Proteins-Structure Function and Genetics* **53**, 693-707 (2003).

18. Heifetz, A. & Eisenstein, M. Effect of local shape modifications of molecular surfaces on rigid-body protein-protein docking. *Protein Engineering* **16**, 179-185 (2003).

19. Fernandez-Recio, J., Totrov, M. & Abagyan, R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins-Structure Function and Genetics* **52**, 113-117 (2003).

20. Chen, R. & Weng, Z. P. A novel shape complementarity scoring function for protein- protein docking. *Proteins-Structure Function and Genetics* **51**, 397-408 (2003).

21. Chen, R., Li, L. & Weng, Z. P. ZDOCK: An initial-stage protein-docking algorithm. *Proteins-Structure Function and Genetics* **52**, 80-87 (2003).

22. Lorber, D. M., Udo, M. K. & Shoichet, B. K. Protein-protein docking with multiple residue conformations and residue substitutions. *Protein Science* **11**, 1393-1408 (2002).

23. Heifetz, A., Katchalski-Katzir, E. & Eisenstein, M. Electrostatics in protein-protein docking. *Protein Science* **11**, 571-587 (2002).

24. Camacho, C. J., Gatchell, D. W., Kimura, S. R. & Vajda, S. Scoring docked conformations generated by rigid-body protein- protein docking. *Proteins-Structure Function and Genetics* **40**, 525-537 (2000).

25. Ewing, T. J. A. & Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *Journal of Computational Chemistry* **18**, 1175-1189 (1997).

26. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

Table 2-1. Example of Bit Reversal[†]

Index k								
In numerical order	0	1	2	3	4	5	6	7
In binary	000	001	010	011	100	101	110	111
In binary bit-reversed[a]	000	100	010	110	001	101	011	111
In bit-reversed numerical order	0	4	2	6	1	5	3	7

[a] The binary bits are reversed with respect to the row above it.

[†] Starting from the numerical value of an index k, the bit-reversal process entails the reordering of index values in their binary format.

Table 2-2. Atomic Radii Determined from Iteration Cycles

	ranges of good atomic radii[a]		
atom type	1a7w	1c9o	final values used[b]
N	1.4~1.6	1.4~1.55	1.4
O	1.3~1.45	1.3~1.35	1.3
C	1.7~1.8	1.7~2.55	1.75
CA	2.15~2.4	2.3~2.4	2.35
CB	2.15	2.2~2.25	2.15[c]

[a] The range is determined from the radii combinations that gave the best 200 docked scores.

[b] Lower bound from the ranges of good atomic radii is used

[c] The C_β atomic radii between 1.9 Å to 2.6 Å should be tested. The suggested value of 2.15 Å is only good for certain cases.

Table 2-3. Docking Results

			Best match to wild-type[b]		First rank with rmsd < 3 Å[b]		Interface[c]	
PDB	Residues	Long axis (Å)[a]	Rank*	RMSD*	Rank*	RMSD*	B/2 (Å)[d]	Area/Res.[e]
12as	327	72.3	—	—	—	—	1989	6.1
1a3c	166	55.6	—	—	—	—	853	5.1
1a4i	285	64.9	—	—	—	—	1353	4.7
1a4u	254	63.1	7	0.291	1	0.656	2547	10.0
1aa7	158	47.7	—	—	—	—	1125	7.1
1ad3	446	124.8	4	0.314	1	0.533	3936	8.8
1ade	431	81.4	17	2.452	17	2.452	2708	6.3
1af5	126	60.5	—	—	—	—	856	6.8
1afw	390	72.7	31	0.333	1	0.558	2400	6.2
1ajs	412	100.9	1	1.124	1	1.124	3401	8.3
1amk	250	59.0	50	0.072	1	0.532	1477	5.9
1aor	605	79.3	—	—	—	—	1180	2.0
1aq6	245	60.9	33	0.716	28	1.097	2232	9.1
1auo	218	53.8	—	—	—	—	662	3.0
1b3a	67	104.0	—	—	—	—	763	11.4
1b5e	241	67.3	2	0.156	1	0.156	2581	10.7
1b67	68	60.5	44	0.537	44	0.537	1607	23.6
1b8a	438	107.4	—	—	—	—	4391	10.0
1b8j	448	81.7	22	0.309	1	0.348	3794	8.5
1bam	200	61.7	—	—	—	—	745	3.7
1bbh	131	54.3	—	—	—	—	771	5.9
1bd0	381	95.0	8	0.699	8	0.699	3091	8.1
1bif	432	85.9	—	—	—	—	858	2.0
1biq	339	77.3	15	0.408	1	1.324	3004	8.9
1bis	146	54.9	16	2.563	16	2.563	1495	10.2
1bjw	381	84.5	25	0.501	2	0.795	2938	7.7
1bkp	278	65.8	—	—	—	—	2206	7.9
1bmd	326	65.8	14	0.311	13	0.312	1564	4.8
1brw	433	82.0	—	—	—	—	1083	2.5
1bsl	323	68.7	5	1.38	5	1.38	1918	5.9
1bsr	124	67.8	13	0.975	3	0.997	1888	15.2
1buo	121	86.3	13	0.259	2	0.536	1972	16.3
1bxg	349	67.3	—	—	—	—	1041	3.0
1bxk	341	79.2	—	—	—	—	1286	3.8
1cdc	96	69.5	2	0.621	1	0.638	3918	40.8
1cg2	389	113.4	—	—	—	—	1298	3.3
1chm	401	88.3	23	0.525	1	0.655	3171	7.9
1cmb	104	53.1	6	1.494	1	1.9	1797	17.3
1cnz	363	89.5	11	1.565	11	1.565	2447	6.7
1coz	126	51.5	—	—	—	—	1050	8.3
1csh	435	88.9	34	0.033	1	0.343	5057	11.6
1ctt	294	65.8	45	0.385	1	0.632	1990	6.8
1cvu	551	146.0	14	1.2	3	1.282	2436	4.4
1czj	110	54.4	—	—	—	—	829	7.5
1daa	277	70.2	15	0.448	15	0.448	2193	7.9
1dor	311	75.9	30	0.368	16	1.326	2189	7.0
1dpg	485	102.8	—	—	—	—	2293	4.7
1dqs	381	77.3	4	0.844	3	1.843	1640	4.3
1dxg	36	31.7	21	0.326	21	0.326	729	20.3
1e98	210	55.2	—	—	—	—	770	3.7
1ebh	436	78.8	43	0.799	23	2.481	1784	4.1
1f13	722	126.8	—	—	—	—	2556	3.5
1fip	73	55.2	23	0.55	4	0.658	1836	25.2

Table 2-3. (Continued)

			Best match to wild-type[b]		First rank with rmsd < 3 Å[b]		Interface[c]	
PDB	Residues	Long axis (Å)[a]	Rank*	RMSD*	Rank*	RMSD*	B/2 (Å)[d]	Area/Res.[e]
1fro	176	68.1	3	0.151	1	0.203	3505	19.9
1gvp	87	54.4	5	1.5	5	1.5	908	10.4
1hhp	99	52.8	2	0.056	2	0.056	1599	16.2
1hjr	158	63.7	—	—	—	—	962	6.1
1hss	111	45.1	44	1.4	11	1.624	1101	9.9
1hxp	340	78.1	22	0.259	1	0.5	3402	10.0
1icw	69	43.4	47	0.898	47	0.898	954	13.8
1imb	273	60.4	36	1.706	9	2.573	1623	5.9
1isa	192	60.1	—	—	—	—	920	4.8
1ivy	452	77.8	—	—	—	—	1601	3.5
1jhg	101	59.1	—	—	—	—	2207	21.9
1jsg	111	60.2	—	—	—	—	794	7.2
1kba	66	47.7	28	2.286	28	2.286	498	7.5
1kpf	126	45.8	5	0.31	3	0.358	1867	14.8
1lyn	125	61.9	—	—	—	—	948	7.6
1m6p	146	53.3	37	1.143	9	1.282	1025	7.0
1mkb	171	58.0	2	0.46	1	0.799	1605	9.4
1mor	366	70.0	50	0.513	21	1.401	2540	6.9
1nox	200	72.3	10	0.397	1	0.741	3033	15.2
1nse	416	84.0	—	—	—	—	2736	6.6
1nsy	271	68.9	19	0.486	19	0.486	2592	9.6
1oac	719	109.3	—	—	—	—	7149	9.9
1opy	123	52.5	28	0.249	23	0.58	1048	8.5
1pgt	209	58.3	15	2.808	15	2.808	1238	5.9
1pre	449	131.0	29	0.844	17	1.272	2300	5.1
1qfh	212	98.5	—	—	—	—	2264	10.7
1qhi	304	68.5	23	0.713	8	2.149	1714	5.6
1qr2	230	75.8	—	—	—	—	1947	8.5
1r2f	283	73.4	48	2.601	36	2.635	1746	6.2
1reg	122	105.3	—	—	—	—	659	5.4
1rfb	119	63.7	9	1.28	1	1.752	2650	22.3
1rpo	61	53.4	44	0.704	1	1.084	1405	23.0
1ses	421	136.3	—	—	—	—	2211	5.3
1slt	133	42.7	—	—	—	—	536	4.0
1smn	241	54.8	26	0.999	1	1.552	866	3.6
1smt	98	77.2	34	0.267	1	0.3	1970	20.1
1sox	463	81.3	—	—	—	—	1404	3.0
1tc1	175	57.7	—	—	—	—	1540	8.8
1tox	515	94.7	1	0.026	1	0.026	3721	7.2
1trk	678	105.6	1	0.371	1	0.371	4476	6.6
1uby	348	78.1	12	2.564	6	2.576	2168	6.2
1utg	70	46.6	44	1.263	9	2.391	1485	21.2
1vfr	217	72.6	1	0.368	1	0.368	3431	15.8
1vok	192	73.4	—	—	—	—	1577	8.2
1wtl	108	58.6	—	—	—	—	698	6.5
1xso	149	47.0	—	—	—	—	662	4.4
2arc	161	56.0	—	—	—	—	765	4.8
2ccy	127	53.9	—	—	—	—	792	6.2
2hdh	286	74.4	—	—	—	—	1524	5.3
2ilk	155	78.0	8	0.314	1	0.745	4542	29.3
2lig	157	89.1	—	—	—	—	1686	10.7
2mcg	215	77.7	—	—	—	—	1646	7.7
2nac	374	75.5	—	—	—	—	3789	10.1

Table 2-3. (Continued)

			Best match to wild-type[b]		First rank with rmsd < 3 Å[b]		Interface[c]	
PDB	Residues	Long axis (Å)[a]	Rank*	RMSD*	Rank*	RMSD*	B/2 (Å)[d]	Area/Res.[e]
2ohx	374	77.8	—	—	—	—	1718	4.6
2spc	106	118.3	—	—	—	—	2508	23.7
2sqc	623	84.8	—	—	—	—	809	1.3
2tct	198	77.1	45	1.198	45	1.198	2675	13.5
2tgi	112	75.6	—	—	—	—	1262	11.3
3dap	320	79.9	—	—	—	—	2661	8.3
3grs	461	79.1	—	—	—	—	3302	7.2
3sdh	145	49.7	—	—	—	—	873	6.0
3ssi	108	56.6	36	0.824	22	2.878	866	8.0
4cha	239	55.5	—	—	—	—	1026	4.3
4kbp	424	79.4	24	2.793	24	2.793	1478	3.5
5csm	250	69.6	25	2.13	9	2.962	2007	8.0
5rub	436	86.0	—	—	—	—	2859	6.6
8prk	282	56.8	—	—	—	—	969	3.4
9wga	170	62.3	2	0.132	1	0.139	2293	13.5

[a] The long axis of a molecule is determined by two times the distance from its geometric center to the furthest atom.

[b] Only the molecules from the highest 50 correlation scores are considered.

[c] Per subunit

[d] Data in this column are taken from Bahadur et al., PROTEINS: Structure, Function, and Genetics 53:708-719 (2003)

[e] The interface area contributed by each subunit divided by the number of residues per subunit.

* — means no match within 3 Å rmsd from the top 50 ranked molecules

Table 2-4. The Area/Residues Ratios and Docking Hits[a]

PDB	Area/Res[b]	Hit[c]	PDB	Area/Res[b]	Hit[c]	PDB	Area/Res[b]	Hit[c]
2sqc	1.3	.	1hjr	6.1	.	1aq6	9.1	+
1aor	2.0	.	1afw	6.2	+	1mkb	9.4	+
1bif	2.0	.	1r2f	6.2	+	1nsy	9.6	+
1brw	2.5	.	1uby	6.2	+	1hss	9.9	+
1bxg	3.0	.	2ccy	6.2	.	1oac	9.9	.
1sox	3.0	.	1ade	6.3	+	1hxp	10.0	+
1auo	3.0	.	1wtl	6.5	.	1b8a	10.0	.
1cg2	3.3	.	5rub	6.6	.	1a4u	10.0	+
8prk	3.4	.	1nse	6.6	.	2nac	10.1	.
4kbp	3.5	+	1trk	6.6	+	1bis	10.2	+
1f13	3.5	.	1cnz	6.7	+	1gvp	10.4	+
1ivy	3.5	.	1ctt	6.8	+	1qfh	10.7	.
1smn	3.6	+	1af5	6.8	.	1b5e	10.7	+
1e98	3.7	.	1mor	6.9	+	2lig	10.7	.
1bam	3.7	.	1m6p	7.0	+	2tgi	11.3	.
1bxk	3.8	.	1dor	7.0	+	1b3a	11.4	.
1slt	4.0	.	1aa7	7.1	.	1csh	11.6	+
1ebh	4.1	+	1jsg	7.2	.	9wga	13.5	+
4cha	4.3	.	3grs	7.2	.	2tct	13.5	+
1dqs	4.3	+	1tox	7.2	+	1icw	13.8	+
1cvu	4.4	+	1czj	7.5	.	1kpf	14.8	+
1xso	4.4	.	1kba	7.5	+	1nox	15.2	+
2ohx	4.6	.	1lyn	7.6	.	1bsr	15.2	+
1dpg	4.7	.	2mcg	7.7	.	1vfr	15.8	+
1a4i	4.7	.	1bjw	7.7	+	1hhp	16.2	+
2arc	4.8	.	1chm	7.9	+	1buo	16.3	+
1isa	4.8	.	1daa	7.9	+	1cmb	17.3	+
1bmd	4.8	+	1bkp	7.9	.	1fro	19.9	+
1pre	5.1	+	3ssi	8.0	+	1smt	20.1	+
1a3c	5.1	.	5csm	8.0	+	1dxg	20.3	+
1ses	5.3	.	1bd0	8.1	+	1utg	21.2	+
2hdh	5.3	.	1vok	8.2	.	1jhg	21.9	.
1reg	5.4	.	1ajs	8.3	+	1rfb	22.3	+
1qhi	5.6	+	3dap	8.3	.	1rpo	23.0	+
1bbh	5.9	.	1coz	8.3	.	1b67	23.6	+
1amk	5.9	+	1qr2	8.5	.	2spc	23.7	.
1pgt	5.9	+	1b8j	8.5	+	1fip	25.2	+
1bsl	5.9	+	1opy	8.5	+	2ilk	29.3	+
1imb	5.9	+	1tc1	8.8	.	1cdc	40.8	+
3sdh	6.0	.	1ad3	8.8	+			
12as	6.1	.	1biq	8.9	+			

[a] Sorted by the Area/Res ratio of each PDB entry in ascending order.

[b] The interface area contributed by each subunit divided by the number of residues per subunit.

[c] Only the top 50 correlation score ranked dimers are considered. "+" means there is at least one docked dimer with an RMSD less than 3 Å to the wildtype. "." means there is no match.

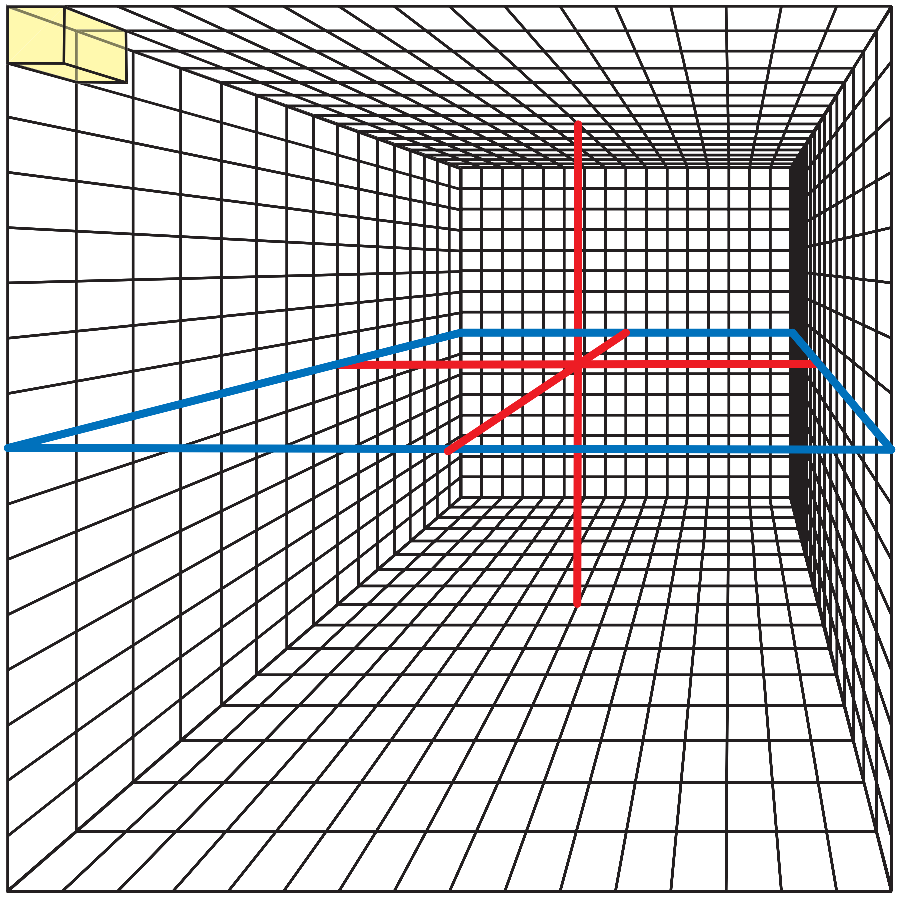


Figure 2-1. 3D grid space represented by memory array. For discretizing protein molecules, a 1D array was created to represent the entire 3D cubic space shown in this figure. Array indices start from the corner, but the actual origin of the coordinate system is in the center. Each array element represents a "box" in space, as represented by the yellow box in the upper left corner. The cubic shape is distorted for visualization clarity.

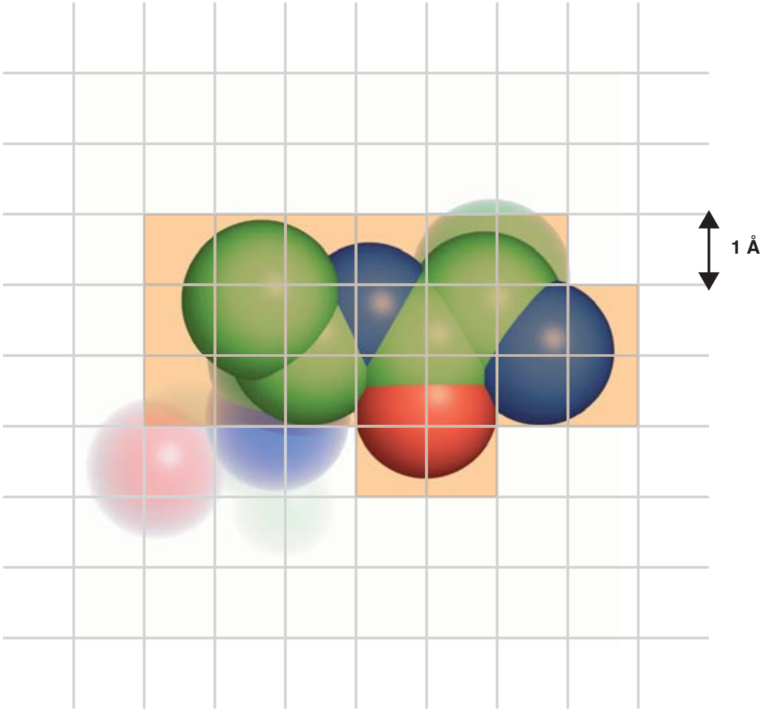


Figure 2-2. Discretizing proteins into 3D grids. The van der Waals spheres of a short protein backbone are shown to illustrate the discretization process. The distances between each atom and the grid boundaries are evaluated. A grid voxel is assigned to an atom if any of its sides or edges falls within the atomic radii of the particular atom. The 2D view of a single 1 Å grid layer is shown. Highlighted in orange are the grids assigned as part of the protein molecule. Notice some of the atoms have no grid voxels assigned to them, as they are off this plane and will be covered by a different grid layer. Resolution will improve if a smaller grid size is used.

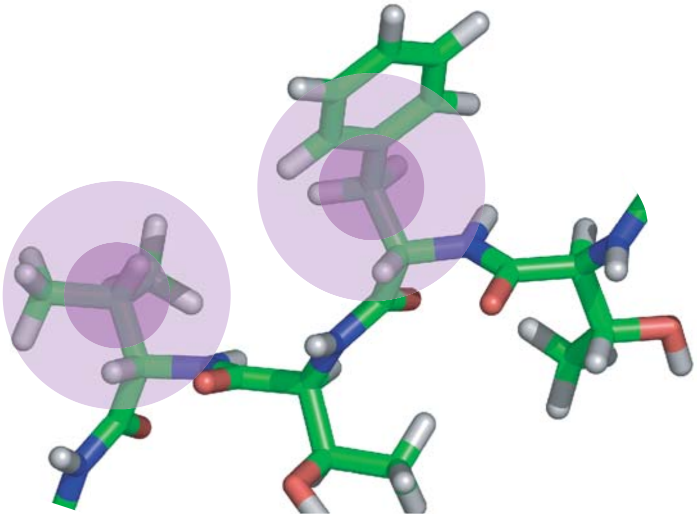


Figure 2-3. Estimating side-chain volumes by spheres. The full atomic details of side-chains beyond C_β are not considered. Instead, a sphere centered at C_β is used to approximate the volumes taken up by the original side-chains.

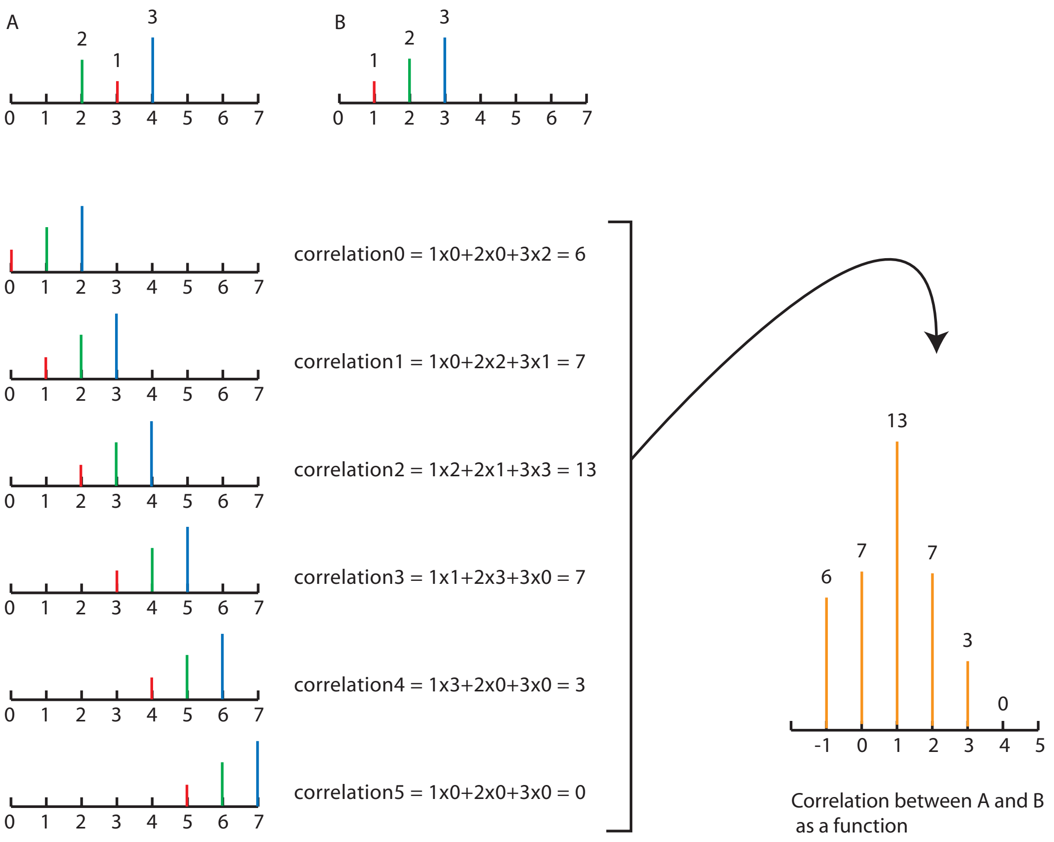

A
2
1
3
0 1 2 3 4 5 6 7
B
1
2
3
0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7
correlation0 = 1x0+2x0+3x2 = 6
0 1 2 3 4 5 6 7
correlation1 = 1x0+2x2+3x1 = 7
0 1 2 3 4 5 6 7
correlation2 = 1x2+2x1+3x3 = 13
0 1 2 3 4 5 6 7
correlation3 = 1x1+2x3+3x0 = 7
0 1 2 3 4 5 6 7
correlation4 = 1x3+2x0+3x0 = 3
0 1 2 3 4 5 6 7
correlation5 = 1x0+2x0+3x0 = 0
13
6
7
7
3
0
-1 0 1 2 3 4 5
Correlation between A and B
as a function

Figure 2-4. Correlation between two functions. The correlation between two functions A and B can be visualized as a correlation map. By moving one function (in this case, B) against the other (A) and evaluating the correlation between the two for each step, a correlation map that contains information about both the correlations and the steps can be obtained.

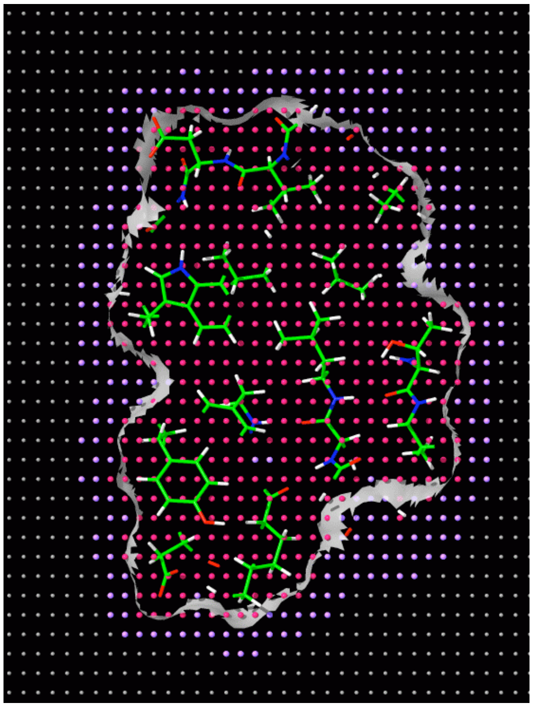


Figure 2-5. Scoring scheme for discretized molecules. Gray spots are assigned the value of "0". The favorable surface layer is shown in purple covering 1.5 Å outside of the core layer, and has a value of "1." The core layer is shown in red, and has a value of "-15" to penalize penetration into this layer. (Figure created by John J. Love)

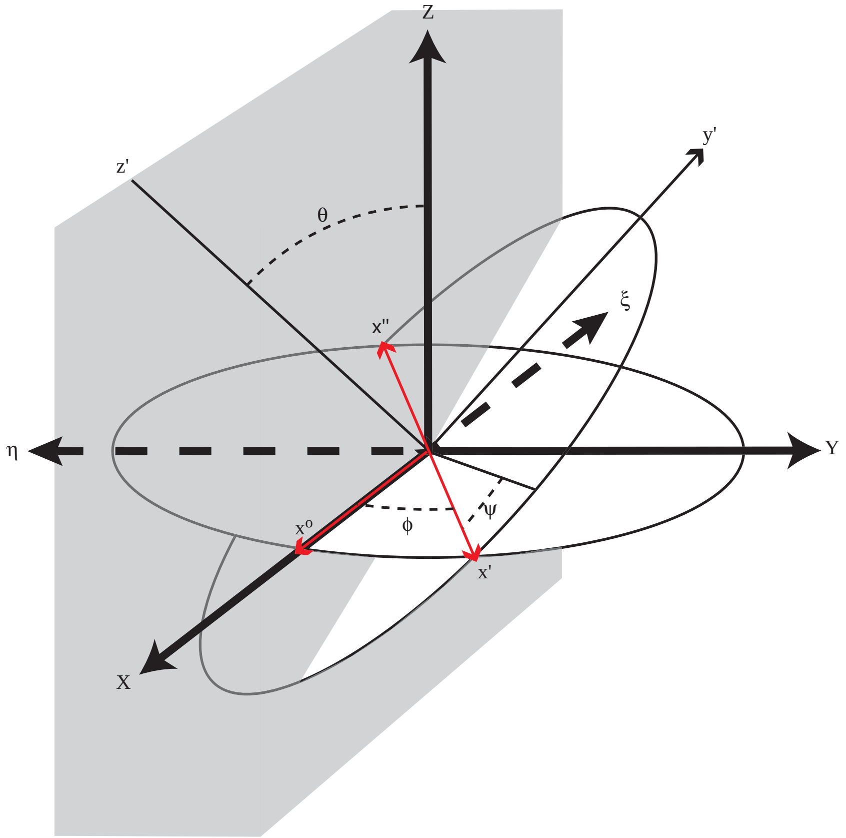


Figure 2-6. Rotational search space reduction with C2 symmetry related dimers. Two symmetry related coordinate systems are shown: XYZ and ξηZ. By the definition of C2 symmetry, any point in the XYZ coordinate system mirrors to a point in the ξηZ system by a 180° rotation. The rotation of a rigid body with respect to a regular XYZ coordinate system can be described using three Euler angles, ϕ, θ, ψ, over the ranges of $0 \leq \phi \leq 2\pi$, $0 \leq \theta \leq \pi$, $0 \leq \psi \leq 2\pi$. Due to the 180° rotational symmetry with respect to the Z axis, the ranges of ϕ and ψ in this XYZ, ξηZ combined coordinate system are both reduced by half to $0 \leq \phi \leq \pi$, $0 \leq \psi \leq \pi$, while the range of θ remains the same. Rotations beyond the range of 0 to π are redundant since the resulting positions can always be folded back to positions within the range of 0 to π, as illustrated by vectors x' and x''.

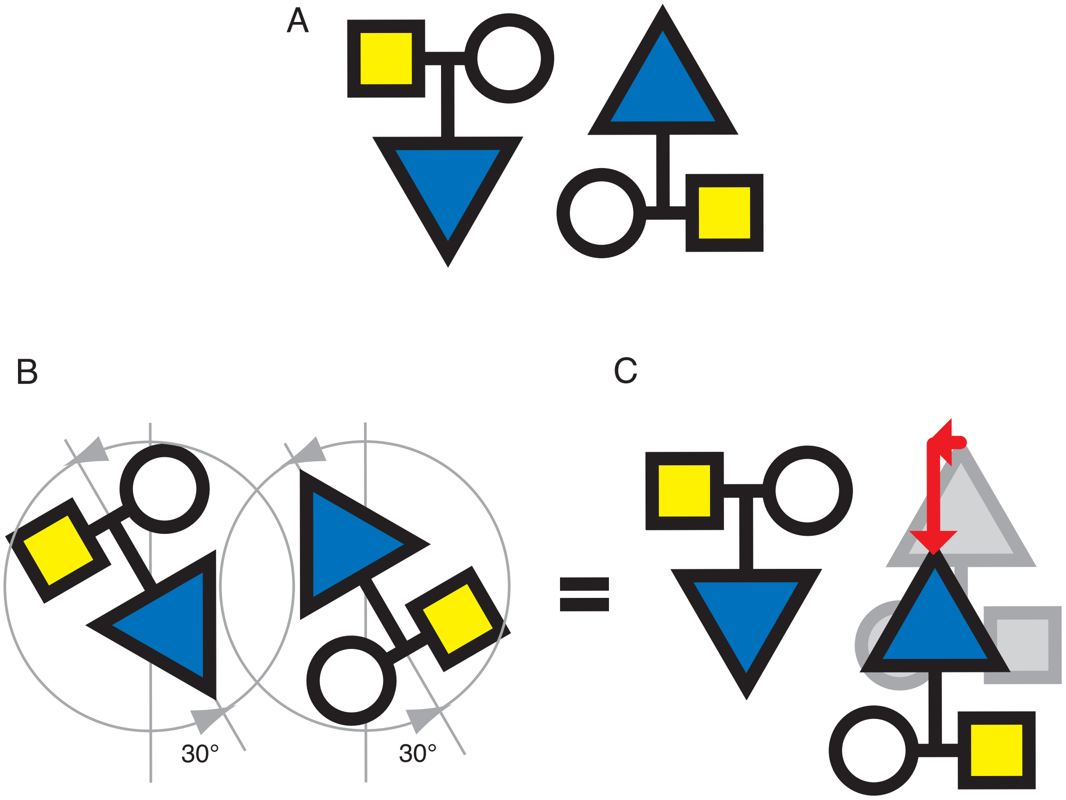
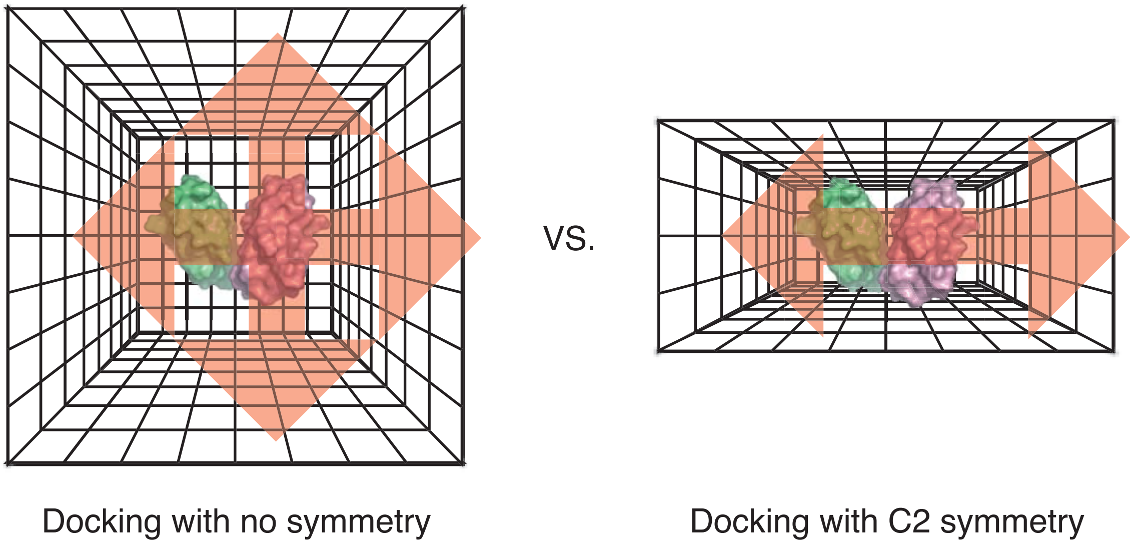


Figure 2-7. Translational equivalence. For any dimeric conformation obtained by rotational manipulations around the symmetry axis, the same conformation can be obtained by translational moves along the axes perpendicular to the symmetry axis. **A.** A pair of C2 symmetry related dimers in its initial position. **B.** Each subunit of the dimer is rotated by 30° to a new C2 symmetry conformation. **C.** The same conformation as in B can be obtained by two translational moves from the initial position (A).



Docking with no symmetry

Docking with C2 symmetry

Figure 2-8. Memory space reduction. Computer memory can be reduced by half when docking C2 symmetry related dimers. When docking dimers with no symmetry restriction, the arena must be large enough to allow complete searches on all sides of the proteins; when C2 symmetry is imposed, translational searches along the dimension parallel to the symmetry axis are not required to maintain the C2 symmetry. This dimension can therefore be reduced to the length of the long axis of the molecule rather than three times the length. Since our FFT implementation requires grid points along each dimension to be a power of 2, we simply reduce the dimension parallel to the symmetry axis by a factor of 2. The number of grid points (and the memory) required for the search therefore becomes half.

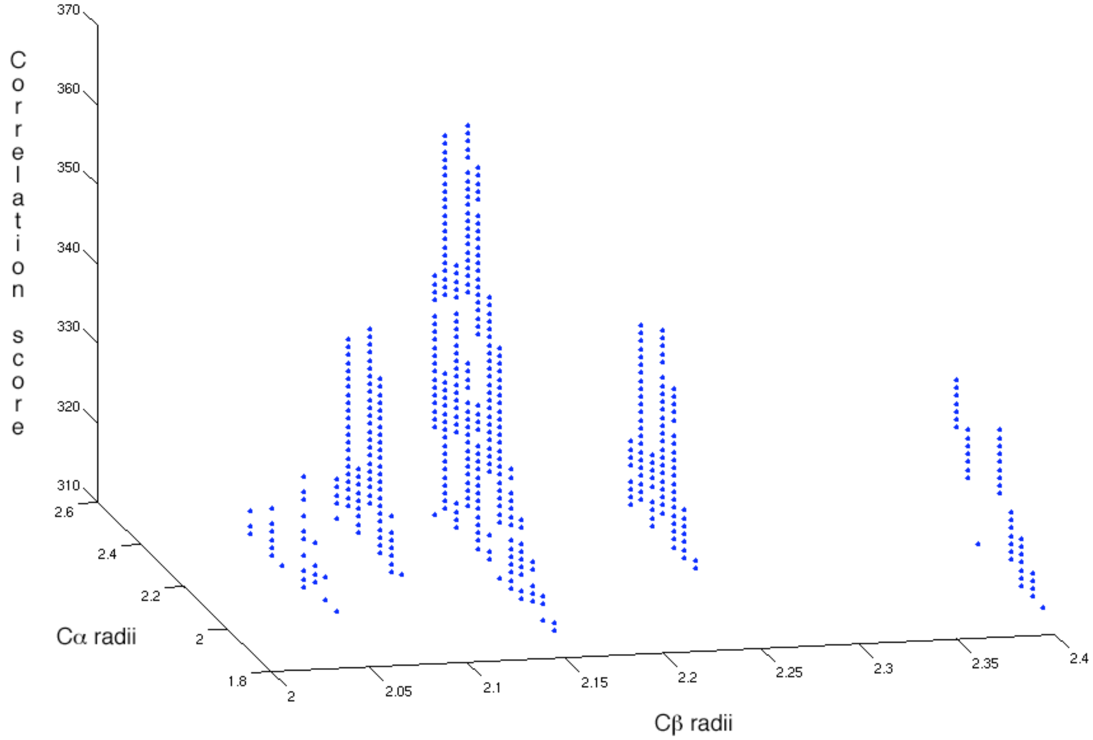


Figure 2-9. Atomic radii determination. Correlation scores from radii iteration cycles are plotted against C_α and C_β radii as an example of how the good radii ranges are determined. The correlation scores are sensitive to changes in C_α and C_β radii, as evidenced by the narrow distribution around a single atomic radius.

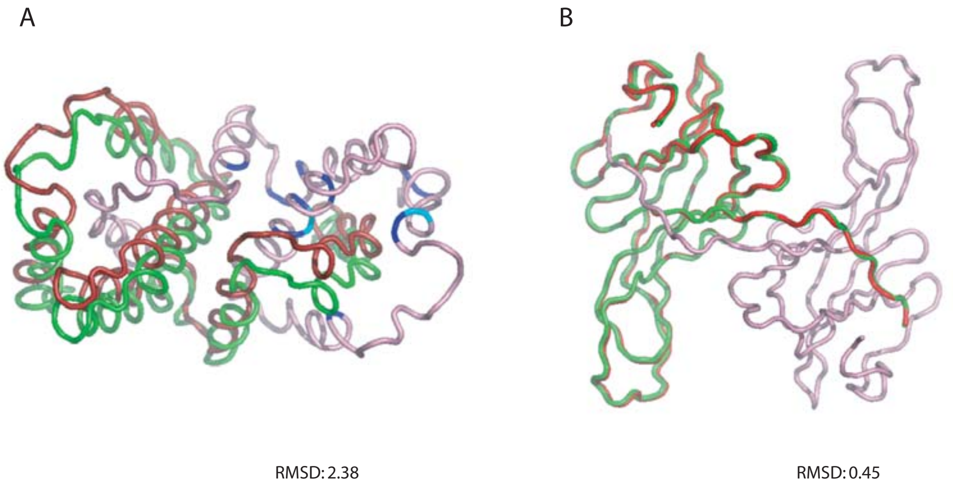


Figure 2-10. Comparison between wild-type and computationally docked conformations. Molecule colored in pink is held stationary during docking; the best docked conformation (in green) is compared with the wild-type conformation (in red). **A.** The two monomers of a 1rfb homodimer are docked to each other. The RMSD between the wild-type and the docked product is 2.38 Å. Positions that are phenylalanines in the wild-type structure are colored in blue; tyrosines are in cyan. **B.** The two monomers of 1ecz are docked to each other; the docked product has an RMSD of 0.45 Å compared to the wild-type.

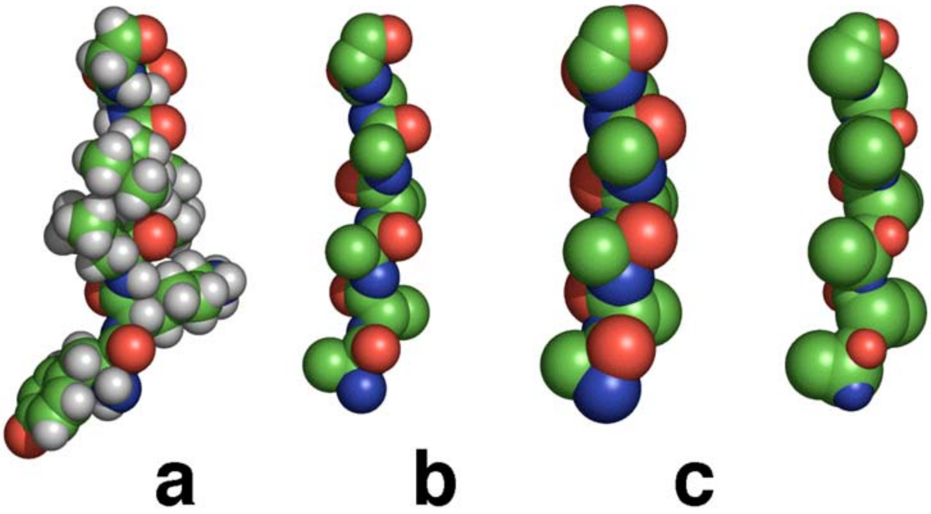


Figure 2-11. Accuracy of three different sphere approximations of the same peptide chain. Carbons are in green, nitrogens in blue, oxygens in red, and hydrogens in white. **a.** Peptide chain with full atom van der Waals spheres (reference). **b.** Poly-alanine model using standard van der Waals radii; peptide volume is underestimated. **c.** Poly-alanine model using a uniform radius of 2.15 Å for all the atoms. **d.** Poly-alanine model using experimentally determined radii for each of the atom types. Hydrogens are not included in models b, c, and d. Modeling a peptide with spheres centered on the C_β atoms gives fairly good approximations for short side-chains, but falls short in estimating aromatic and other long side-chains.

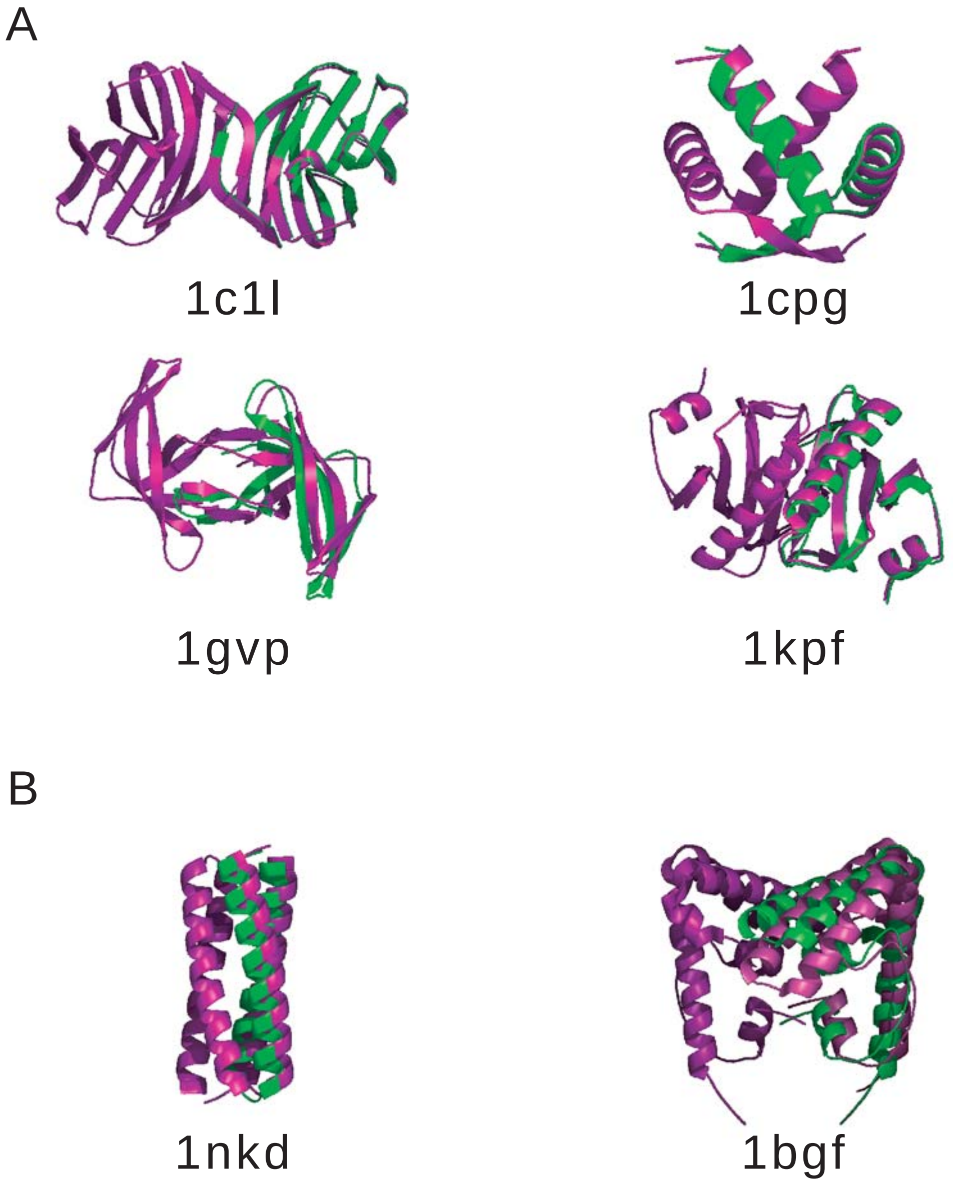


Figure 2-12. The types of amino acids in the interface influence the docking results. Wild-type conformations are shown in magenta and docked conformations are shown in green. For each of the dimers, one of the subunits of the docked structure was superimposed on the corresponding wild-type subunit (shown here on the left); the overlapping or offset portions of the docked structure relative to the wild-type can be seen by the amount of green displayed. **A.** Proteins that use backbone hydrogen bonds in the interface to form cross-dimer β-sheets. **B.** Helical proteins that primarily use side chain-side chain interactions to form the interface. Even though the overall dimer configurations are correctly identified in the examples in B, the dimer conformations are more closely reproduced for the molecules in A.

Chapter 3

A *De Novo* Designed Protein/Protein (Heterodimer) Interface*

*Adapted from manuscript prepared by Possu Huang, John J. Love and Stephen L. Mayo.

(to be submitted)

ABSTRACT

The goal of this work was to redesign a small monomeric protein (Protein G) such that it self-assembled into a dimer complex of specific structure. The first step in driving *de novo* self-assembly was the computational docking of the proteins together in a predefined orientation. To achieve this goal we modified an established docking algorithm, the Geometric Recognition Algorithm (GRA). The GRA treats the molecules as rigid bodies and rigorously assesses interfacial surface complementarity as a function of translational and rotational position. This process is computationally intensive yet was rendered tractable by utilizing the Fourier Correlation Theorem. Upon obtaining the optimal intermolecular atomic coordinates the two molecules were treated as one and a suite of highly developed protein design algorithms, which utilize advanced molecular mechanics force fields, was used to computationally repack the interfacial side-chains in a manner analogous to the cores of well folded proteins. Total gene synthesis was used to introduce the mutations and physically produce the redesigned proteins. The extent and specificity of binding were confirmed with analytical ultracentrifugation and heteronuclear NMR.

INTRODUCTION

Molecular self-assembly is the spontaneous association of molecules into stable, structurally well-defined complexes joined by noncovalent bonds. All major cellular processes depend on the precise, highly specific self-assembly of proteins into functional complexes. In addition, extracellular protein/protein interactions (*e.g.,* growth factor/receptor and antibody/antigen interactions) are also highly specific and governed by the same physical parameters that drive protein folding. Understanding and controlling these parameters is a major goal of protein biochemistry. To date, much progress has been made in this area by analyzing the large body of data collected on natural protein interfaces[1-9] and references therein. The field of protein design is uniquely positioned to complement these efforts with an inverse approach, *i.e.,* instead of predicting how native complexes form we can explore the essential binding parameters by driving the *de novo* self-assembly of previously monomeric proteins. Moreover, the ability to direct a designed protein to bind a target protein in a site-specific manner has potential therapeutic as well as technological applications.

Advances in protein/protein interface design have thus far come primarily from directed evolution methods in which diverse libraries of proteins are physically generated and screened for candidates that either bind a target protein *in vivo* or an immobilized target *in vitro*[10]. These methods include, for example, two-hybrid screens, yeast surface display, phage display, mRNA display and ribosome display[3,11-15]. Though highly effective, these methods have inherent limitations due to the practical considerations associated with the physical generation of protein libraries as well as limiting *in vivo* factors (*e.g.,* transformation efficiency). These limitations prevent library diversity from exceeding a complexity of ~10^{14}. More importantly, directed evolution methods do not result in the designed protein forming a site-specific complex with the target protein (*i.e.,* the binding site on the target cannot be decided beforehand and is therefore indiscriminate). Conversely, computational protein design methods can screen virtual libraries of much larger complexity (*e.g.,* ~10^{80})[16] and the binding of the designed protein to the target protein results in a complex of specific structure.

Computational protein design methods have recently made significant progress towards engineering novel protein interfaces. Chevalier *et al.* generated an artificial endonuclease by fusing two domains from different endonucleases followed by reengineering of the newly formed interface[17]. They combined computational redesign and an *in vivo* protein-folding screen to create an endonuclease with novel sequence specificity and, upon solving the crystal structure of the engineered endonuclease, proved the accuracy of the protein interface redesign algorithm. To explore binding specificity at protein/protein interfaces, the simplest interface, the heptad repeat of well-studied coiled-coil of GCN4[18], was redesigned using both positive and negative design features[19]. Homodimers of coiled-coils were successfully generated and thus indicate that, in the context of heptad repeats, a multistate framework is necessary for engineering specificity and selecting against undesired competitor conformations. Finally, an extensive redesign of the interface between the protein calmodulin and a bound peptide (derived from smooth muscle myosin light chain kinase) revealed the importance of emphasizing intermolecular versus intramolecular interactions during the computational design and optimization of interfacial residues[20,21]. Additionally, it was discovered that the best redesigned calmodulin variant showed an increase in binding specificity relative to wild-type calmodulin and that the increase was not due to increased binding to the peptide but instead due to decreased binding to alternative targets.

Here we report the *de novo* design of a heterodimer interface that was generated by docking the β1 domain of the streptococcal Protein G to itself in a structurally specific fashion followed by mutation of specific interfacial side-chains so as to drive complex formation. The 56 amino acid Protein G domain (Figure 3-1) was chosen because it expresses well in *E. coli*, it is monomeric and well behaved in solution, it has been extensively redesigned and biophysically analyzed[22] and its small compact structure has been determined to high resolution[23,24]. In addition, the functional use of full-length streptococcal Protein G to immunoprecipitate IgG antibodies provides natural evidence that the β1 domain has the ability to form multiple intermolecular contacts. Crystal structures of single β1 domains in complex with the constant F_c region of two different IgG antibodies (*i.e.,* 1FCC and 1IGC) revealed that the β1 domain does indeed bind in a distinct manner to different proteins[25-27] and is thus a reasonable candidate for *de novo* docking.

The overall *de novo* docking process naturally divides into two steps. The first step entails the computational docking of the backbone coordinates of the two proteins together in a general starting orientation. After choosing the general starting orientation local rotational and translational space is rigorously searched to determine the specific intermolecular orientation that corresponds to the docked complex of highest surface complementarity (*i.e.,* within the context of the general starting orientation). To obtain the positional backbone coordinates that correspond to the docked complex of highest surface complementarity we modified an established geometric recognition algorithm[28,29] and used it to rigorously assess surface complementarily of two Protein G monomers as a function of intermolecular position. Even though we had dictated the general starting orientation, and thus did not need to search all rotational and translational space, a high-resolution calculation of the correlation function was still desired and consequently remained computationally intensive. We therefore implemented the Fourier correlation theorem (FCT) originally described in[28]. The FCT reduces the number of translational calculations (*i.e.*, ~N^6) down to an order of N^3 log_2 (N^3)[30,31].

In the second step of the design process the two docked proteins were treated as one and a suite of highly advanced protein design algorithms were used to computationally mutate and repack the side-chains at the protein-protein interface in a manner similar to that observed in the cores of well-folded proteins. The algorithms used have previously been applied to protein design and stabilization and are contained in the ORBIT (optimization of rotomers by iterative techniques) suite of algorithms[16,22,32-34]. Since symmetry restraints were not imposed during the first step of the docking process, the second step resulted in a pair of protein monomers that had different sets of mutations and therefore resulted in the formation of a heterodimer complex (the protein monomers are referred to herein as monomer-A and monomer-B). The designed mutant monomers were each physically generated and purified and complex formation was confirmed by analytical ultracentrifugation and heteronuclear NMR.

RESULTS

Computational Docking

A modified version of a geometric recognition algorithm (GRA) was used to computationally dock Protein G to itself in a structurally specific manner. The GRA, in its original form, has been traditionally used to dock proteins together in an attempt to predict the quaternary structures of dimeric complexes[28,29]. Surface complementarity is assessed with the GRA by projecting the molecules into a 3-dimensional grid of *N x N x N* points where they are represented by the following discrete functions:

$$\text{molecule } a_{l,m,n} = \begin{cases} 1 \text{ surface of molecule} \\ \rho \text{ core of molecule} \\ 0 \text{ outside the molecule} \end{cases}$$

$$\text{molecule } b_{l,m,n} = \begin{cases} 1 \text{ inside of molecule} \\ 0 \text{ outside the molecule} \end{cases}$$

Matching of complementary surfaces is then accomplished by computing the following correlation function:

$$\text{Correlation Function:} \quad c_{\alpha,\beta,\gamma} = \sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{l=1}^{N} a_{l,m,n} \bullet b_{l+\alpha,m+\beta,n+\gamma} \qquad \text{Eq. 3-1}$$

Protein Discretization: To illustrate the modifications to the GRA necessary for *de novo* docking a general description of the original discretization process is provided. The means by which the two molecules are discretized is as follows - the atomic coordinates for molecule $a_{l,m,n}$ are centered in a 3-dimensional grid and held stationary. Each grid point is then assigned a value based on its proximity to molecule $a_{l,m,n}$ atoms. If a grid point falls within 1.8 Å of a molecule $a_{l,m,n}$ atom it is assigned the value ρ (*i.e.,* -15) and if a grid point falls between 1.8 Å and 3.3 Å of an atom it is given a value of +1. This process effectively digitizes the structure of the protein into a 3-dimensional grid where grid points located within the core of the protein (*i.e.,* ρ) are assigned a value of -15 and those on the surface a value of +1, thus creating a surface layer around the stationary molecule with a thickness of 1.5 Å. For the second protein, molecule $b_{l,m,n}$, if a grid point falls within 1.8 Å of an atom it was given a value of +1 (there is no surface layer for molecule $b_{l,m,n}$). All other grid points that are not in proximity to the corresponding protein atoms are assigned a value of zero.

At the end of the discretization process each molecule is individually represented by a 3-dimensional array of digitized values. The coordinates for molecule $b_{l,m,n}$ are translated through the grid and at each shift vector position they are discretized and the correlation value assessed. The correlation value is calculated by obtaining the product of the corresponding array elements for each molecule and summing the products over the entire grid (equation 1). If the translational shift vector is such that molecule $b_{l,m,n}$ is not in proximity to molecule $a_{l,m,n}$ then all non-zero values for each molecule are multiplied against zero values from the other and the correlation appropriately sums to zero. At various points within the grid the translational shift vectors are such that molecule $b_{l,m,n}$ will penetrate and significantly overlap molecule $a_{l,m,n}$. To penalize against this physically unrealistic interaction the product of the corresponding grid point values from both molecules (*i.e.,* –15 for molecule $a_{l,m,n}$ and +1 for molecule $b_{l,m,n}$) is negative and, depending on the extent of penetration, will sum to a large negative correlation. Finally when the surfaces of both proteins are in favorable proximity to one another the product of the surface layer values of molecule $a_{l,m,n}$ (+1) and the non-zero values of molecule

$b_{l,m,n}$ (+1) is positive and thus sum to a positive correlation. At the end of the entire calculation the intermolecular position that corresponds to the highest correlation is the orientation of highest surface complementarity.

Assessment of *de Novo* Docking Parameters: To not bias the docking results with wild-type residues it was necessary to prune all side chains to the C_β atoms (excluding glycines) and therefore it was not possible to use the GRA in its original form. Use of the original discretization parameters (without side chains) would result in a dimer complex with unrealistically small interfacial volume (*i.e.,* the monomers within the dimer complex would be too tightly packed). Unrealistically small interfacial volume would not be conducive for the side chain selection process performed in the subsequent step. Therefore the docking parameters listed above were reassessed and recalculated for the case in which the wild-type side chains were pruned to the C_β atom. To ascertain optimal discretization values an extensive analysis was performed on the crystal structures of a number of natural complexes. The goal of which was to extract from the natural complexes optimized parameters that would provide proper interfacial volume for successful side chain selection.

Eighteen Brookhaven Protein Data Bank files that contain the coordinates of protein complex structures, and comprise twenty-three unique protein interfaces, were analyzed for the purpose of extracting *de novo* docking parameters (see Materials and Methods for PDB entry codes). All complexes were projected and centered into a 3-dimensional grid where one protein (or protein pair) was treated as molecule $a_{l,m,n}$ and the other as molecule $b_{l,m,n}$. Many of the PDB structures analyzed consisted of antibody/antigen complexes and, when present, the antibody light chain and heavy chain were treated as one protein while the antigen was treated as the other. The intermolecular positions of both molecules were held constant (*i.e.*, positioned as found in the crystal structures) and prior to discretization all side chains were pruned to the C_β atoms. For the discretization analysis the surface skin thickness of molecule $a_{l,m,n}$ was kept constant and the same as that described above for natural docking (*i.e.*, 1.5 Å). The primary parameter varied in this analysis was the radial distance within which a grid point is assigned the non-zero values of either -15 for molecule $a_{l,m,n}$ or +1 for molecule $b_{l,m,n}$ (for natural docking with full side chains the radial distance used was 1.8 Å). While the position of each molecule was held stationary the radial distance was varied from 1.5 Å to 3.5 Å in 0.05 Å increments. At each increment the molecules were discretized and fast Fourier transform (FFT) and non-FFT correlations calculated. The distances for each complex that corresponded to the largest correlation were statistically analyzed and resulted in an average value of 2.05 Å ±0.48 Å. Due to the relatively large variance in the values for the different complexes a series of Protein G/Protein G GRA/FCT-derived dockings were separately calculated with radial distances of 2.00, 2.05, 2.10, 2.15 and 2.20 Å. The resulting docked complexes were analyzed with 3-dimensional molecular visualization tools (*e.g.*, GRASP, molmol, POV-Ray) and it was concluded that the complex that corresponded to the 2.15 Å radial distance had optimal interfacial volume.

Generation of the Starting Orientation: The overall goal of this project was to engineer a dimer complex of specific structure (as opposed to an indiscriminately formed complex). Therefore we chose to dock Protein G to itself in an orientation in which the surfaces of the α-helices were packed against one another. To generate two monomers oriented helix-face to helix-face the Protein G backbone coordinates were centered in a square grid (64 Å per side) and rotated so that the average plane of the β-sheet was approximately parallel to the x-y plane. The coordinates were further rotated so that long axis of the α-helix was roughly parallel to the y-axis and then translated 12 Å in the negative z-dimension resulting in the β-sheet surface positioned near to and relatively flush with the rear wall of the grid. This monomer was held stationary and is therefore referred to a monomer-A. The second monomer (monomer-B) was generated by copying the coordinates of the first, rotating them 180° about the y-axis, and then translating them in the positive direction along the z-axis so that the helices from both monomers faced one another. In addition, the second monomer was rotated 180° about the z-axis causing it to be flipped head-to-tail relative to monomer-A (Figure 3-2).

An additional modification to the GRA is related to the fact that, in the case of *de novo* docking, not all degrees of rotational freedom must be searched. Therefore, from the general translational starting position described above, the starting rotational position was generated by rotating monomer-B 45° about all three principle axes. To obtain the correlations as function of rotational position, a triple-nested loop was used to rotate monomer-B 90° about each axis (in 5° increments) and the modified GRA-FCT algorithm was used to calculate the correlation at each increment resulting in 5,832 separate correlation calculations. The top 15 correlations for each rotational position were stored and rank ordered upon completion to obtain the specific rotations and translational shift vectors that corresponded to the dimer complex of highest surface complementarity.

Docking Results: The rotations and translational shift vectors that correspond to the top 1000 complexes were used to generate PDB files containing coordinates of the docked complexes without side chains (except for the C_β atoms). Total buried surface areas and interfacial volumes were measured and gap indices calculated for the top 1000 complexes. The gap index is defined as the interfacial volume divided by the total buried surface area [7,8]. Programs used to measure these parameters were modified to handle proteins with truncated side chains. No discernable trend in the gap indices, as a function of docking scores, was observed (data not shown) although there were slight trends towards lower buried surface areas and lower interfacial volumes as a function of decreased docking score.

The complexes that corresponded to the top 50 docking scores were analyzed with 3-dimensional molecular visualization tools and it was concluded that the top scoring complex was the best candidate for the next step. A correlation map that corresponds to the rotational values of this complex is provided in the supplemental material and its structure is depicted in Figure 3-3. The coordinates of this docked complex were fed into the second overall step of the *de novo* docking process - the side chain selection process.

Interfacial Side Chain Selection via the ORBIT Suite of Algorithms

The ORBIT algorithms were used to mutate and repack residue positions located at the interface of the top docked complex. For the first step in this process the RESLASS algorithm (which classifies residues as core, boundary or surface based on their position in the molecule) was used to determine which residues are buried (or partially buried) in the complex relative to the free monomers. 15 positions, previously classified as surface or boundary, were reclassified as core and 7 surface positions were reclassified as boundary (Table 3-1). ORBIT was used to assess the energy of and select primarily hydrophobic side-chains for the 15 interfacial core positions and hydrophilic side-chains for the 7 reclassified boundary positions. Due to potentially favorable interfacial proximity 2 additional surface positions were included in the calculation. Figure 3-4 illustrates the side chains that were selected for by the ORBIT algorithms for the 24 calculated positions. The total redesign resulted in a 20-fold mutant (12 for monomer-A and 8 for monomer-B, 4 remained wild-type). Upon complex formation (with mutant side chains) the combined total buried surface area is ~1560 Å^2 (~76% of which is hydrophobic).

Protein Expression Levels and Thermal Stability

The genes for both monomers were synthesized with standard molecular biology methods, expressed in bacteria and purified to homogeneity. Interestingly, the protein expression levels for the two proteins, determined after HPLC purification, differed substantially with monomer-A expressing at levels similar to wild-type Protein G (*i.e.,* ~40 mg/L) and monomer-B at levels 10-fold lower (*i.e.,* ~4 mg/L). This finding may be related to the different thermal stabilities of each monomer and variable proteolytic resistance during bacterial expression.

Introduction of the specific mutations that resulted in the unique monomer-A and monomer-B amino acid sequences altered the physical properties of each monomer relative to wild-type Protein G. Standard thermal melts (monitored with circular dichroism) were performed on both monomers (data not shown) and it was determined that the 12 mutations that resulted in the monomer-A sequence inadvertently stabilized it to a hyperthermophile (*i.e.,* T_m > 100° C) while the 8 for monomer-B proved to be destabilizing, resulting in a T_m ~ 37°C (the T_m for wild-type Protein G is ~87° C). It is unclear as to why the monomer-A mutations resulted in a protein variant with a $T_m > 100^o$ C but, according to published results, it is fairly clear that the Y45A mutation introduced in monomer-B is the most likely cause of its reduced T_m. A short C-terminal fragment of wild-type Protein G is known to form a β-hairpin under physiological conditions. To understand the cooperative folding of the β-hairpin, structural stabilities were determined for eight different alanine substituted peptides[35,36]. The measured thermodynamic parameters indicated that the nonpolar residues Tyr 45 and Phe 52 and the polar residues Asp 46 and Thr 49 are crucial for β-hairpin folding and therefore also crucial for the stability of the complete β1-domain. In the monomer-B sequence only one of these destabilizing positions were mutated from wild-type (*i.e.,* Y45A) and therefore it is most likely this mutation that resulted in reduced thermal stability.

Analytical Ultracentrifugation

Sedimentation equilibrium experiments were performed on the *de novo* designed dimer complex. Runs were carried out at 28,000, 40,000 and 48,000 rpm, at 20° C on free monomer-A, free monomer-B and the monomer-A/monomer-B complex. Global nonlinear least-squares analysis of the data from the lowest speed in the initial run was consistent with weak dimerization, with a putative Kd of ~300 μM (data not shown). A complete global analysis of the data at all speeds and concentrations could not conclusively show that dimerization was occurring, as monomer B alone showed evidence of nonideality. Although there are difficulties in the analysis of this data, the initial results indicated that dimerization was occurring, in agreement with preliminary

1D NMR analysis. We therefore pursued multidimensional heteronuclear NMR analysis on the *de novo* complex.

NMR Spectroscopy

NMR spectra of wild-type Protein G are well dispersed and exhibit homogeneous line widths and peak shapes that are indicative of a globular, well-folded protein. Well-dispersed peak shapes were observed in preliminary NMR spectra collected at 20° C on both designed monomers while free in solution, although the monomer-B peaks were not quite as sharp or homogeneous as those for monomer-A. These findings indicate that both monomers adopt the general tertiary fold of Protein G but monomer-B may be more mobile and dynamic in accordance with its lower thermal stability relative to monomer-A and wild-type Protein G.

Chemical shift perturbation analysis of preliminary 1D NMR spectra on non-isotopically labeled monomer-A in the presence of unlabeled monomer-B indicated successful complex formation. Isolated peaks in the upfield aliphatic region for monomer-A were shifted in the presence of monomer-B in comparison to the spectrum of free monomer-A (data not shown).

To further confirm successful complex formation monomer-A was selectively labeled with ^{15}N and 2D [^{1}H, ^{15}N] HSQC spectra were collected on two different samples - one sample contained free ^{15}N-monomer-A while the other contained both ^{15}N-monomer-A and an equimolar quantity of unlabeled monomer-B (Figure 3-5). For the heteronuclear experiments monomer-A was isotopically labeled because of its superior expression levels relative to monomer-B. In a manner similar to the 1D analysis, complex formation was confirmed by chemical shift perturbation of a number of [^{1}H, ^{15}N] peaks. In addition to the observed chemical shift perturbations a number of peaks also broadened to the point where they were no longer detectable.

To ascertain whether the monomers were associating in the target orientation it was necessary to determine the [^{1}H, ^{15}N] assignments of the peaks that had shifted and broadened upon complex formation. To accomplish this we determined the complete backbone [^{1}H, ^{15}N] resonance assignments of free ^{15}N-monomer-A by analyzing 3D-[^{1}H, ^{15}N]-NOESY-HSQC and 3D-[^{1}H, ^{15}N]-TOCSY-HSQC collected on the free monomer.

The assigned 2D-[^{1}H, ^{15}N]-HSQC peaks of free ^{15}N-monomer-A were qualitatively compared to those of the spectra in the presence of equimolar amounts of unlabeled monomer-B. With few exceptions (Y3, K13 and A48) the peaks that exhibited chemical shift perturbations mapped in close proximity to the putative interface of the target orientation (Figure 3-6).

DISCUSSION

Computational Docking with the Modified GRA

The goal of the computational docking calculations was to obtain the Protein G/Protein G dimer orientation of highest surface complementarity in the context of a general starting orientation. A number of factors were necessarily taken into account when choosing the starting orientation of the dimer complex. When viewed down its long axis the overall structure of Protein G is roughly triangular in shape (Figure 3-1). The tertiary structure consists of an α-helix that overlays and packs against a four-stranded β-sheet. One vertex of the overall triangular shape is formed by the outer surface of the α-helix while the other two vertices consist of the edges of the β-sheet. For the general starting position we chose to dock either of two helical faces against one another (*i.e.,* the β-sheet surface was not considered as a possible interfacial surface). This general orientation was chosen because amino acid positions on or near the surface of the α-helix have been shown to be relatively permissible to mutation without inducing gross tertiary structural perturbations[22,37].

Although we had dictated the general starting orientation, and therefore did not need to search all rotational and translational space, there was still a need to rigorously search local space to find the optimal surface-to-surface geometric fit. To accomplish this we modified a well-established algorithm used extensively in the field of native protein docking; the geometric recognition algorithm (GRA)[28,29]. The GRA treats the two molecules as rigid bodies and uses surface complementarity as the criteria for goodness of fit. Traditionally, it has been used to dock the crystal structures of single proteins together that are known to form complexes in solution. The goal of which is to predict the structure of the actual protein/protein complex as it would exist in solution. Initially the use of the GRA led to modest success due to the fact that only purely geometric factors were used to assess the multitude of docked candidate complexes[28]. It was recognized that physical chemical factors also needed to be taken into account when assessing the validity of docked candidates[29].

Our approach differed from that described above in that we used the GRA solely in the first step of the design where it was used to dock the protein molecules together (and where only shape complementarity is used as the criteria for goodness of fit). It is in the second step (*i.e.,* side chain selection using the ORBIT suite of design algorithms) that physical chemical forces were explicitly calculated to determine the optimal set of mutated interfacial amino acids. To insure the success of the second step it was necessary to modify the GRA from its originally published form. A main function of the first step was to determine the intermolecular positions of the two Protein G monomers in which the greatest set of amino acid positions are in proximity across the *de novo* interface. In addition, the docking calculations must result in a complex with interfacial spacing conducive to side chain selection. We did not want to bias the results with wild-type Protein G side chains and therefore all side chains were pruned down to the C_β atoms (excluding glycines). Due to this fact that it was necessary to reevaluate the docking parameters originally described for native docking[28,29]. To this end a series of natural complexes were extensively analyzed and parameters unique to *de novo* docking extracted. The primary outcome of the analysis was that the radial distance used to assign core values for both molecule $a_{l,m,n}$ and molecule $b_{l,m,n}$ was increased from 1.8 Å to 2.15 Å; an increase that reflects the loss of the side chain atoms. Further analysis revealed that the optimal the skin layer thickness for molecule $a_{l,m,n}$ was that described for native docking (*i.e.,* 1.5 Å). Using these values, in combination with the other modifications described above, Protein G was docked to itself and resulted in a dimer complexes of high surface complementarity.

In an attempt to ascertain whether the highest docking score corresponded to the best candidate for interfacial side chain selection the top 1000 complexes (without side chains) were generated and subjected to additional analysis. The gap index, which is calculated by dividing the interfacial volume by the total buried surface area, was determined for the top 1000 complexes. No discernable trend in the gap indices, as a function of docking scores, was observed. Not surprisingly there were slight trends towards lower buried surface areas and interfacial volumes as a function of decreased docking score. These results reflect the fact that, for the top 1000 scores, there are only minor differences in the structural characteristics of the complex interfaces. This may be related to the translational and rotational resolution used or to the fact that, aside from gross tertiary structure features, the surfaces of both molecules (especially with truncated side chains) are relatively featureless and therefore comparatively homogeneous. In addition, the surface layer thickness of molecule $a_{l,m,n}$, set to 1.5 Å to allow for interfacial plasticity, may act to further attenuate the features of both molecules and thus reduce the calculated differences in the interfacial characteristics. Regardless, the complex that corresponds to the top docking score exhibited excellent geometric fit between the docked molecules as illustrated in Figure 3-3. The coordinates of this complex were used in the next step – the interfacial side chain selection process.

Interfacial Side Chain Selection via the ORBIT Suite of Design Algorithms

An integral step in the *de novo* docking process entailed the use of the ORBIT suite of protein design algorithms[16]. The primary function of the ORBIT algorithms is to return a mutated protein sequence optimized for a given three-dimensional backbone structure[38]. They employ an unbiased, quantitative design method based on the physical chemical properties that determine protein structure and stability[39].

The coordinates of both monomers from the top docked complex obtained in the first step were treated as a single protein and fed into the second step where interfacial residue positions were mutated and repacked in a manner similar to that observed for the cores of well-folded proteins. The rationale for this approach is based in the fact that natural interfaces have physical characteristics that are similar to protein cores. Examination of the crystal structures of protein-protein complexes provided strong evidence that hydrophobic residues play a principal role for in protein-protein association[7-9]. For example, alanine scanning performed on the receptor for human growth hormone revealed a central hydrophobic region, dominated by two tryptophan residues, that accounted for more than three-quarters of the binding free energy[4]. In general, and depending on the degree of association, it has been observed that protein interfaces are more hydrophobic than protein exteriors, yet are slightly more polar than protein interiors. This is probably due to the need to promote association without greatly destabilizing the unassociated monomers[7-9]. In the calculated *de novo* complex approximately 76% of the buried surface area is hydrophobic.

In addition to the general hydrophobicity of natural interfaces, the interface area per subunit (A_i) calculated for crystal structure complexes was found to range from 368

Å^2 to 4746 Å^2, which represents 6.5% - 29.4 % of the total exposed surface area per subunit[7,8].

A_i is defined as: $A_i = ([A_{sa} + A_{sb}] - A_{sab}) / 2$

where A_{sa} and A_{sb} denote the total surface area of each disassociated subunit and A_{sab} denotes the total surface area of the subunits associated in a dimer. The total buried area upon complex formation for the *de novo* calculated complex is approximately 1,560 Å^2 and therefore A_i = 780 Å^2. This represents 22.8% of the total exposed surface area per subunit for the *de novo* complex; a percentage well within the range described above.

Modest Affinity as Measured by Analytical Ultracentrifugation

A complete analysis of the analytical ultracentrifugation runs performed on each of the separate monomers and the monomer-A/monomer-B complex was not possible due to the fact that free monomer-B exhibited evidence of nonideality. Even though this was the case, global nonlinear least-squares analysis of the data from the lowest speed in the initial run was consistent with modest dimerization, with a putative Kd of ~300 _M (data not shown). Subsequent NMR analysis and determination of the thermal stabilities of both monomers revealed a potential cause for the nonideal behavior observed for monomer-B (see below).

Thermal Stability and Fiber Formation

Thermal stability was measured with CD temperature melts for both monomers. The 12 mutations that resulted in the monomer-A sequence stabilized it to a hyperthermophile (*i.e.,* T_m > 100° C) while the 8 for monomer-B proved to be destabilizing, resulting in a T_m ~ 37° C (the T_m for wild-type Protein G is ~87° C). Interestingly, at the concentrations required for NMR studies, monomer-B alone was observed to form macroscopic fibrils in the NMR tube (Figure 3-7). This, in itself, was not surprising, since many proteins upon destabilization can be induced to form fibers[40-45]. What is interesting is that in the presence of monomer-A no monomer-B fibrils were observed. This finding provided macroscopic evidence that a complex does indeed form between these designed proteins and also provided an excellent model system for studying protein-based fibril inhibition. A more extensive analysis of monomer-B fibrilliation, monitored with electron microscopy and thioflavine-T fluorescence, confirmed that monomer-A does completely block monomer-B fiber formation and most likely does so by binding to and stabilizing the correctly folded form of monomer-B (manuscript in preparation). The fact that free monomer-B is prone to self-association, giving rise to visible fibers, provides a likely explanation for the nonideal behavior observed for monomer-B during the analytical ultracentrifugation analysis.

Interface Site Mapped on Monomer-A surface with NMR Spectroscopy

To determine if the monomers were binding in the general target orientation multidimensional NMR spectra were collected on free ^{15}N-monomer-A and compared to spectra collected on ^{15}N-monomer-A in complex with an equimolar amount of unlabeled monomer-B.

A 2D-[^{1}H, ^{15}N] HSQC spectrum of free ^{15}N-monomer-A (Figure 3-5, panel **a**) contains the expected number of backbone amide peaks (~55), exhibits significant dispersion and is quite homogeneous in line width and peak intensity. The observed homogeneous line widths indicate that monomer-A has a relatively compact, stable and well-folded structure which in good agreement with the measured T_m for monomer-A ($T_m > 100^\circ$ C). In addition, the homogeneous line widths indicate that monomer-A is tumbling at rate appropriate for its molecular weight (~6200 D) and, in agreement with analytical ultracentrifugation results, is monomeric even at relatively high NMR concentrations (*i.e.,* between 1 and 2.5 mM). In contrast, line widths in the 2D-[^{1}H,^{15}N] HSQC spectrum of the ^{15}N-monomer-A in complex with unlabeled monomer-B are broader and not nearly as homogeneous as that for free monomer-A (Figure 3-5, panel **b**). This behavior is indicative of a change in the NMR relaxation properties usually associated with a decrease in the tumbling rate caused by an increase in the apparent molecular weight of the molecule. The broader, less homogeneous peaks provide indirect evidence that a dimeric complex does form between monomer-A and monomer-B. In addition, the chemical shifts of a subset of peaks shifted upon complex formation and some peaks were rendered completely non-observable. The peaks rendered non-observable are probably the result of exchange broadening due to fluctuations between two or more states on a microsecond-millisecond time scale and imply that the complex may experience a relatively rapid exchange between free and bound states. This explanation appears likely especially in light of the modest binding affinity measured with analytical ultracentrifugation (Kd ~300 μM).

To ascertain if monomer-B was binding at the targeted surface of monomer-A it was necessary to determine the complete backbone [^{1}H, ^{15}N] resonance assignments of free ^{15}N-monomer-A by analyzing 3D-[^{1}H, ^{15}N] NOESY-HSQC and 3D-[^{1}H, ^{15}N] TOCSY-HSQC collected on free monomer-A. Complete backbone assignments were made for free ^{15}N-monomer-A which enabled the assignment of the peaks that shifted and broadened upon complex formation. Almost all the peaks that exhibited chemical shift perturbation corresponded to interfacial residue positions (Figure 3-6). The three exceptions (*i.e.,* Y3, K13 and A48) may be due to a number of factors such as the observed changes in relaxation parameters described above for monomer-A or they could represent an alternative binding mode that may be sampled to a lesser degree.

Conclusion

This initial attempt at the *de novo* docking of a normally monomeric protein to itself proved to be successful at generating a structurally specific dimer complex, albeit with a modest dissociation constant (~300 μM). The GRA was successfully modified and adapted to rigorously search local rotational and translational space of two protein monomers (with truncated side chains) and return a complex docked in the confines of a general starting orientation. The coordinates of the docked complex were treated as one protein and the ORBIT suite of protein design algorithms were used to mutate and repack interfacial side chains with the goal of selecting a sequence that would ultimately drive complex formation. Protein interactions between the two monomers were confirmed with analytical ultracentrifugation, fiber inhibition and heteronuclear NMR spectroscopy. The fact that the binding affinity is modest, combined with the associated exchange broadening of interfacial NMR peaks, precludes structure elucidation by multidimensional NMR. Work is currently underway to improve the binding affinity with additional rounds of docking and ORBIT-based interfacial side chains selection.

MATERIALS AND METHODS

Computational Docking

The algorithm used to dock the backbone (plus C_β atoms) of Protein G to itself is a modified version of that originally described in [28] and [29]. The necessary modifications are described in detail in the Results section. The algorithm, complete with modifications, was implemented using the C programming language. All real 3-dimensional discrete Fourier transforms (both forward and inverse), that were implemented to significantly reduce the complexity of the docking calculations, are those described in Numerical Recipes in C[46].

The entry codes of the 18 PDB files used to extract *de novo* docking parameters are - 1ATN, 1BRS, 1DQJ, 1DZB, 1FCC, 1FDL, 1HRP, 1IGC, 1JHL, 1JTO, 1LPA, 1MLC, 1NCD, 1VFB, 2BTF, 2JEL, 3HFL, 3HFM.

The arrays used for docking were N x N x N points with N = 128. The grid was square with each side spanning 64 Å resulting in a translation resolution of 0.5 Å. The coordinates of molecule $b_{l,m,n}$ were rotated 90° about all three principle axes and the correlation calculated for every 5° increment resulting in 5,832 separate correlations. The top 15 correlations were stored for each rotational position and the total of all scores were rank ordered to obtain the highest score. The rotational and translational shift vectors that corresponded to the highest score were used to generate a dimer complex that was used in the interfacial side chain selection process.

Interfacial Side Chain Selection (ORBIT)

Standard ORBIT design parameters were used.

Protein Expression and Purification

Synthetic DNA oligos were used for recursive PCR synthesis of the genes for monomer-A and monomer-B. One modification to the ORBIT selected sequence was made: position 28 of monomer-A was changed from tryptophan to tyrosine so as to reduce the hydrophobicity of regions that were not fully buried at the interface. The genes were cloned into pET-11a (Novagen) and recombinant protein was expressed by IPTG induction in BL21(DE3) hosts (Invitrogen). The proteins were isolated using a freeze/thaw method [47] and purification was accomplished with reverse-phase HPLC using a linear 1% min^{-1} acetonitrile/water gradient containing 0.1% TFA. The yield of purified protein from expression in rich media was ~4 mg/L of bacterial growth for monomer-B and ~40 mg/L for monomer-A. Labeled monomer-A protein, for NMR studies, was prepared with standard M9 minimal media using ^{15}N-ammonium sulfate (2 g/L). Protein purity was verified with standard SDS-PAGE and reverse phase HPLC and the correct molecular weight was confirmed by mass spectrometry.

Analytical Ultracentrifugation

Sedimentation equilibrium experiments were conducted in a Beckman XL-I Ultima analytical ultracentrifuge equipped with absorbance optics. Runs were carried out at 28,000, 40,000 and 48,000 rpm, at 20^{o} C. Three separate solutions consisting of different concentrations for both free proteins plus the complex were prepared to achieve

OD_{280} readings of approximately 0.15, 0.25 and 0.4. The protein concentrations were ~36 μM, ~60 μM and ~96 μM in a total volume of 110 μl containing 50 mM NaCl and 50 mM NaPi (pH~6.5).

NMR Spectroscopy

NMR spectra were collected at 293 K on a Varian UnityPlus 600 MHz spectrometer equipped with an HCN-triple-resonance probe with triple-axis pulse field gradients. Protein concentrations were ~1.5 mM in 50 mM sodium phosphate, pH ~6.5.

References

1. Nooren, I. M. & Thornton, J. M. Diversity of protein-protein interactions. *Embo J* **22**, 3486-92 (2003).

2. Nooren, I. M. & Thornton, J. M. Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol* **325**, 991-1018 (2003).

3. Wells, J. A. Binding in the growth hormone receptor complex. *Proc Natl Acad Sci U S A* **93**, 1-6 (1996).

4. Clackson, T. & Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383-6 (1995).

5. Janin, J. & Wodak, S. J. Protein modules and protein-protein interaction. Introduction. *Adv Protein Chem* **61**, 1-8 (2002).

6. Janin, J. & Seraphin, B. Genome-wide studies of protein-protein interaction. *Curr Opin Struct Biol* **13**, 383-8 (2003).

7. Jones, S. & Thornton, J. M. Protein-protein interactions: a review of protein dimer structures. *Prog Biophys Mol Biol* **63**, 31-65 (1995).

8. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20 (1996).

9. Stites, W. E. Protein-Protein Interactions: Interface Structure, Binding Thermodynamics, and Mutational Analysis. *Chemical Reviews* **97**, 1233-1250 (1997).

10. Amstutz, P., Forrer, P., Zahnd, C. & Pluckthun, A. In vitro display technologies: novel developments and applications. *Current Opinion in Biotechnology* **12**, 400-405 (2001).

11. Matsuura, T., Ernst, A., Zechel, D. L. & Pluckthun, A. Combinatorial Approaches To Novel Proteins. *Chembiochem* **5**, 177-182 (2004).

12. Serebriiskii, I. G. et al. Detection of peptides, proteins, and drugs that selectively interact with protein targets. *Genome Res* **12**, 1785-91 (2002).

13. Feldhaus, M. & Siegel, R. Flow cytometric screening of yeast surface display libraries. *Methods Mol Biol* **263**, 311-32 (2004).

14. Atwell, S., Ridgway, J. B., Wells, J. A. & Carter, P. Stable heterodimers from remodeling the domain interface of a homodimer using a phage display library. *J Mol Biol* **270**, 26-35 (1997).

15. Takahashi, T. T., Austin, R. J. & Roberts, R. W. mRNA display: ligand discovery, interaction analysis and beyond. *Trends Biochem Sci* **28**, 159-65 (2003).

16. Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82-7 (1997).

17. Chevalier, B. S. et al. Design, activity, and structure of a highly specific artificial endonuclease. *Mol Cell* **10**, 895-905 (2002).

18. O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* **254**, 539-44 (1991).

19. Havranek, J. J. & Harbury, P. B. Automated design of specificity in molecular recognition. *Nat Struct Biol* **10**, 45-52 (2003).

20. Shifman, J. M. & Mayo, S. L. Modulating calmodulin binding specificity through computational protein design. *J Mol Biol* **323**, 417-23 (2002).

21. Shifman, J. M. & Mayo, S. L. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc Natl Acad Sci U S A* **100**, 13274-9 (2003).

22. Malakauskas, S. M. & Mayo, S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* **5**, 470-5 (1998).

23. Gronenborn, A. M. et al. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **253**, 657-61 (1991).

24. Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* **33**, 4721-9 (1994).

25. Sauer-Eriksson, A. E., Kleywegt, G. J., Uhlen, M. & Jones, T. A. Crystal structure of the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG. *Structure* **3**, 265-78 (1995).

26. Derrick, J. P., Davies, G. J., Dauter, Z., Wilson, K. S. & Wigley, D. B. Crystallization and preliminary X-ray analysis of the complex between a mouse Fab fragment and a single IgG-binding domain from streptococcal protein G. *J Mol Biol* **227**, 1253-4 (1992).

27. Derrick, J. P. & Wigley, D. B. Crystal structure of a streptococcal protein G domain bound to an Fab fragment. *Nature* **359**, 752-4 (1992).

28. Katchalski-Katzir, E. et al. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A* **89**, 2195-9 (1992).

29. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* **272**, 106-20 (1997).

30. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in Fortran* (Cambridge University Press, Cambridge, 1986).

31. Bracewell, R. N. Numerical Transformations. *Science* **248**, 697-704 (1990).

32. Shimaoka, M. et al. Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat Struct Biol* **7**, 674-8 (2000).

33. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A* **98**, 14274-9 (2001).

34. Marshall, S. A. & Mayo, S. L. Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* **305**, 619-31 (2001).

35. Kobayashi, N., Honda, S., Yoshii, H. & Munekata, E. Role of side-chains in the cooperative beta-hairpin folding of the short C-terminal fragment derived from streptococcal protein G. *Biochemistry* **39**, 6564-71 (2000).

36. Honda, S., Kobayashi, N. & Munekata, E. Thermodynamics of a beta-hairpin structure: evidence for cooperative formation of folding nucleus. *J Mol Biol* **295**, 269-78 (2000).

37. Street, A. G. & Mayo, S. L. Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* **3**, 253-8 (1998).

38. Street, A. G. & Mayo, S. L. Computational protein design. *Structure Fold Des* **7**, R105-9 (1999).

39. Gordon, D. B., Marshall, S. A. & Mayo, S. L. Energy functions for protein design. *Curr Opin Struct Biol* **9**, 509-13 (1999).

40. Morozova-Roche, L. A. et al. Amyloid fibril formation and seeding by wild-type human lysozyme and its disease-related mutational variants. *J Struct Biol* **130**, 339-51 (2000).

41. Ramirez-Alvarado, M. & Regan, L. Does the location of a mutation determine the ability to form amyloid fibrils? *J Mol Biol* **323**, 17-22 (2002).

42. Khurana, R. et al. Partially folded intermediates as critical precursors of light chain amyloid fibrils and amorphous aggregates. *Biochemistry* **40**, 3525-35 (2001).

43. Kheterpal, I., Williams, A., Murphy, C., Bledsoe, B. & Wetzel, R. Structural features of the Abeta amyloid fibril elucidated by limited proteolysis. *Biochemistry* **40**, 11757-67 (2001).

44. Lee, S. & Eisenberg, D. Seeded conversion of recombinant prion protein to a disulfide-bonded oligomer by a reduction-oxidation process. *Nat Struct Biol* **10**, 725-30 (2003).

45. West, M. W. et al. De novo amyloid proteins from designed combinatorial libraries. *Proc Natl Acad Sci U S A* **96**, 11211-6 (1999).

46. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1988).

47. Johnson, B. H. & Hecht, M. H. Recombinant proteins can be isolated from E. coli cells by repeated cycles of freezing and thawing. *Biotechnology (N Y)* **12**, 1357-60 (1994).

48. Sanner, M. F., Olson, A. J. & Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305-20 (1996).

49. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281-96 (1991).

Table 3-1. Amino Acid Sequence for Wild-Type Protein-G and the ORBIT Selected Sequences for Monomer-A and Monomer-B

	1 10 20 30 40 50
	| | | | | |
WT –	MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
	c c ss c cc scc scs
Mon-A –	MTYKLILNGKTLKGE**F**T**A**EA**E**DAA**L**AE**YI**F**RAL**A**KAQ**GVDGEWTYDDATKTFTVTE
	scc cs c c s c c s
Mon-B –	MTYKLILNGKTLKGETTTEAVD**I**ATA**AD**VF**A**QYA**A**DNGV**K**GEWT**A**D**E**ATKTFTVTE

c = core, s = surface, bold = a mutated position, underline = calculated position that is wild-type

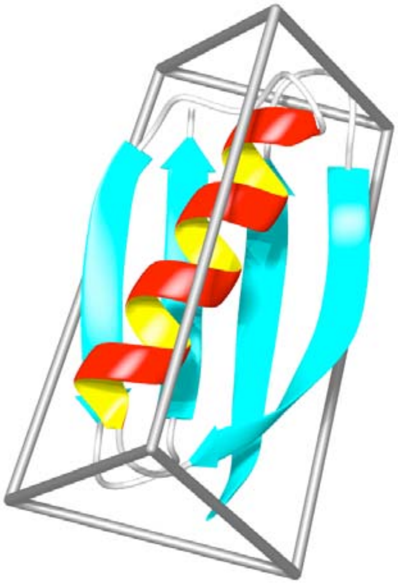


Figure 3-1. Secondary Structure of the β1 Domain of Protein G. A ribbon diagram illustrating the relatively simple secondary structure arrangement of the β1 domain of the Streptococcal Protein G. An α-helix overlays and packs against a β-sheet made up of four strands. The triangular prism shaped bars are meant to illustrate the global triangular shape of the β1 domain where the α-helix makes up one vertex and the edges of the β-sheet the other two.

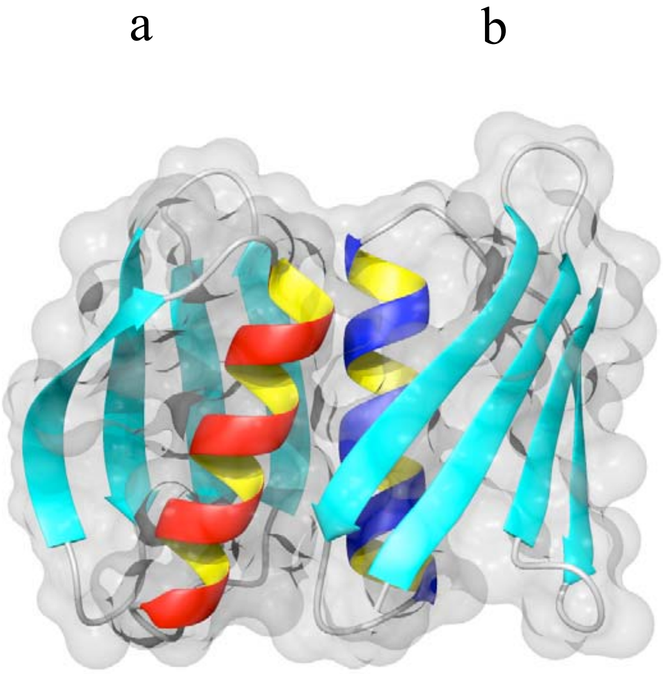


Figure 3-2. General Starting Orientation. A ribbon diagram illustrating the general starting orientation used in the first step of *de novo* docking, *i.e.,* application of the modified GRA to computationally dock the Protein G backbones together in the context of the general starting orientation. A solvent accessible surface was generated around each monomer with the program MSMS [48] and used to create the ghosted image of the actual surface of the backbone atoms. The figure on the left (**a**) corresponds to monomer-A while that on the right (**b**) corresponds to monomer-B.

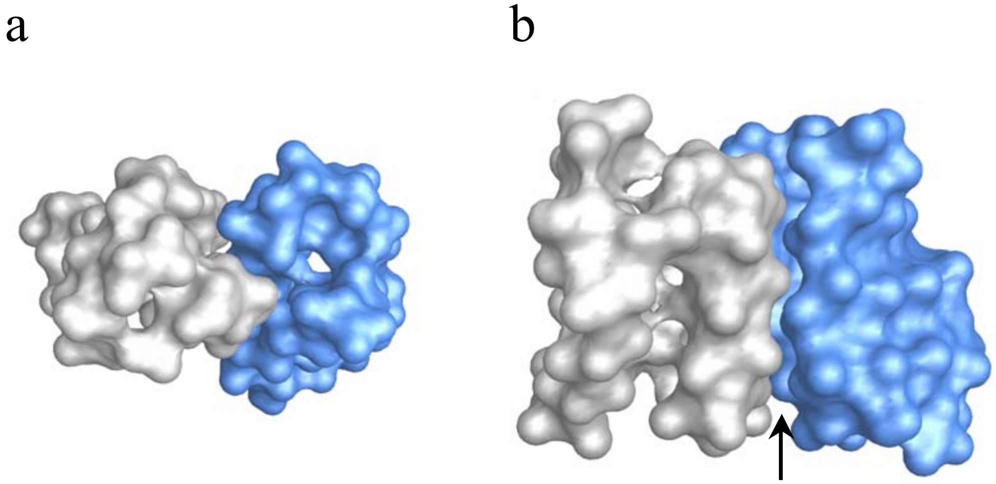


Figure 3-3. GRA Docking Results, the Top Docked Complex. A GRASP surface image [49] of the docked complex that exhibited the highest surface complementarity (*i.e.,* the highest GRA docking score). Image **a**, a top down view of that shown in **b**, illustrates the tight geometric fit between the monomers where protruding knobs from one monomer (generated, for example, by the C_β atoms) fit into the holes (or concave compressions) on the other. The arrow in image **b** points to the interface between the two monomers and is meant to draw attention to the high degree of interdigitation exhibited at the dimer interface. The coordinates of this docked complex were used in the next step of the computational docking process, *i.e.,* ORBIT-based side chain selection.

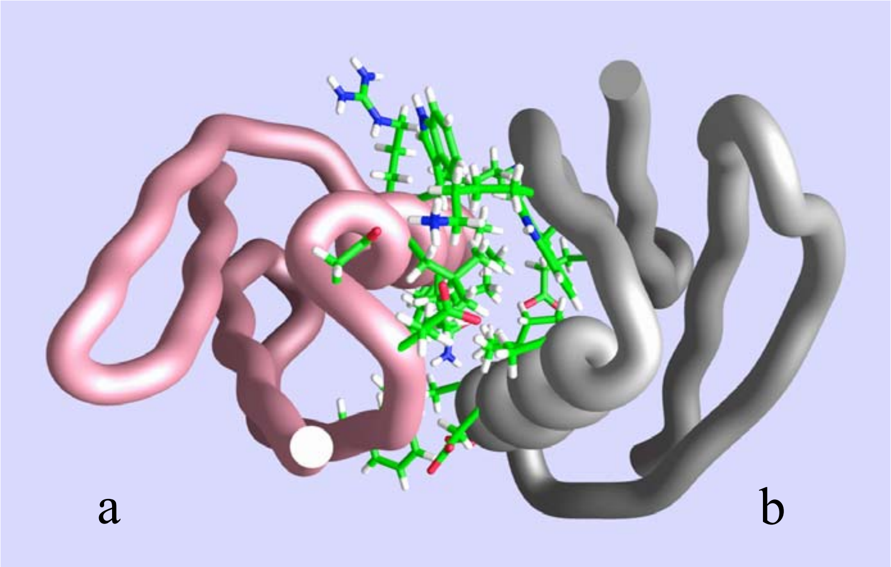


Figure 3-4. Side Chain Selection Results. This image only displays the side-chains (as balls and sticks) of the 24 positions for which ORBIT was used to select for interfacial side chains. The total redesign resulted in a 20-fold mutant with 12 mutations for monomer-A (**a**) and 8 for monomer-B (**b**) (ORBIT selected wild-type amino acids for 4 positions). Upon complex formation the mutant monomers bury ~1560 Å^2 of surface area (~76% of which is hydrophobic).

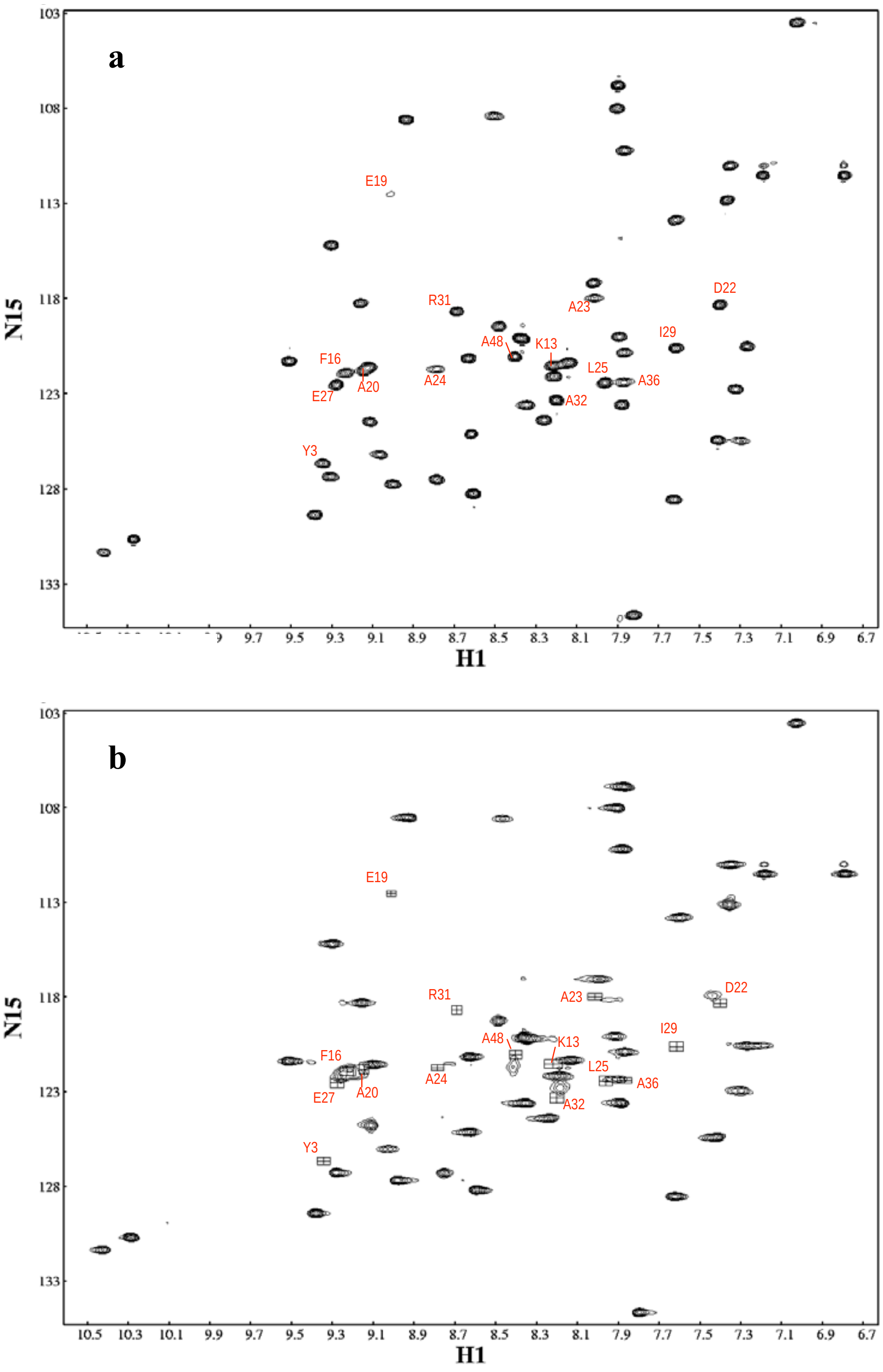

a
b
N15
H1
103
108
113
118
123
128
133
10.5
10.3
10.1
9.9
9.7
9.5
9.3
9.1
8.9
8.7
8.5
8.3
8.1
7.9
7.7
7.5
7.3
7.1
6.9
6.7
E19
R31
A23
D22
I29
A48
K13
F16
L25
A36
A20
A24
E27
A32
Y3

Figure 3-5. [^{1}H, ^{15}N] HSQC Spectra. (**a**) [^{15}N, ^{1}H] HSQC spectrum of uniformly enriched ^{15}N-monomer-A alone and (**b**) with equimolar quantities of unlabeled-B. The ^{15}N-monomer-A peaks that are non-observable or exhibit chemical shift perturbations upon complex formation are labeled red. For peaks that shifted, or were rendered non-observable, the original peak positions from (**a**) are illustrated with a black crossed box in (**b**). (Figure on previous page)

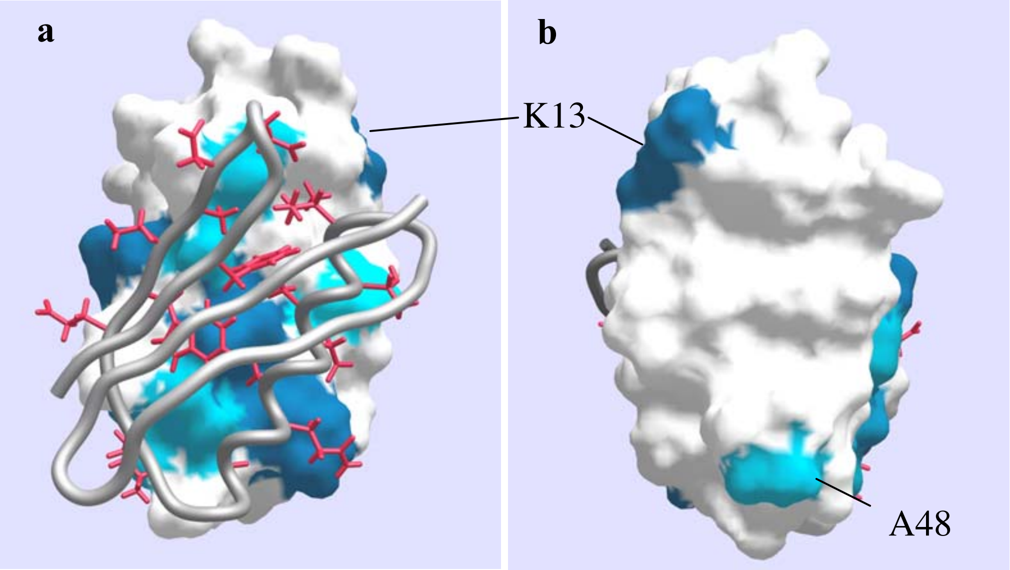


Figure 3-6. Chemical Shift Perturbations Mapped to the Surface of Monomer-A. The program GRASP [49] was used to generate the above images where chemical shift perturbations are mapped onto the surface of ^{15}N-monomer-A. Monomer-B is depicted as a gray backbone worm with interfacial side-chains colored red. On the monomer-A surface, residues that have [^{15}N, ^{1}H]-HSQC peaks that are not detectable in the complex are colored dark blue and those that exhibit chemical shift changes are colored lighter blue. The image on the left (**a**) corresponds to the interface of the putative target orientation while that in (**b**) is the surface of β-sheet of monomer-A (*i.e.,* **b** is a 180° rotation relative to **a**).

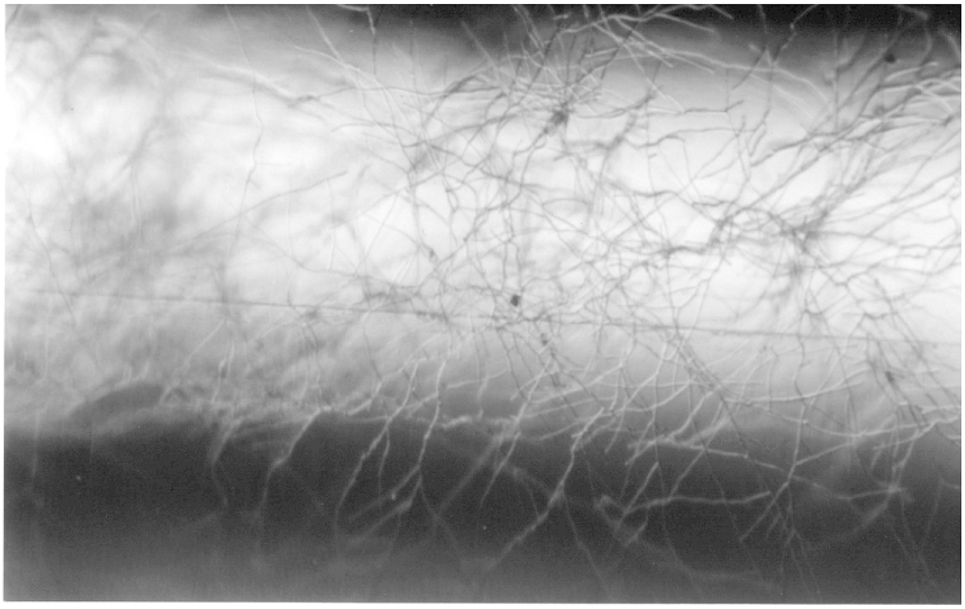


Figure 3-7. Macroscopic Monomer-B Fibers. This image is a photograph of the fibers that were observed to grow in an NMR tube that contained free unlabeled monomer-B. The concentration of monomer-B for NMR analysis was approximately 1.5 mM. Fibers were observed to spontaneously form in the NMR tube after approximately three days.

Chapter 4

Computational Design and Experimental Characterization of *De Novo* Homodimers

ABSTRACT

C2 symmetry related homodimers are the simplest form of self-associating protein oligomers. The goal of this project is to see whether we can generate novel protein homodimers through the combined use of a protein docking algorithm and a protein design algorithm. For this purpose, we developed a specialized docking algorithm that examines all possible homodimer conformations and rank orders them based on their surface complementarities. During the docking process, amino acid side-chains are pruned to C_β, and the remaining atoms are represented by spheres with experimentally determined radii to estimate the volume originally taken up by the side-chains. A docked dimer scaffold was chosen for the subsequent design step. We fine-tuned interfacial hydrophobicity by adjusting the penalty for polar surface area burial. Two engrailed homeodomain variants carrying the designed interface were experimentally constructed and characterized by CD, NMR, analytical ultracentrifugation, and X-ray crystallography methods. The results suggest that we have successfully created proteins that can self-assemble into dimers.

INTRODUCTION

In the most fundamental processes of life, protein molecules interact with each other and with other constituents of the living cell through highly regulated and specific binding and dissociation events. Cellular biochemistry is largely carried out by proteins and their functional amino acid moieties. The most striking examples of protein-protein interactions are found in antibodies, which bind numerous targets mainly through six short cannonical loops with variable amino acid composition. The twenty amino acids found in natural proteins have different properties; some are hydrophobic, some are hydrophilic, and each has a particular shape and size. The variations among the amino acid side-chains allow for the diversity observed in protein binding interfaces. The simplest dimers, C2 symmetry related homodimers, are the most abundant protein assemblies found in nature, and their binding interfaces are more hydrophobic than other types of dimers. It is plausible that hydrophobicity drives dimer formation in naturally-ocurring protein dimers, and that this phenomenon might be relevant to computational protein design.

Previous successes in protein design benefit largely from the practice of assigning amino acid positions to surface, boundary and core categories based on the amount of burial. This practice effectively partitions the protein into a well-defined hydrophobic interior and polar exterior, and increases the likelihood of a successful design by reducing the complexity of the calculation.

We are interested in generating novel protein homodimers starting from a monomeric protein crystal structure. Borrowing from the same methods used for protein core redesigns, we employed hydrophobicity as the driving force for the design. We developed a specialized docking algorithm that finds the most geometrically complementary homodimers among all possible homodimer configurations. Designing hydrophobic interfaces using dimer scaffolds that bury the most surface area partially mimics the homodimers found in nature; in addition, this procedure helps to reduce the problem size. The physical and chemical principles we follow are well established, but to our knowledge, this is the first dimer design effort that incorporates docking and design algorithms.

RESULTS

Modeling C2 Symmetry Related Homodimers

Our strategy of designing novel protein homodimers from monomeric proteins requires a modeling step in which the monomeric protein crystal structure is positioned against a copy of itself by a two-fold symmetry operation to form the dimer scaffold. To achieve this, we developed a fast Fourier transform (FFT)-based docking algorithm parameterized against known high-resolution homodimer structures. The parameterization consists of (1) a reduced representation of the amino acid side-chains that uses spheres on the C_β atoms instead of an all-atom description, (2) a set of experimentally determined atomic radii for the backbone atoms that is validated against known homodimer crystal structures, and (3) a 3D grid discretization strategy that emphasizes rigid protein backbones. The reduced side-chain representation ensures the proper spacing between the two docked subunits and smoothes the protein surface to reduce noise in the docking searches. Similarly, the use of experimentally determined atomic radii (including the radius used for the C_β spheres) favors docked dimer configurations that bury interface volumes similar to those buried by natural homodimers. Because our protein design methodology does not allow backbone flexibility, strong penalties are incorporated in our docking algorithm's scoring function to exclude docked molecules with overlapping backbones.

The surface complementarities of all possible C2 related homodimer conformations are evaluated through FFT correlation calculations, originally developed to predict protein complexes from their unbound members[1]. A list of possible dimers are generated and ranked according to their surface complementarity scores. High-scoring dimers can then be parsed into clusters, each containing only dimers of the same conformation. These clusters of distinct dimer conformations are then visually inspected to identify the most plausible conformation as our design target.

We chose the *Drosophila* engrailed homeodomain structure (PDB code: 1enh) as our starting monomer because of its helical structure and the availability of several thermally stable variants[2,3]. The helical structure offers several possible binding faces and is easier to stabilize with protein design than β-sheet structures. From the top 200 docked engrailed homeodomain dimers, we obtained 11 distinct dimer conformations. One of them was chosen as our final design scaffold based on the relative orientation of the helices, the amount of buried surface area, and the topology of the interface. Despite being ranked second in the docking correlation scores, the selected dimer conformation has the flattest and most tightly arranged interface. In this particular conformation, helix-1 and helix-2 from each of the subunits form a novel parallel helix bundle (Figure 4-1). This four-helical bundle resembles a naturally occurring coiled-coil in that it has a similar interhelix distance. The two helix-1 helices in the docked dimer are separated by 10 Å, similar to the 9.8 Å found in a coiled-coil dimer[4,5]. However, the helices in the engrailed homeodomain structure are straight, so there is no superhelical structure in our model. This homodimer model created by docking two engrailed homeodomain monomers is the template used in all the subsequent design steps.

Design Strategies

We use the ORBIT (Optimization of Rotamers by Iterative Techniques) protein design tools to generate the amino acid sequences for the dimers. Several details regarding the designed proteins must be considered. We are testing the idea of generating novel protein homodimers with hydrophobic dimer interfaces, using hydrophobicity as the main driving force for self-association. This procedure creates a hydrophobic patch on the surface of a soluble monomeric protein; however, this unshielded hydrophobic surface is destabilizing and usually results in molten globules and insoluble aggregates. Explicitly accounting for these unfavorable states through negative design is impractical for our system, due to the number of residues involved. Instead, we optimize the positive design under the premise that explicitly defining more favorable features for the target state will decrease the probability of population by an alternate state (Figure 4-2). By optimizing sequences for the dimer conformation, we design sequences that can follow an alternate folding pathway in the free-energy landscape. Instead of falling into the undesired states, these proteins could self-associate to minimize exposure of the hydrophobic surface area, forming a dimer consisting of two properly folded subunits.

A balance must be maintained between the polar and hydrophobic amino acids in the interface. We achieved this by tweaking a parameter used in our surface area based solvation term (σ_p, see below). Our solvation model is described by the equation:

$$E_{as} \approx \sigma_{np} A_{np,b} + \sigma_{np} A_{np,e} + \sigma_p A_{p,b} \qquad \text{Eq. 4-1}$$

where the atomic solvation energy (E_{as}) consists of a benefit term for burial of nonpolar amino acids ($\sigma_{np} A_{np,b}$) and penalty terms for polar amino acid burial ($\sigma_p A_{p,b}$) and nonpolar amino acid ($\sigma_{np} A_{np,e}$) exposure. Each of these terms contains a solvation energy factor (σ) and a surface area component (A), because atomic solvation is proportional to solvent-accessible surface area[6]. The value of σ_p sets the magnitude of penalties for burying polar atoms, and by adjusting this factor we can fine-tune the hydrophobicity of the interface (Table 4-1). Previously, we used a different type of solvation penalty to avoid burial of polar residues in the core; it penalized buried hydrogens that do not participate in hydrogen bonds. However, in initial dimer designs, penalizing polar hydrogens in the interface resulted in overpopulation of aromatic amino acids. We therefore decided to use σ_p to fine-tune hydrophobicity. In Table 4-1, the first interfacial design ID1, with a σ_p of 0.02, buries two aspartic acids (Asn10 and Asn13) without satisfying all the potential hydrogen bond donors and acceptors. We increased σ_p until the hydrophobicity of the interface improved. ID3 shows a reasonable balance between the polar and hydrophobic contents in the interface and was subsequently used as the sequence for the interface.

The sequences predicted for the interface, ID1, ID2 and ID3, shown in Table 4-1, were designed on the background of a thermostable variant of engrailed homeodomain, SC1[3]. The interfacial positions were selected based on their locations in the modeled dimer and the unbound monomer structures. Residues were first categorized into surface, boundary, and core classes for both structures, based on their C_α and C_β distances from the surface. Positions with classifications that switched from surface (in the monomer) to core or boundary (in the dimer) were considered interfacial positions. Design calculations were run to optimize the interface and core residues of the two subunits simultaneously in the dimer conformation. During the design process, amino acid and rotameric symmetry were imposed such that the equivalent positions on the two subunits would always remain identical. We found that the sequences obtained at the interface are sensitive only to the relative positions of the subunits and are indifferent to the sequences comprising the rest of the structure (referred to as background sequences). We obtained identical interfacial sequences with two different engrailed homeodomain backgrounds when the same docked dimer model was used (data not shown).

Because amino acids at the interfacial positions are mostly surface residues on the wild-type structure and do not have strong interactions with the protein core, the designed interface can be easily grafted to different structural backgrounds. For this purpose, side-chain placement calculations were performed to ensure the compatibility of the designed interface with the protein backgrounds. The stable backgrounds used were reported previously by Marshall et al., and these sequences show exceptional thermostability and NMR chemical shift dispersions[2,3]. Sequences derived from engrailed homeodomain mutants SC1 and NC3-Ncap were used in our design to host the designed binding interface (ID3). Two sequences, one derived from SC1 (C2G2) and the other from NC3-Ncap (C2G2CAP) were constructed and characterized experimentally (Figure 4-3).

C2G2 is derived directly from SC1 as a result of the interface design. The design calculation for constructing the interface included the ten positions in the interface as well as those classified as core, for a total of 19 positions; these were optimized under a different set of parameters from those used originally to generate the SC1 sequence. The sequence for the core is therefore modified in the context of the dimer interface. Four mutations from SC1, namely L7Y, L29V, I35L, L39Y, were suggested by ORBIT, and these are all mutations for core positions. Except for the modified core and the surface hydrophobic patch (the ID3 mutations), C2G2 shares the same surface and boundary sequences with SC1.

The C2G2CAP sequence was generated by grafting the ID3 design to NC3-Ncap. NC3-Ncap is also a stable and well-behaved engrailed homeodomain variant with a reported ΔG of unfolding of 5.6 kcal/mol. By grafting the interface sequence to a different host, we can test whether dimer formation is influenced by factors other than the interface sequence itself.

Assessment of the Designed Model

The most favorable design, ID3, has a protein core-like interface with hydrophobic amino acids at positions 13, 14, and 17 burying over ninety percent of their surface areas. Although the polar amino acid, threonine, at position 10 is 98.8% buried, it can potentially make a hydrogen bond with its symmetry related threonine from the other subunit to not only avoid energetic penalty but also to contribute to binding specificity (Figure 4-4). Among the favorable features found in this design, the phenylalanines at position 17 are the most striking; the exchange between the two phenylalanines from the two subunits create a knob-and-hole fit between the two molecules (Figure 4-5). The core of the interface is composed of six hydrophobic amino acids (three from each subunit), which form the basis of surface complementarity.

Two other favorable design features involve arginines forming salt-bridges and C-terminal capping hydrogen bonds across the dimer interface (Figure 4-6). Our designed hydrophobic interface is surrounded by three arginines (Arg9, Arg25, Arg32) and one aspartic acid residue (Figure 4-7). Arginine at position 25 forms two hydrogen bonds with the carbonyl oxygen of residue 16, capping the C-terminal of helix-1 from the binding partner. A cross-dimer salt-bridge is found between an arginine and an aspartic acid at position 32 and 12 respectively. Electrostatics and helix-dipole driven interactions like these are highly specific and in the case of C-terminal capping interactions can be energetically favorable and stabilizing[7].

Characterization of the Designed Proteins

To test the validity of the design, sequences C2G2 and C2G2CAP were synthesized, expressed, purified and characterized experimentally. C2G2 is expressed as inclusion bodies in the *E. coli* cytoplasm, and its solubility is highly dependent upon pH (Figure 4-8) while C2G2CAP is readily soluble in aqueous buffers. At near neutral pH, the solubility of refolded C2G2 is less than 30 μM.

Far-ultraviolet (UV) circular dichroism (CD) was used to measure the secondary structure content of the two designed variants, and the characteristic minima at 208 nm and 222 nm of an engrailed homeodomain were preserved, suggesting that C2G2 and C2G2CAP maintain the helical fold (Figure 4-9). However, the thermal denaturation curves for the two proteins (Figure 4-9B and Figure 4-9D) show that the two proteins have different stabilities and thermodynamic characteristics. In Figure 4-9B, the thermal denaturation curve for C2G2 shows two transitions. Comparison of the curve in Figure 4-9B with Figure 4-9A, in which the far-UV wavelength scan is monitored over the entire course of the thermal denaturation, reveals that the helical structure disappears during the first transition, and the signal becomes more β-sheet-like. The first derivative of the melting signal suggests an inflection point for the first transition at approximately 39 °C. The β-sheet-like CD signals at high temperatures are not from irreversible amyloid aggregates because the thermal denaturation curve can be reproduced from the same transition is unknown, we cautiously follow the thermal denaturation curve and perform all other analyses on C2G2 at 10 °C.

For C2G2CAP, on the other hand, even at 99 °C the protein does not lose all of its secondary structure, as illustrated in Figure 9C; the minima at 208 and 222 nm are retained over the course of the experiment. Due to the lack of a well-defined post-transition, the melting temperature (T_m) is approximated by taking the first derivative of the data; it is about 70 °C. Analyses of C2G2CAP were carried out at room temperature (25 °C).

It should be noted that mutations were introduced to C2G2's core by using σ_p instead of polar hydrogen burial penalties in the solvation parameters; therefore, the drastic loss of stability compared to its background variant, SC1, is likely due to changes in the protein core and not the hydrophobic interface. This idea is supported by the fact that C2G2CAP, whose core remains the same as that of NC3-Ncap, largely retains its thermostability; the small interaction energies observed between the interface and the core also suggest that the interface is not interfering with core packing.

Despite being thermostable, C2G2CAP shows dynamic characteristics when examined by NMR spectroscopy. C2G2CAP's 1D proton line shapes are broadened and its 2D N15-proton HSQC shows only a fraction of the peaks (Figure 10). These results suggest that many factors can contribute to conformational heterogeneity, and one of the possibilities is self-association. The broadened line widths and missing peaks, however, hinder structure determination by NMR; therefore, X-ray crystallography was used instead.

Self-association Properties of the Designed Proteins

Our designed proteins should prefer a dimeric assembly in solution as they were created by energetically optimizing the interface for a dimeric conformation. Sedimentation experiments were carried out to examine the hydrodynamic properties, oligomeric states, and in the case of self-assembly, binding constants of the two designs. Sedimentation velocity data modeled by finite element solutions to the Lamm equation with maximum entropy regularization provide information for the size distribution of the sample. We use the program SEDFIT, written by Schuck, for this analysis[8]. Due to the small molecular weight of the proteins, sedimentation experiments were carried out at relatively high rotor speeds. For velocity sedimentation monitored by absorbance optics, the data were collected at rotor speeds of 48,000 rpm, and for those monitored by Rayleigh interference optics, rotor speeds of 60,000 rpm were used.

In Figure 4-11, the continuous distribution of sedimentation coefficients, c(s), is shown for both C2G2 and C2G2CAP with a monomeric engrailed homeodomain variant (NC0) as a control. Because the s-values reported for the weight averaged monomers and dimers are independent from changes in loading concentration, the monomer and the dimer are in slow equilibrium; the time-scale for monomer-dimer exchange is slow compared to their sedimentation rates. Although the c(s) model should only be used for discrete non-interacting species, due to the slow equilibrium observed, the monomer and the dimer can thus be fitted as discrete species.

Reported in Figure 4-11 for the experimental conditions at 10 °C, the peaks at 0.8 Svedberg (S) correspond to monomeric engrailed homeodomain species, and the existence of peaks at around 1.2 S (shown in Figure 4-11 for C2G2CAP as an unresolved peak shoulder) indicates that there are dimers. The dimer species were further confirmed by fitting the data with the continuous molar mass distribution model, c(M), suggesting that the species with the higher molecular weight is twice the mass of the monomer. The unresolved second peak for C2G2CAP is the result of the equilibrium between the monomer and the dimer, because the dimer species is clearly resolved when using interference optics to monitor the sedimentation time course of C2G2CAP at a higher loading concentration (Figure 4-12). The increase in loading concentration shifts the equilibrium to favor dimer formation. In Figure 4-12, the ratios of the monomer, dimer, and higher order oligomer concentrations to the total loading concentration are 27%, 72% and 1%, respectively; fitting s-value ranges from 1 S to 8 S includes all observed species in the sample.

The size-distribution results obtained from sedimentation velocity experiments indicate successes in designing self-associating dimers. For both C2G2 and C2G2CAP, the AUC experiments provide strong evidence for dimer formation from purified monomers. Based on the loading concentrations in the sedimentation velocity experiments and the resulting monomer-dimer distributions, it can also be concluded that C2G2 forms tighter binding dimers than those of C2G2CAP since a higher loading concentration is required in order to detect C2G2CAP dimers.

We further characterized the designs with sedimentation equilibrium experiments in order to determine the oligomeric states and the association/dissociation constants. Sedimentation equilibrium data were collected from three loading concentrations at three speeds. The concentration gradients established at equilibrium for C2G2 can be described by non-linear least squares fitting with monomer and dimer components, and the dissociation constant (Kd) is determined as 12.2 μM (Figure 4-13 and Table 4-2).

For C2G2CAP, however, the data fit poorly to either a single-state or a two-state model, but fit to the monomer-dimer-tetramer (M-D-Tet) model. Fitting the data with the M-D-Tet model is plausible because when floating the reduced buoyant molecular weight (σ) during the fitting process, the fitted σ value remains close to the theoretical σ for C2G2CAP. Although the overall fit of the equilibrium sedimentation data is good with the M-D-Tet model, the concentrations of the component monomer, dimer, and tetramer species extrapolated from the fit disagree with the distribution obtained from sedimentation velocity experiments. Specifically, in Figure 4-12, there are significant amounts of the monomer and dimer in the c(s) distribution (27% monomer, 72% dimer, 1% oligomer, mentioned previously), but this is not reflected in the M-D-Tet fit. In the M-D-Tet equilibrium fit, the tetramer species dominates over the monomer species, and this is not observed in the c(s) distributions. The M-D-Tri model, however, captures the concentration profiles for each of the components, but it does not fit the data as well as the M-D-Tet model. The M-D-Tri model fits are shown in Figure 4-14 with the M-D-Tet model for comparison. The difference in the size-distributions observed in the equilibrium and velocity experiments is significant, but in either the M-D-Tet or the c(s) distribution fits, the dimer remains the dominant species (Figure 4-14 and Table 4-3). Further tests are needed to precisely determine the cause of the differences in the AUC experiments; our current knowledge of the Kd is shown in Table 4-3.

Initial X-ray Diffraction Results

The X-ray diffraction data was collected at the SSRL synchrotron to 1.8 Å resolution, and the diffraction pattern of one of the frames is shown in Figure 4-15. The peaks were indexed with DENZO and SCALEPACK[9], suggesting that C2G2CAP is in the space group R32. There is one molecule in each of the asymmetric units, and the initial molecular replacement solutions indicate that C2G2CAP proteins are in C3 arrangement as trimers (Figure 4-16). But due to the relatively high R_{free} (47%), we will need to refine the structure further or obtain experimental phases to finalize the structure.

DISCUSSION

The specialized docking algorithm developed to model C2 symmetry related protein dimers from monomeric proteins was used to generate the dimeric backbone model for protein design. One of the resulting dimer conformations from docking the crystal structure backbone of the engrailed homeodomain was used because of its high surface complementary and properly positioned helices. The helices from the two subunits stack to form the interface. Although the helices are in parallel arrangements, there is no resemblance to known coiled-coils. Taking the docked dimer structure as the scaffold for protein design, the ORBIT protein design programs were used to design the binding interface using hydrophobicity as the main driving force for self-association. We generated and characterized two designs experimentally, and our designed proteins self-assemble to form dimers.

Designing interfaces using model backbones generated by an approximated docking method has not been attempted before, and in order to achieve the design objectives, several important assumptions must be made (Figure 4-2). We have no knowledge of other competing states; also, we do not know *a priori* whether a sequence exists in sequence space that could adopt the docked dimer conformation. Furthermore, certain published methods to deal with negative designs are not applicable due to the relatively large number of positions involved[10,11]. By carefully selecting our design backgrounds and interface design parameters, we were able to design self-associating dimers by optimizing only the positive design states. The significantly reduced solubility and stability of C2G2, compared to its background protein, SC1, could potentially be due to our failure to capture the aggregated state, since no negative designs are explicitly considered. This should be further investigated by grafting the sequence for the binding interface to SC1 without including the modified protein core. The results from such a study would indicate whether interfaces should be treated differently from protein cores and would help determine if the use of σ_p throughout the design positions is appropriate. Our success with C2G2CAP is indicated by AUC experiments, but a detailed assessment of the designed interface still requires confirmation from X-ray structures.

The disagreement between the velocity and equilibrium AUC experiments should be addressed. Obtaining Kds from sedimentation equilibrium experiments is more robust than integrating curves from velocity experiments, but fitting equilibrium data can have difficulties when multiple equilibrium states are considered. The size-distribution observed from the sedimentation velocity experiments suggests that C2G2CAP forms mostly dimers, and this is in agreement with the sedimentation equilibrium results. Non-linear least squares fitting of the equilibrium exponentials in our case may not have sufficient information in discriminating the relative concentrations of the monomer and the tetramer, but despite the difficulties in analyzing the AUC experiments, the dimer is the dominant species in the equilibrium.

Conclusion

We report the first computationally designed *de novo* protein homodimers using a combination of protein docking and protein design tools. The greatest uncertainty in such an undertaking is the quality of the dimer model, because the rotamer selection criterion in the protein design algorithm is very sensitive to changes in backbones. In spite of the approximate nature of our docking algorithm in using reduced backbones for docking, our protein design algorithm was able to predict a sequence that is plausible for the particular dimer conformation. Both proteins with the designed interface sequence self-assemble into dimers, each with a different affinity that is, in the first order approximation, inversely correlated with their thermal stabilities. The strategy of combining docking and design algorithms for generating novel oligomers should be applicable to a broader range of structures. Even if the dimer conformation is not completely novel, the docking procedure facilitates exploration around the local binding site for alternative conformations. We have described a general method for creating novel dimers.

MATERIALS AND METHODS

Generation of Dimer Models

The specialized docking algorithm for generating C2 related homodimers by docking the backbone of a monomeric protein to itself, and the parameterization of the reduced backbone representation used in the docking process are described in detail in Chapter 2. The procedure for generating the engrailed homeodomain dimer is described, in addition to the description given in the Results section, as follows. The arrays used for the docking calculation were 64 by 128 by 128 in length for each of the x, y and z dimensions. Each element of the arrays corresponds to a 1 Å^3 cube in space. The backbone of the wild-type engrailed homeodomain crystal structure (PDB code: 1enh) was first centered in the space represented by the 3D grids (the initial state), duplicated to generate the coordinates of the binding partner, which was in turn rotated 180° about the x-axis, followed by docking using FFT correlations. After each round of translational evaluation, the docking program returns to the initial state, and the centered backbone coordinates are rotated by 1° increments about the y- and the z-axes (following the rules described in Chapter 2 for docking C2 related dimmers) before a subsequent round of coordinate duplication and translational evaluation take place. The top 20 correlations were stored for each rotational position and the final list of all correlations were ranked. The top 200 conformations generated from the ranked list were clustered based on their structural similarities to 11 distinct dimer conformations, and each of the dimers were visually inspected to pick out the final candidate, which in this case, is the conformation with the second highest correlation score.

Computational Protein Design

The positions on the docked dimer were categorized into core, boundary and surface classes. When compared with the wild-type classifications, the positions with classifications that switched from surface to core, surface to boundary, or boundary to core in the context of the dimer conformation were included in the design; this covers positions on either subunit. For all of the design positions except for position 13 and 29, the following types of amino acids were allowed: Ala, Val, Leu, Ile, Phe, Tyr, Trp, Asp, Asn, Glu, Gln, Lys, Ser, Thr, His, Arg. Position13 and 29 allow everything listed above but Glu and Lys. The rotamer conformations on the equivalent positions for each of the subunits were linked, so not only was the symmetry in amino acid identity reserved but also the exact rotamer for each of the equivalent positions. For positions not included in the design calculations, their wild-type conformations were left unaltered.

The following lists all the positions designed as "core" positions: 10, 13, 14, 17, 25, 32, 7, 11, 29, 33, 35, 39, 40, 43, 44. Those designed as "boundary" positions are: 9, 16, 18, 28. The difference between designing positions as "core" and as "boundary" is in the multiplication factors used for the different residue classes when computing their solvent-accessible surface areas[12].

The rotamer library used for the interface design is based on the backbone dependent library by Dunbrack and Karplus[13]. We used rotamers that cover the discrete conformations reported by Dunbrack and Karplus and their $\chi 1$ and $\chi 2$ dihedral angle expansions for all amino acid types except for Gln, Glu, Arg, and Lys – no expansions were used for these amino acids (it is called the e2QERK0 library).

The rotamer optimization procedure for the designed positions involves using the HERO and FASTER algorithms in sequential order[14,15]. The pairwise rotameric interaction energy matrix was first reduced by HERO until HERO stalled without converging, and the remaining portion was optimized by FASTER to completion.

Gene Synthesis

The DNA primer sequences used to construct C2G2 and C2G2CAP were designed using the PRImer DEsign (PRIDE) program written by Po-Ssu Huang, which automatically produces primer sequences for recursive PCR total gene synthesis purposes[16]. Each of the genes was constructed from four overlapping DNA oligos. The synthetic DNA oligos were purchased from Sigma-Genosys and Integrated DNA Technologies, Inc. The sequences of the constructed genes were verified by DNA sequencing performed by the Caltech DNA sequencing facility. The genes were cloned into BL21(DE3) host strain for IPTG induced expression.

Expression and Purification

The proteins were expressed by IPTG induction at the late log phase of growth for 4 hours at 37 °C in LB media. The C2G2 protein was expressed in the inclusion bodies from the BL21(DE3) hosts (Stratagene), and was extracted from the cells by using the EmulsiFlex-C5 homogenizer (Avestin) followed by solubilization by the addition of 8M guanidine hydrochloride (Sigma-Aldrich). After extensive centrifugation and filtration, the solubilized C2G2 in guanidine hydrochloride was purified directly with reverse-phase

HPLC (C8) using a linear 0.5 % min^{-1} 2%/95% blend acetonitrile gradient in the presence of 0.1% Trifluoroacetic acid (TFA) as the pairing ion.

C2G2CAP was expressed in the soluble portion of the cell extract, and was isolated by the freeze-thaw method[17] between dry ice ethanol slurry and 42 °C water bath. The same 0.5 % min^{-1} gradient was used for purification on the reverse-phase HPLC C8 column (Vydac). For the 2D NMR HSQC studies, C2G2CAP was expressed in standard M9 minimal media using 15N-ammonium sulfate (2g/L).

The masses of the purified C2G2 and C2G2CAP were confirmed by mass spectrometry using the matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) method, and the purity of the samples were verified by standard SDS-PAGE. The protein concentrations were determined by measuring the optical density of guanidine chloride denatured sample at 280 nm, using extinction coefficients determined from Tyr and Trp contents. The extinction coefficients for C2G2 and C2G2CAP at 280 nm are 13520 and 11120 (M∗cm)$^{-1}$ respectively.

Refolding of C2G2

Lyophilized C2G2 was refolded by solubilization in 8M guanidine hydrochloride, followed by buffer exchange. The refolding conditions for C2G2 were based on the FoldIt screen (Hampton Research). The solubilized C2G2 in guanidine hydrochloride was first rapidly diluted 50 to 100 fold into the stabilization buffer containing 55 mM MES, 220 mM sucrose, 275 mM L-Arginine, 150 mM NaCl and 440 mM KCl, followed by dialyzing against the final refolding buffer containing 55 mM MES, 150 mM NaCl and 440 mM KCl, until at least one million fold dilution of the L-Argining and sucrose components was achieved. The refolded sample was concentrated by putting dry PEG 800 over the dialysis tubing to draw the buffer out by osmosis. For CD and AUC experiments, NAP-25 columns were used to exchange the buffer to 50mM sodium phosphate, 150 mM sodium chloride, pH 6.2.

Circular Dichroism

The circular dichroism data for measuring the secondary structure contents and thermal stabilities of the proteins were collected on an Aviv 62A DS spectropolarimeter. The C2G2 thermal melting curve shown in Figure 9 was collected at a protein concentration of approximately 12 μM in a 0.1 cm pathlength cuvette using 2 °C increment from 1 °C to 99 °C. The C2G2CAP curve was collected with a protein concentration of approximately 16 μM in a 1 cm pathlength cuvette, using also 2 °C temperature steps. At each temperature the samples were equilibrated for two minutes before the wavelength scans between 200 and 250 nm were collected. The signal averaging time at each wavelength was one second. The T_m was defined as the temperature where the maximum of the first derivative of the signal occurs. For all of the CD experiments, 50mM sodium phosphate and 150 mM NaCl buffer at pH 6.2 was used.

NMR experiments

NMR spectra were collected at 4, 10, or 25 °C on a Varian UnityPlus 600 Mhz spectrometer. Protein concentrations were ~12 μM for C2G2 1D spectra and ~750 μM for C2G2CAP experiments. Buffer used for these experiments contains 50 mM sodium phosphate, pH 6.2.

Analytical Ultracentrifugation

Protein samples were analyzed on an XL-I analytical ultracentrifuge equipped with an AnTi60 rotor (Beckman Coulter). For sedimentation velocity experiments, two-channel epon-filled centerpieces were used, and the cells were torqued to 130 lb-inch for experiments running at rotor speeds of 60,000 rpm. For C2G2 velocity experiments, the data were acquired at 230 nm, 10 °C, in continuous scanning mode and at rotor speeds of 48,000 rpm. For C2G2CAP velocity experiments, the data were acquired at 10 or 25 °C using either absorbance optics at 230, 250 or 280 nm or Rayleigh interference optics in continuous scanning mode (in the cases where interference optics were used, data were collected at 5 minute intervals) and at rotor speeds of 48,000 or 60,000 rpm. The sedimentation boundaries were fitted to the continuous distribution of sedimentation coefficient, c(s), model using SEDFIT[8]. Data were fitting using 100 sedimentation coefficient increments in the range of 0.1 to 5 S for 10 °C runs and 1 to 8 S for 25 °C runs. Time invariant noises and baseline offsets were also corrected. Maximum entropy regularization confidence level of 0.95 was used in all of the size-distribution analysis.

Sedimentation equilibrium experiments were carried out in 6 channel epon-filled centerpieces. For C2G2, the sedimentation equilibrium experiments were carried out using three loading concentrations (13 μM, 6.5 μM, 3.25 μM) at three rotor speeds (30,000, 35,000 and 48,000 rpm) monitored at 230 nm using absorbance optics. For C2G2CAP, the equilibrium sedimentation data were collected using also three concentrations (0.6 mM, 0.2 mM, 0.066 mM) at three rotor speeds (48,000, 54,000 and 60,000 rpm) monitored with Rayleigh interference optics. We subtracted both the machine blanks (during data collection) and water blanks from the interference data before fitting. Sedimentation equilibrium data were fitted with the program WinNONLIN (v1.06), and the fitting statistics were listed in Table 4-2 and Table 4-3. The theoretical reduced buoyant molar masses, σ, were calculated using the SEDNTERP (v1.08) program.

Crystallization and Data Collection

C2G2CAP crystals were grown under the following buffer conditions: 0.2 M calcium chloride dihydrate, 0.1 M HEPES – Na (pH 7.5), 28% v/v Polyethylene Glycol 400 at 4 °C. Mixing 2 μl of C2G2CAP in 10 mM Tris and 100 mM NaCl at pH 7.6 with 2 μl of the crystallization buffer above in sitting drops forms microcrystals initially, but large crystals will eventually grow from the drops. Crystals of space group R32 (a = 69.242 Å, b = 69.242 Å, c = 69.222 Å) was obtained and was flash frozen in liquid nitrogen before mounting on X-ray diffractometers. Native crystal data were collected to 1.8 Å resolution at the SSRL synchrotron (40556 collected reflections, 10930 unique reflections, 94.9 % complete; R_{merge} = 4.7). Diffraction data were analyzed using DENZO and SCALEPACK[9].

References

1. Katchalskikatzir, E. et al. Molecular-Surface Recognition - Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 2195-2199 (1992).
2. Marshall, S. A., Morgan, C. S. & Mayo, S. L. Electrostatics significantly affect the stability of designed homeodomain variants. *Journal of Molecular Biology* **316**, 189-199 (2002).
3. Marshall, S. A. & Mayo, S. L. Achieving stability and conformational specificity in designed proteins via binary patterning. *Journal of Molecular Biology* **305**, 619-631 (2001).
4. Oshea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. X-Ray Structure of the Gcn4 Leucine Zipper, a 2-Stranded, Parallel Coiled Coil. *Science* **254**, 539-544 (1991).
5. Harbury, P. B., Kim, P. S. & Alber, T. Crystal-Structure of an Isoleucine-Zipper Trimer. *Nature* **371**, 80-83 (1994).
6. Eisenberg, D. & Mclachlan, A. D. Solvation Energy in Protein Folding and Binding. *Nature* **319**, 199-203 (1986).
7. Strop, P. & Mayo, S. L. Contribution of surface salt bridges to protein stability. *Biochemistry* **39**, 1251-1255 (2000).
8. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. *Biophysical Journal* **78**, 1606-1619 (2000).

9. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Macromolecular Crystallography, Pt A* **276**, 307-326 (1997).

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

11. Bolon, D. N., Wah, D. A., Hersch, G. L., Baker, T. A. & Sauer, R. T. Bivalent tethering of SspB to ClpXP is required for efficient substrate delivery: A protein-design study. *Molecular Cell* **13**, 443-449 (2004).

12. Street, A. G. & Mayo, S. L. Pairwise calculation of protein solvent-accessible surface areas. *Folding & Design* **3**, 253-258 (1998).

13. Dunbrack, R. L. & Karplus, M. Backbone-Dependent Rotamer Library for Proteins - Application to Side-Chain Prediction. *Journal of Molecular Biology* **230**, 543-574 (1993).

14. Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. Exact rotamer optimization for protein design. *Journal of Computational Chemistry* **24**, 232-243 (2003).

15. Desmet, J., Spriet, J. & Lasters, I. Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER) as a new method for protein structure optimization. *Proteins-Structure Function and Genetics* **48**, 31-43 (2002).

16. Prodromou, C. & Pearl, L. H. Recursive Pcr - a Novel Technique for Total Gene Synthesis. *Protein Engineering* **5**, 827-829 (1992).

17. Johnson, B. H. & Hecht, M. H. Recombinant Proteins Can Be Isolated from Escherichia-Coli-Cells by Repeated Cycles of Freezing and Thawing. *Bio-Technology* **12**, 1357-1360 (1994).

Table 4-1. Design of The Dimer Interface

	σ_P	Designed Positions[a]									
		9	10	13	14	16	17	18	25	28	32
ID1	0.02	R	N	N	L	W	F	D	R	W	S
ID2	0.03	-	-	-	-	-	-	-	-	-	-
ID3*	0.04	-	T	L	-	L	-	-	-	-	R

[a] The postions whose residue classifications change from exposed to buried (surface to core) and from exposed to partially buried (surface to boundary) are considered interfacial postions and are redesigned.

* The interface used in the experimentally characterized dimers

Table 4-2. C2G2 Sedimentation Equilibrium Analysis

Fix/Float				Exp. Molar		95% conf.			95% conf.	
σ[a]	Model[b]	SRV[c]	σ[a]	Mass (Da)[d]	LnK2	Interval		Kd (μM)[e]	Interval	
Float	M-D	3.21E-03	0.7105	6174.96	0.8140	0.4972	1.123	12.22	16.77	8.97
Fix	M-D	3.23E-03	0.7628		0.3179	0.0140	0.6218	20.07	27.20	14.81

[a] Reduced buoyant molecular weight.

[b] M-D refers to the monomer-dimer equilibrium model.

[c] Squre root of variance (or RMSD) of the fit.

[d] Experimental molar mass determined from σ by the SEDNTERP program for data taken at 10 °C, 30,000 rpm.

[e] Dissociation constant determined from LnK2.

Table 4-3. C2G2CAP Sedimentation Equilibrium Analysis

Fix/Float σ[a]	Model[b]	SRV[c]	σ[a]	LnK2	95% conf. interval		Kd (μM)[d]	95% conf. interval	
Fix	M-D-Tet	1.77E-02	1.832	1.13	0.96	1.30	30	35	25
Float	M-D-Tet	1.77E-02	1.733	1.53	1.34	1.72	20	24	16
Fix	M-D-Tri	2.26E-02	1.832	-0.46	-0.66	-0.27	145	177	120
Float	M-D-Tri	1.76E-02	2.532	-1.44	-1.59	-1.27	384	449	327

Table 4-3 (continued)

Fix/Float σ[a]	Model[b]	LnK3	95% conf. interval		LnK4	95% conf. interval	
Fix	M-D-Tet				-0.39	-0.64	-0.13
Float	M-D-Tet				0.68	0.18	1.17
Fix	M-D-Tri	-0.56	-0.69	-0.43			
Float	M-D-Tri	-4.00	-4.28	-3.72			

[a] Reduced buoyant molecular weight.

[b] M-D refers to the monomer-dimer equilibrium model.

[c] Squre root of variance (or RMSD) of the fit.

[d] Dissociation constant determined from LnK2.

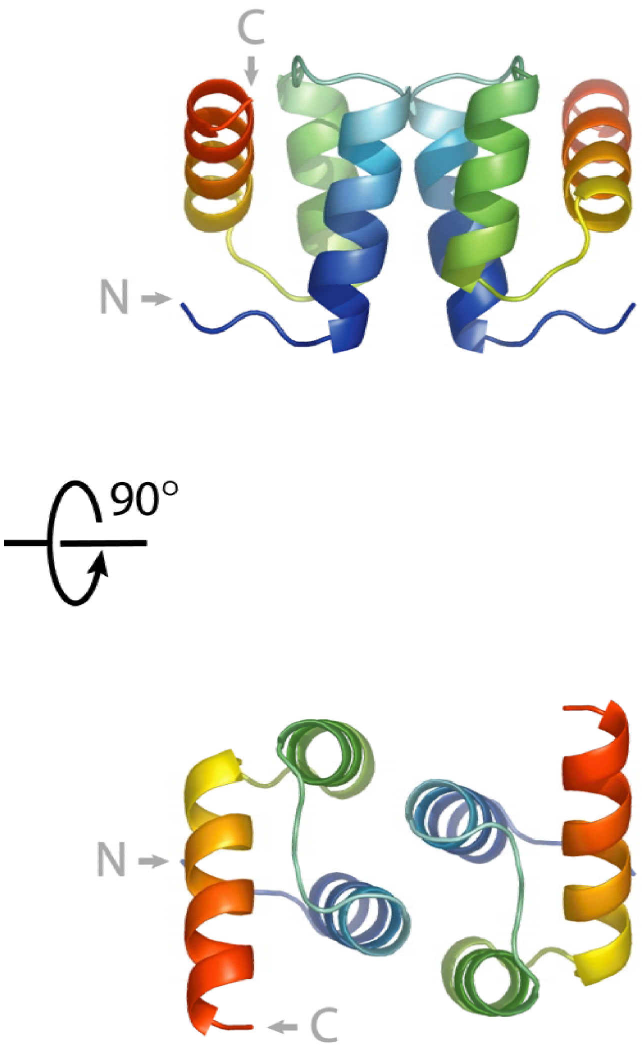


Figure 4-1. The engrailed homeodomain structure based dimer. The dimer shown is generated by docking engrailed homeodomain structure (PDB code: 1enh) to itself with C2 symmetry. The dimer interface is formed between the four helices, helix-1 and helix-2 from each of the subunits.

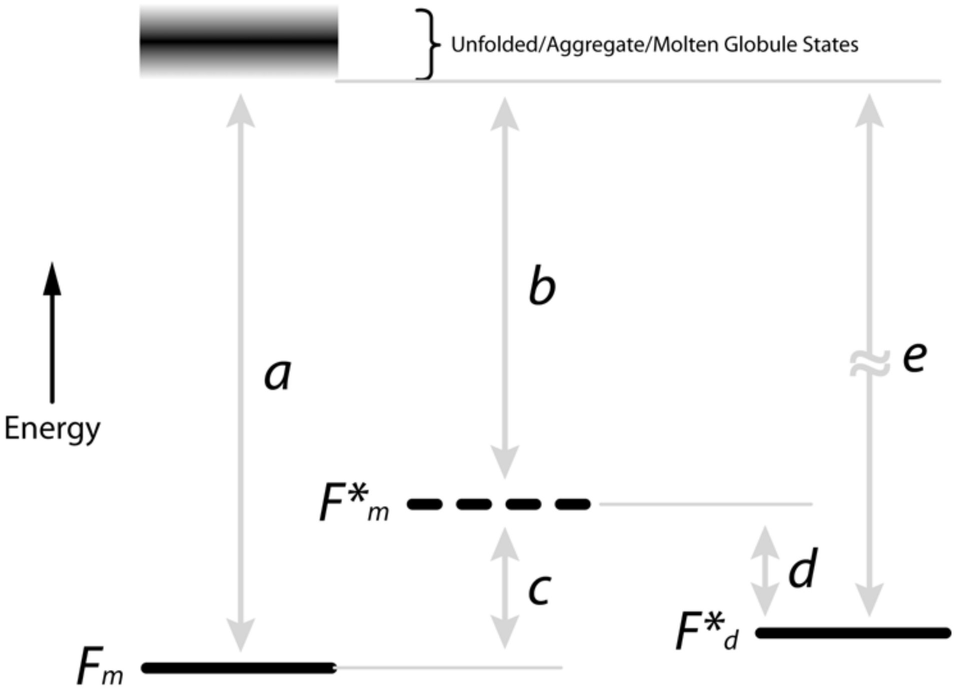
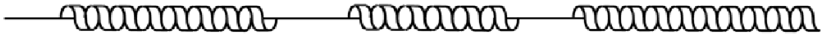


Figure 4-2. Energy diagram for comparison between different associative states. States indicated with an "*" have hydrophobic patches. F_m is the folded monomer. F^*_m is the folded but destabilized monomer with a surface hydrophobic patch. F^*_d is the folded dimer with a designed hydrophobic interface. The relative energies between these states and the unfolded/aggregate/molten globule states are illustrated with energies *a*, *b*, *c*, *d*, and *e*. Solid lines indicate those states that can be modeled explicitly, and a dash line is shown for the state that is not explicitly considered. In our design model, the hydrophobic patches are designed on the background of a stable monomeric variant, and thus *a* is sufficiently deep to accomodate the destablizing effects introduced by the surface patch. Therefore, $a > c$. We avoid the undesirable states under this assumption. By introducing a modeled F^*_d state, where $e > b$ since the exposed hydrophobic surface areas are buried, we optimize *e* to indirectly maximize *d*. This strategy drives dimer formation by positively designing both *a* and *e* and assumes that the destabilization of F^*_m caused by the exposed hydrophobic patch indirectly optimizes *d*.

```
                       1         2         3         4         5
              ----|----0----|----0----|----0----|----0----|----0-
SC1           TKFDEQLKRRLEEEFKRDRRLTNQRRHDLSQKLGINEELIEDWFRRKEQQI
C2G2          TKFDEQYKRTLELLFLFDRRLTNQRRHWVSQRLGLNEEYIEDWFRRKEQQI
NC3-Ncap      TEFSEEQKRRLDEEFRRDRRLTEERRRDLSQKLGLNEEQIERWFRRKEQQI
C2G2CAP       TEFSEEQKRTLDLLFLFDRRLTEERRRWLSQRLGLNEEQIERWFRRKEQQI
```



```
Residue class bsbssscssbcssbcsssbsbsssbbsscbsscscssscc ssccssbsssb
```

Figure 4-3. Sequence alignment for SC1, C2G2, NC3-Ncap and C2G2CAP. The locations of the three helices in the engriled homeodomain structure are shown below the C2G2CAP sequence. Residue classifications as either core (c), boundary (b) or surface (s) are shown at the bottom**.**

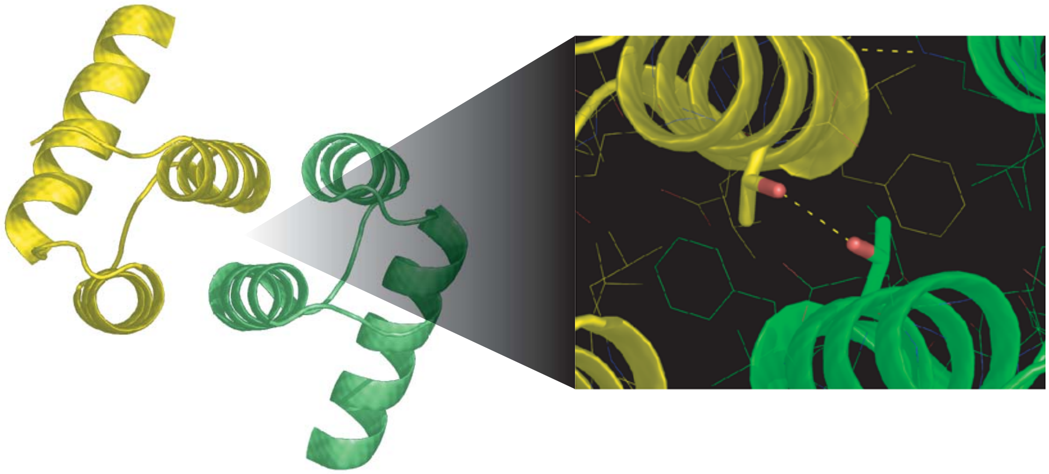


Figure 4-4. Thr 10 burial. The 2-fold symmetry related threonines (Thr10) in the center of the dimer interface forms a buried hydrogen bond.

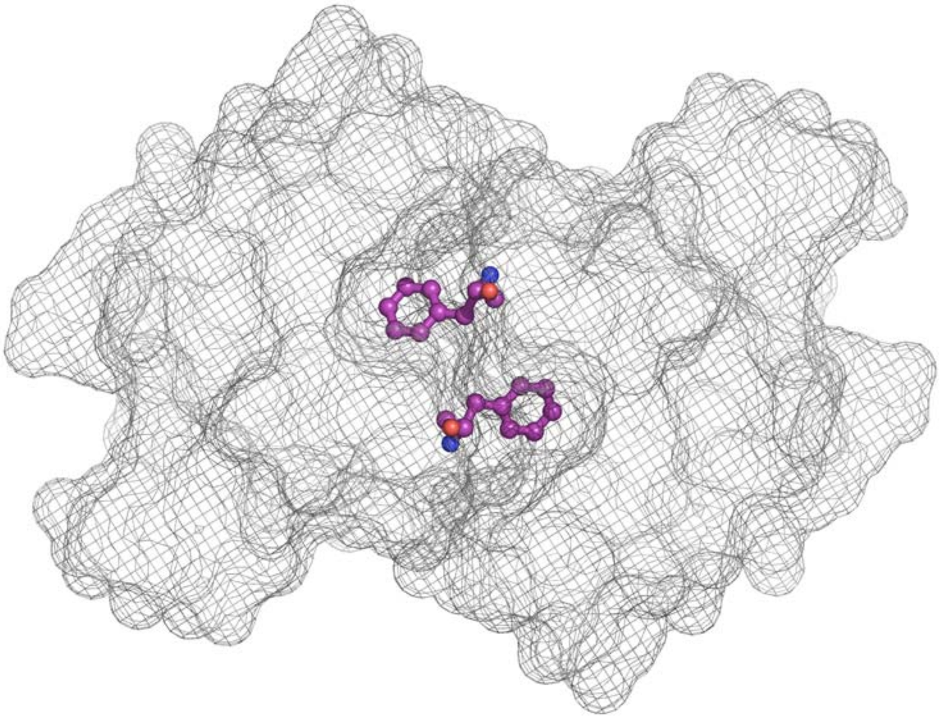


Figure 4-5. The phenylalanine pocket. The exchange of phenylalanines at the interface between the subunits creates a highly complementary hydrophobic core.

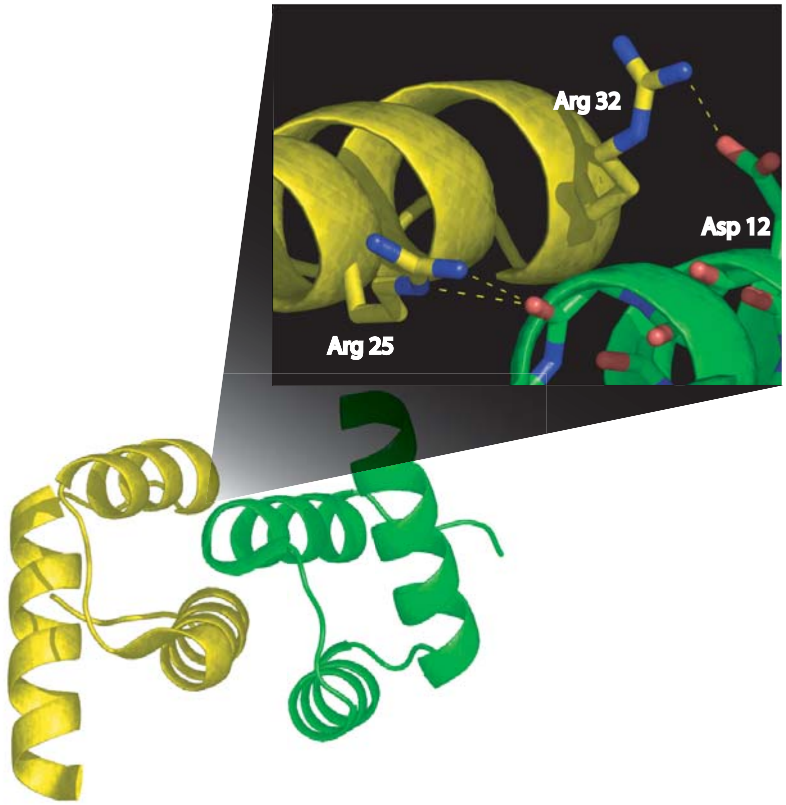


Figure 4-6. C-terminal capping and salt bridges. Arginines (Arg 25) from each member molecule form C-terminal helical caps across the dimer interface. There are also two pairs of surface salt bridges between the arginines and aspartic acids (Arg 32 and Asp 12). Only one is shown for each interaction type.

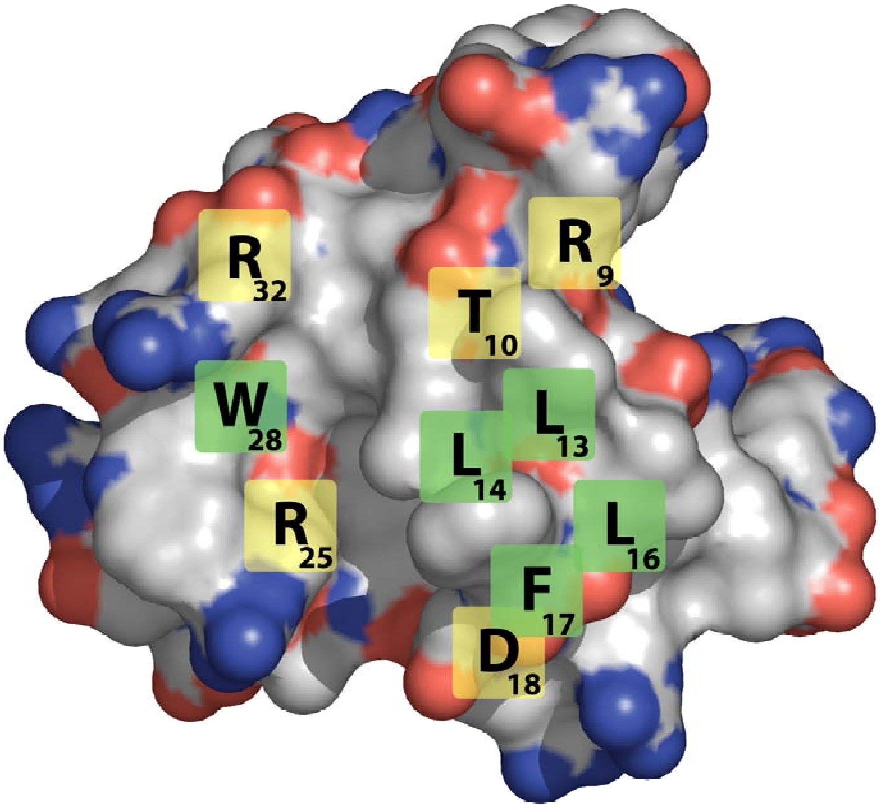


Figure 4-7. The designed interface. Positions with polar residues are highlighted in yellow, and those with hydrophobic residues in green. The hydrophobic residues form the core of the interface and are surrounded by polar residues shielding the core from solvent.

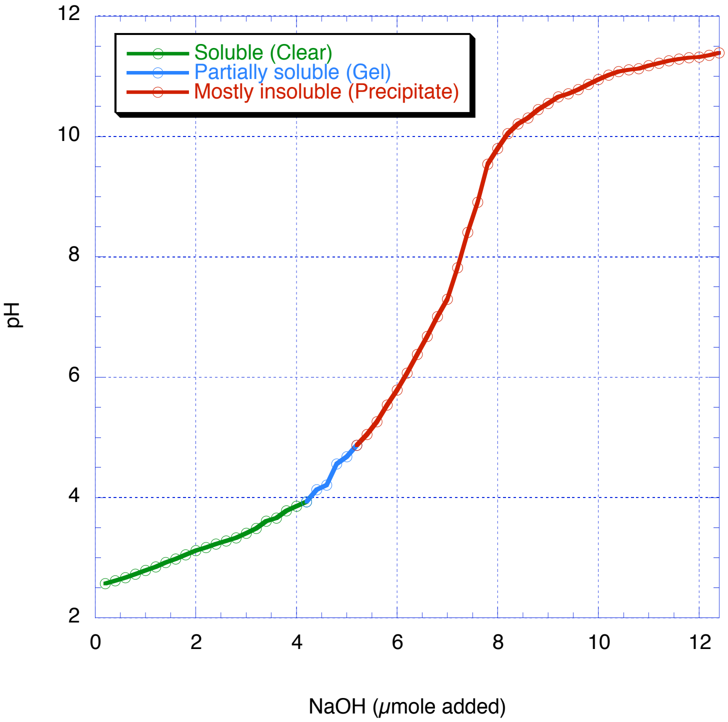


Figure 4-8. Base titration of C2G2. The solubility of C2G2 at different pHs is measured by titrating NaOH. At 1 mM protein concentraion, C2G2 is highly soluble at low pH (pH 2.6 to 3.8). At pH around 3.9 to 4.8, C2G2 is partially soluble and forms a clear but viscous gel. At pH beyond 4.8, the solubility is very low, with most of the proteins precipitated out of the solution.

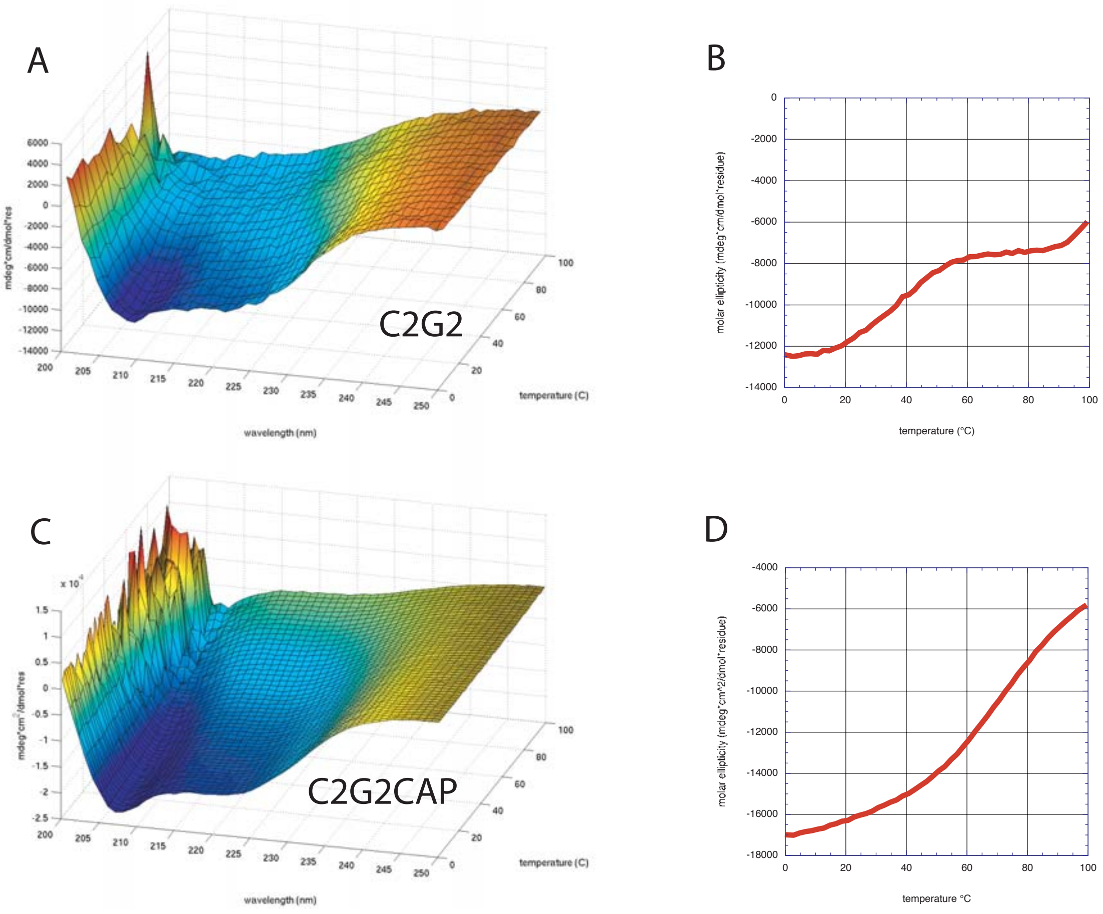


Figure 4-9. Thermal denaturation of C2G2 and C2G2CAP. A. CD wavelength scans of C2G2 as a function of temperature. **B.** Thermal denaturation of C2G2 monitored by CD at 222 nm. **C.** CD wavelength scans of C2G2CAP as a function of temperature. **D**. Thermal denaturation of C2G2CAP monitored by CD at 222 nm.

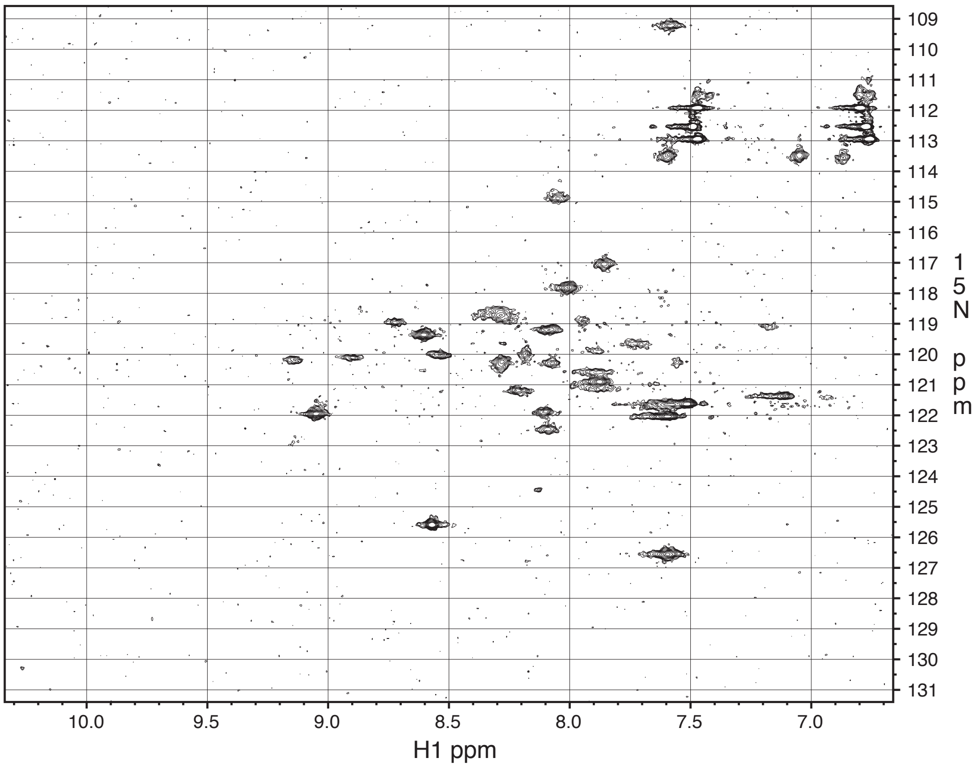


Figure 4-10. N15-H1 HSQC spectrum of C2G2CAP. There are 51 residues in C2G2CAP but only about 40 peaks are shown. This is likely due to dimer formation since the folding of the molecule is confirmed by X-ray crystallography.

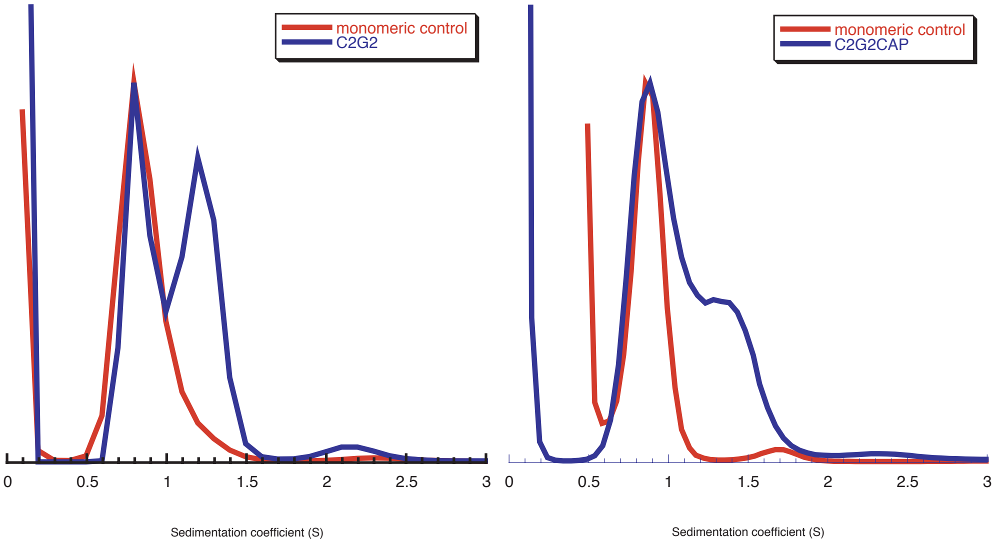


Figure 4-11. The c(s) distribution of C2G2 and C2G2CAP at 10 °C. The size-distribution analysis of C2G2 and C2G2CAP shows a peak at ~0.8 S corresponding to the monomer and a higher molecular weight species sedimenting at ~1.2 S. The relative molar mass of the two peaks is confirmed by fitting the data to c(M) model suggesting a 2:1 ratio. This is strong evidence of dimer formation.

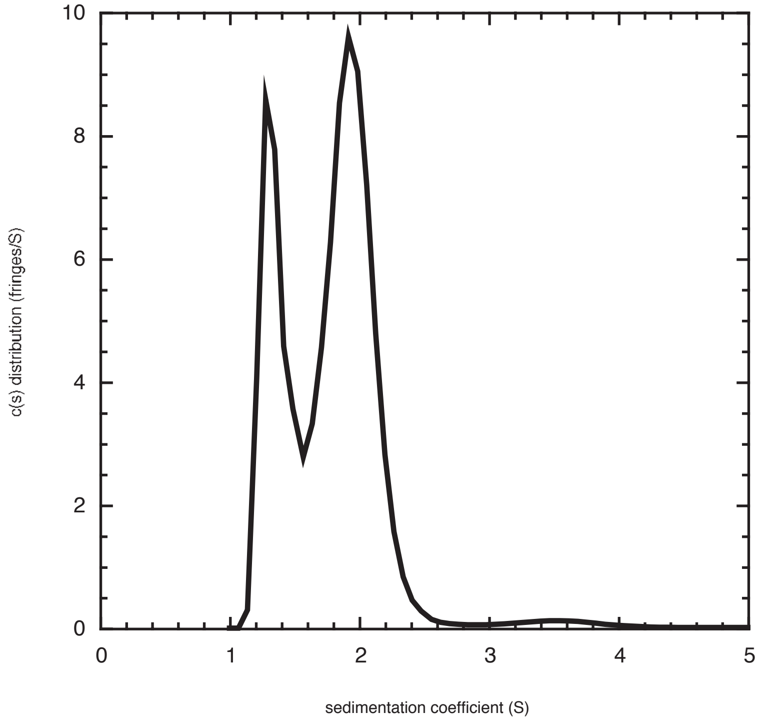


Figure 4-12. The c(s) distribution of C2G2CAP at 25°C. The loading concentration of C2G2CAP is 2.5 mg/ml. The dimer species becomes a more prominant peak when compared with Figure 4-11 due to the higher loading concentration. Integrating the area under the peaks suggests that ~27% are monomer, ~72% dimer and ~1% higher order oligomer (possibly tetramer).

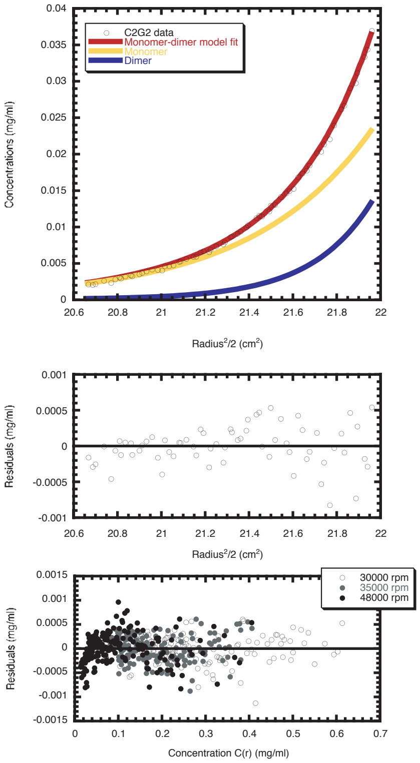


Figure 4-13. Sedimentation equilibrium of C2G2. The concentrations of the monomer and dimer species that constitute the concentration exponentials are shown at the top, and the fitting residuals are shown in the center. At the bottom are the fitting residuals for all of the datasets (three concentrations, three speeds).

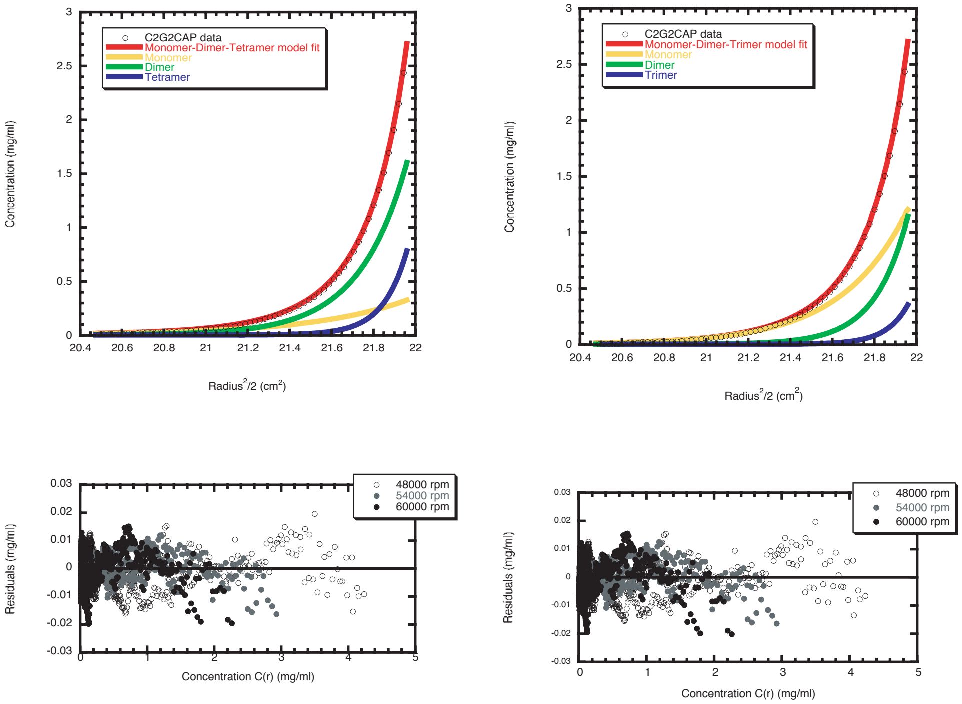


Figure 4-14. Sedimentation equilibrium of C2G2CAP. Data fitting to both M-D-Tri and M-D-Tet models are shown. M-D-Tri fits are on the left, and M-D-Tet fits are on the right. Both fits show reasonable residuals when the reduced buoyant molar mass was allowed to change, but M-D-Tet fits the data best (as illustrated by the statistics in Table 4-3). M-D-Tri model is shown here for comparison between the different species concentration ratios obtained from equilibrium and velocity sedimentation experiments.

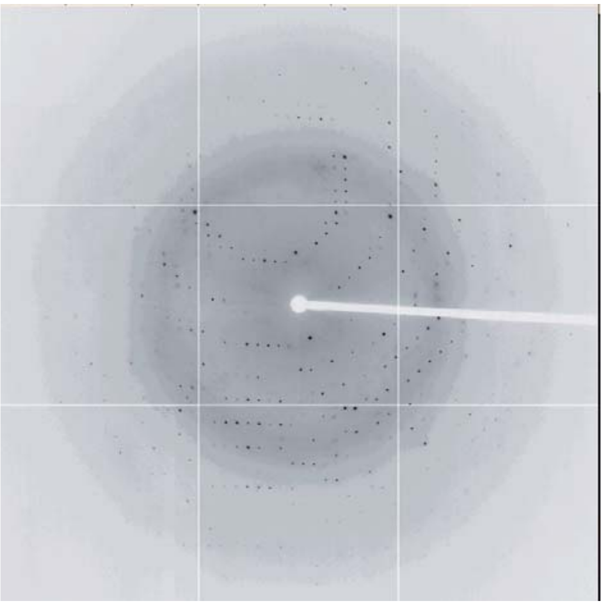


Figure 4-15. Diffraction pattern from synchrotron X-ray source. Data were collected at the SSRL synchrotron to 1.8 Å resolution. From the diffraction map, the crystal is indexed by DENZO to space group R32 (see text for details).

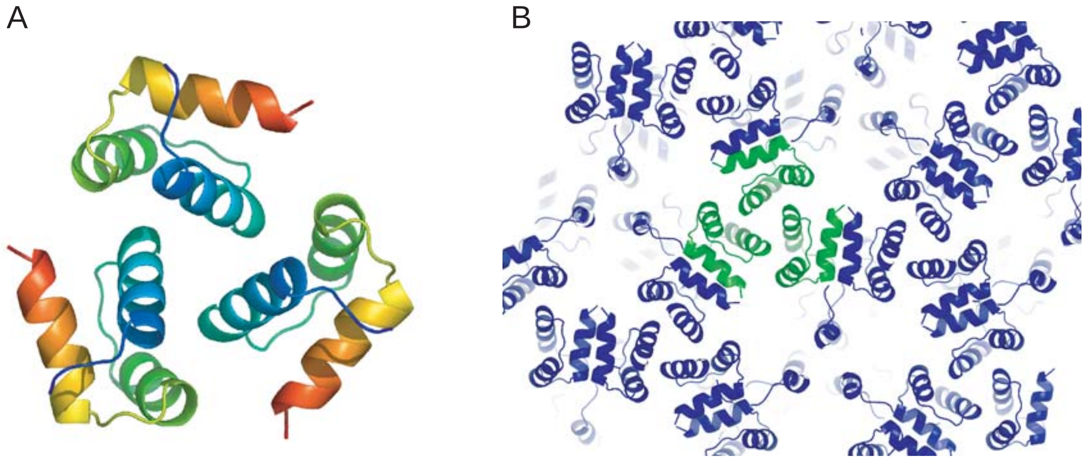


Figure 4-16. Preliminary X-ray crystallography results. **A.** The current best molecular replacement solution suggests a trimer arrangement of C2G2CAP. **B.** The crystal packing of our current model.

Appendix

Designing Good Primers with PRIDE*

*Adapted from ORBIT manual, Chapter 18, written by Marie Ary

Abstract

The PRIDE (PRImer DEsign) program can be used to design good primers for gene construction with the two-step recursive PCR method described at the end of this appendix. The sequence of the protein to be made is input, along with the number of primers desired and the restriction enzymes to be used for the 5′ and 3′ cloning sites. The primers are designed so that the overlaps are all the same length and have the same number of GC pairs. This gives a similar degree of annealing at all overlaps and allows a single optimal annealing temperature to be used for all the primers. The program checks for multiple occurrences of the restriction sites and warns the user of terminal nonspecific annealing or if primer dimers are likely to form. Primer sequences are regenerated or modified or the user can rerun the program until all requirements are met. The final primer sequences can be printed out at the end along with an estimate of the cost.

How PRIDE Works

The user specifies the sequence of the protein to be made in an ASCII file. The number of primers desired and the restriction enzymes to be used for the 5′ and 3′ cloning sites are specified interactively by answering the prompts or via an input file (see Specifying the Parameters below). The program takes the protein sequence input, translates it into a corresponding DNA sequence, then adds the GC clamps, the START codon (N-terminal methionine), the STOP codon (TAA), and the sequences for the restriction enzymes specified at the 5′ and 3′ ends to generate an initial full length sequence for the insert. This is then divided up into the specified number of primers.

The overlapping regions are designed first so that they are exactly 22 nucleotides long and have a GC content of exactly 10. This is done by checking for GC content, and either changing the third position of a degenerate codon to G or C if possible until there are 10 or shifting the location of the overlap region. This gives a similar degree of annealing at all overlaps and allows a single optimal annealing temperature to be used for all the primers.

Once the overlapping regions are set, the sequences of the nonoverlapping regions are randomized. That is, the program generates these sequences by randomly selecting the DNA codon corresponding to each amino acid from a set of degenerate codons preselected for good expression (> 15% occurrence) in the specified host (*E. coli*) (see Table A-1 below). Instead of always using the same codon for a given amino acid (e.g. the one with the highest occurrence rate), randomization increases the diversity of the primers thereby decreasing the chances of nonspecific annealing. You can specify your own random number seed for this process or have the program choose one for you.

The program then checks for multiple occurrences of the specified restriction sites and repeats the randomization until any multiples are eliminated. If a restriction site is found in an overlapping region, an error message is output and the program is aborted.

The final primer sequences are then listed, aligned at the overlaps. The program checks these for terminal nonspecific annealing and the possibility of primer dimer formation. If more than three terminal base pairs match, these are indicated in the output along with a warning.

The possibility of primer dimer formation is indicated with a primer dimer score. Each GC pair counts 3, each AT pair counts 2, and a geometrical sum is used to account for the number of consecutive matches. A higher score is worse. Scores greater than 42 are listed in the output with a warning. The user should always look at the alignment to make sure that matches outside the overlap region are not likely to compete with the overlap region. If so, the program should be rerun using a different random number seed or a different number of primers. If desired, the user can print out a listing of the final primer sequences and an estimate of their cost with and without PAGE purification.

Creating the Protein Sequence File

The file containing the amino acid sequence of the protein to be made is in ASCII format and can be created or modified using a text editor. The amino acids are specified starting from the N-terminus using the standard 1-letter abbreviations (all uppercase). Non natural amino acids cannot be used. N-terminal methionine, a STOP codon, and GC clamps will be added automatically, so do not include these when specifying your protein sequence. Currently, the protein sequence length is limited to 500 amino acids (minimum length is 8). The protein sequence file for protein G, for example, should begin as follows:

```
>  TTYKLILN.....
```

If running the program on an SGI machine (e.g., corona), make sure to remove any new line characters in the file. Put the sequence file in a convenient place in your home directory.

Specifying the Parameters

The PRIDE parameters may be specified interactively by answering the following prompts:

1. Name of the protein sequence file (include path if needed).
2. Number of primers desired (even integer).
3. Restriction enzyme for the 5′ end: (1) *Hind*III, (2) *Nde*I, (3) *Bam*HI, (4) *Pst*1.
4. Restriction enzyme for the 3′ end: (1) *Hind*III, (2) *Nde*I, (3) *Bam*HI, (4) *Pst*1.
5. Random number seed (integer or 0).
6. Printout for ordering desired (y or n). If yes (y), then:
7. Name of the printout for ordering file. This file lists the final primer sequences, their cost, and the final random number seed (which may have been incremented by the program).
8. Primer group name. A numerical identifier will be appended to the group name to identify each primer (e.g., group_name-1, group_name-2, etc.).

The parameters may also be specified in an input file created previously. They must be listed in the order given above, each one on a new line, left justified. Details on how to choose the number of primers, restriction enzymes, and random number seed are given below.

Number of Primers

The value for the number of primers (#primers) must be even and should be chosen so that the length of each primer (including the overlaps) is less than 100 (less than 80 is preferred). The number of primers can be calculated from the following:

$$\text{primer\_length} = \text{midsection} + (2 \times \text{overlap}) \leq 100$$

$$\text{primer\_length} = \frac{\text{full\_length} + [\text{overlap} \times (\#\text{primers} - 1)}{\#\text{primers}}$$

$$\text{full\_length} = \text{full length of the insert}$$

$$= (3 \times \#\text{residues}) + 5'\text{terminal\_length} + 3'\text{terminal\_length})$$

$$5'\text{terminal\_length} = 5'\text{cloning\_site\_length} + \text{GCclamp} = 6 + 6 = 12$$

$$3'\text{terminal\_length} = 3'\text{cloning\_site\_length} + \text{GCclamp} = 6 + 6 = 12$$

$$\text{overlap} = 22$$

Restriction Enzymes

The Mayo lab is currently using two vectors, pET11-M and pET11a. Three unique cutting sites are available for pET11-M (*Hind*III, *Pst*I, and *Bam* HI), and two unique cutting sites are available for pET11a (*Nde*I and *Bam* HI). You can choose a single restriction enzyme for both the 5′ and 3′ ends of your insert, or you can specify two different ones. In general, it is better to pick two different enzymes. Make sure they are specified in the proper order.

Restriction Enzyme	**Recognition Sequence**
(1) *Hin*dIII	5′ A▾AGCTT 3′, 3′ TTCGA▾A 5′
(2) *Nde*I	5′ CA▾TATG 3′, 3′ GTAT▾AC 5′
(3) *Bam*HI	5′ G▾GATCC 3′, 3′ CCTAG▾G 5′
(4) *Pst*I	5′ CTGCA▾G 3′, 3′ G▾ACGTC 5′

Enter the appropriate number (1, 2, 3, or 4) in response to the prompt. The enzyme to be used at the 5′ end is specified first, followed by the 3′ end. The following is recommended:

Vector	**5′ end**	**3′ end**
pET11-M	*Hin*dIII	*Pst*I
pET11a	*Nde*I	*Bam* HI

Random Number Seed

You can have the program choose the random number seed used to select the codons for the non-overlapping regions (a different one each time), or you may specify your own random number seed. Using your own seed ensures that you will get the same results if you decide to run the program again. Enter your own favorite random number seed (integer) or enter 0 to have the program choose one for you.

Running PRIDE

Currently, ra does not include a C++ compiler, so PRIDE cannot be run on ra. It should therefore be run on corona. Log in to corona and go to the directory containing the protein sequence file to be used. To run PRIDE interactively, enter the UNIX command:

```
~possu/bin/PRIDE
```

and answer the prompts. The results will automatically be displayed on the screen. If you have created an input file, specify that it be read in. The PRIDE results may also be saved in an output file:

```
~possu/bin/PRIDE < input_file > output_file
```

You can specify that the job be run in the background by including two "&" as follows:

```
~possu/bin/PRIDE < input_file >& output_file &
```

Looking at the Output — the Primer Dimer Report

PRIDE generates an output file that does the following (see Sample PRIDE Output File below):

1. Echoes input for first five parameters.
2. Lists overlap scan results for each set of primers, including the number of positions with fixed nucleotide identity and the number of positions with optional nucleotide identities.
3. Notes when all primer overlaps achieve GC content of 10.
4. Lists randomization process for non-overlaps.
5. Lists number of restriction sites found for the two restriction enzymes chosen. If there is more than one occurrence for an enzyme, randomization is repeated and the random number seed is incremented by 1.
6. Lists sequences of the final primers (aligned at the overlaps if screen is wide enough to view it).
7. Reports any nonspecific annealing and primer dimers found.

You should always look at the alignment to make sure that matches outside the overlap region are not likely to compete with the overlap region. If so, the program should be rerun using a different random number seed or number of primers. If specified, the printout for ordering file is output. This file lists the sequences of the designed primers, their estimated cost, and the final random number seed.

Sample PRIDE Output File

The following PRIDE output was obtained for a protein G variant.

```
*************************************************
*
*RECURSIVE PCR PRIMER DESIGN PROGRAM
*a software to make our life a little bit easier
*
*-by possu@caltech, 2000

************************************************


Please enter your filename:pgb
Please enter the number of primers desired:4

You have 55 residues in your target sequence

TYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE

starting magical reverse translation....
your sequence is 165 nucleotides long.

Which enzyme would you use for your 5' end cloning site:
(1) HindIII
(2) NdeI
(3) BamHI
(4) Pst1
And your selection:1

Which enzyme would you use for your 3' end cloning site:
(1) HindIII
(2) NdeI
(3) BamHI
(4) Pst1
And your selection:3


Please enter your favorite random number seed (or enter 0 for system pick):1234
Setting nucleotide overlap to be 22
----

shifting = 0
primer 1 and 2
You have 7 A.
You have 1 T.
You have 4 C.
You have 4 G.
You have 0 ATG.
You have 2 AG.
You have 4 CT.
----
```

```
shifting = 0
primer 2 and 3
You have 7 A.
You have 3 T.
You have 2 C.
You have 3 G.
You have 2 ATG.
You have 3 AG.
You have 2 CT.

----

shifting = 0
primer 3 and 4
You have 6 A.
You have 2 T.
You have 1 C.
You have 7 G.
You have 0 ATG.
You have 1 AG.
You have 5 CT.

You have 7 A.
You have 1 T.
You have 4 C.
You have 4 G.
You have 0 ATG.
You have 2 AG.
You have 4 CT.

8 = init count
9
10
congratulations, your oligoe 1 design is now done with no sweat.

==============

You have 7 A.
You have 3 T.
You have 2 C.
You have 3 G.
You have 2 ATG.
You have 3 AG.
You have 2 CT.

5 = init count
6
7
You have 7 A.
You have 3 T.
You have 4 C.
You have 3 G.
You have 2 ATG.
You have 3 AG.
You have 0 CT.

7 = init count in second round check
8
9
10
congratulations, your oligoe 2 design is now done with no sweat.

===============

You have 6 A.
You have 2 T.
You have 1 C.
You have 7 G.
You have 0 ATG.
You have 1 AG.
You have 5 CT.

8 = init count
9
10
congratulations, your oligoe 3 design is now done with no sweat.
===============

randomization... making your oligoes unique!
0 = random7
0 = random7
0 = random6
1 = random7
1 = random7
1 = random6
1 = random7
0 = random6
1 = random5
0 = random5
1 = random7
0 = random5
2 = random5
0 = random7
0 = random5
1 = random7
1 = random7
0 = random5
0 = random7
0 = random5
0 = random7
1 = random6
0 = random7
0 = random7
0 = random7
0 = random5
1 = random7
0 = random6
```

```
HindIII occured 1 time(s) in your sequence.
BamHI occured 1 time(s) in your sequence.



primer design finished, printing results:

GGCGGCGCAAGCTTATGACCTACAAACTGATTCTGAACGGTAAGACCCTGAAAGGTGAAACCACTA
TGGGACTTTCCACTTTGGTGATGACTTCGACATCTACGTCGCT
GGCGACTCTTCCAAAAGTTCGTT
CGCTGAGAAGGTTTTCAAGCAATACGCAAACGATAACGGTGTAGACGGTGAATGGACCTATGATGA
TGCCACTTACCTGGATACTACTGCGTTGGTTCTGGAAGTGGC
ATTGACTTATTCCTAGGGCCATCGA


primer 1 has undesired 3' annealing with 2 at x offset = 9.
But polymerases do not extend this type of primer dimers, skipping....
matchsumupper = 44 y_upperstart = 0
Primer dimer is likely to happen for upper primer 1 and primer 2
  GGCGGCGCAAGCTTATGACCTACAAACTGATTCTGAACGGTAAGACCCTGAAAGGTGAAACCACTA->
    |     | |   |    ||||  |     |     | |  |   |    |||||       | |
<-TGGGACTTTCCACTTTGGTGATGACTTCGACATCTACGTCGCTGGCGACTCTTCCAAAAGTTCGTT

matchsumlower = 44 x_lowerstart = 0
Primer dimer is likely to happen for lower primer 1 and primer 2
  GGCGGCGCAAGCTTATGACCTACAAACTGATTCTGAACGGTAAGACCCTGAAAGGTGAAACCACTA->
    |     | |   |    ||||  |     |     | |  |   |    |||||       | |
<-TGGGACTTTCCACTTTGGTGATGACTTCGACATCTACGTCGCTGGCGACTCTTCCAAAAGTTCGTT

matchsumlower = 45 x_lowerstart = 1
Primer dimer is likely to happen for lower primer 1 and primer 2
  GGCGGCGCAAGCTTATGACCTACAAACTGATTCTGAACGGTAAGACCCTGAAAGGTGAAACCACTA->
    |   | ||| | |   |         | |   ||         | |||| | |     | |  |
 <-TGGGACTTTCCACTTTGGTGATGACTTCGACATCTACGTCGCTGGCGACTCTTCCAAAAGTTCGTT

matchsumlower = 212 x_lowerstart = 44
Primer dimer is likely to happen for lower primer 1 and primer 2
  GGCGGCGCAAGCTTATGACCTACAAACTGATTCTGAACGGTAAGACCCTGAAAGGTGAAACCACTA->
                                               |||||||||||||||||||||||
                                             <-TGGGACTTTCCACTTTGGTGATGACTTCGACATCTACGTCG
CTGGCGACTCTTCCAAAAGTTCGTT

matchsumupper = 44 y_upperstart = 1
Primer dimer is likely to happen for upper primer 1 and primer 3
But polymerases do not extend this type of primer dimers, skipping....
primer 1 has undesired 3' annealing with 4 at x offset = 12.
But polymerases do not extend this type of primer dimers, skipping....
primer 2 has undesired 3' annealing with 3 at x offset = 21.
But polymerases do not extend this type of primer dimers, skipping....
matchsumupper = 48 y_upperstart = 1
Primer dimer is likely to happen for upper primer 2 and primer 3
But polymerases do not extend this type of primer dimers, skipping....
matchsumupper = 212 y_upperstart = 44
```

```
Primer dimer is likely to happen for upper primer 2 and primer 3
But polymerases do not extend this type of primer dimers, skipping....
matchsumlower = 44 x_lowerstart = 1
Primer dimer is likely to happen for lower primer 2 and primer 3
  TTGCTTGAAAACCTTCTCAGCGGTCGCTGCATCTACAGCTTCAGTAGTGGTTTCACCTTTCAGGGT->
   |  |   |   |   ||| ||    |  | |   |   | |        |   | | |||   |
 <-AGTAGTATCCAGGTAAGTGGCAGATGTGGCAATAGCAAACGCATAACGAACTTTTGGAAGAGTCGC

matchsumlower = 44 x_lowerstart = 8
Primer dimer is likely to happen for lower primer 2 and primer 3
  TTGCTTGAAAACCTTCTCAGCGGTCGCTGCATCTACAGCTTCAGTAGTGGTTTCACCTTTCAGGGT->
            | | |     |  || | |  |  || | |   ||    || |  | |   |   |
        <-AGTAGTATCCAGGTAAGTGGCAGATGTGGCAATAGCAAACGCATAACGAACTTTTGGAAGAGTCGC

matchsumlower = 46 x_lowerstart = 10
Primer dimer is likely to happen for lower primer 2 and primer 3
  TTGCTTGAAAACCTTCTCAGCGGTCGCTGCATCTACAGCTTCAGTAGTGGTTTCACCTTTCAGGGT->
             | |  |  |      |  | |||||| | | |    | |  |     || |
          <-AGTAGTATCCAGGTAAGTGGCAGATGTGGCAATAGCAAACGCATAACGAACTTTTGGAAGAGTCGC

matchsumlower = 46 x_lowerstart = 13
Primer dimer is likely to happen for lower primer 2 and primer 3
  TTGCTTGAAAACCTTCTCAGCGGTCGCTGCATCTACAGCTTCAGTAGTGGTTTCACCTTTCAGGGT->
               |  |||  ||||  | || |  |  |  |  ||  | || |   || |  |
             <-AGTAGTATCCAGGTAAGTGGCAGATGTGGCAATAGCAAACGCATAACGAACTTTTGGAAGAGTCGC

primer 2 has undesired 3' annealing with 4 at x offset = 39.
But polymerases do not extend this type of primer dimers, skipping....
matchsumlower = 45 x_lowerstart = 2
Primer dimer is likely to happen for lower primer 2 and primer 4
  TTGCTTGAAAACCTTCTCAGCGGTCGCTGCATCTACAGCTTCAGTAGTGGTTTCACCTTTCAGGGT->
     |    ||     ||  |  | |||  |    ||     | |||   |  | |   | | ||
  <-TGCCACTTACCTGGATACTACTGCGTTGGTTCTGGAAGTGGCATTGACTTATTCCTAGGGCCATCGA

primer 3 has undesired 3' annealing with 4 at y offset = 18.
  CGCTGAGAAGGTTTTCAAGCAATACGCAAACGATAACGGTGTAGACGGTGAATGGACCTATGATGA->
                                                   |            ||||
                                               <-TGCCACTTACCTGGATACTACTGCGTTGGTTCTGGAAGT
GGCATTGACTTATTCCTAGGGCCATCGA

matchsumlower = 49 x_lowerstart = 2
Primer dimer is likely to happen for lower primer 3 and primer 4
  CGCTGAGAAGGTTTTCAAGCAATACGCAAACGATAACGGTGTAGACGGTGAATGGACCTATGATGA->
      |   |      | |  |  |||||| | | | |      |    |||||     |      |
  <-TGCCACTTACCTGGATACTACTGCGTTGGTTCTGGAAGTGGCATTGACTTATTCCTAGGGCCATCGA

matchsumlower = 212 x_lowerstart = 44
Primer dimer is likely to happen for lower primer 3 and primer 4
  CGCTGAGAAGGTTTTCAAGCAATACGCAAACGATAACGGTGTAGACGGTGAATGGACCTATGATGA->
                                              ||||||||||||||||||||||
                                            <-TGCCACTTACCTGGATACTACTGCGTTGGTTCTGGAAGTGGC
ATTGACTTATTCCTAGGGCCATCGA

Do you need a printout for ordering? (y/n):y
```

```
please enter output filename:pgb.out
Primer group name:pgb

Output finished.
```

Sample Printout for Ordering File

The following printout for ordering was obtained after running PRIDE on a protein G variant.

```
pgb-1
5'-GGCGGCGCAAGCTTATGACCTACAAACTGATTCTGAACGGTAAGACCCTGAAAGGTGAAACCACTA-3'
66mer
0.55 dollars/base * 66mer = 36.3

pgb-2
5'-TTGCTTGAAAACCTTCTCAGCGGTCGCTGCATCTACAGCTTCAGTAGTGGTTTCACCTTTCAGGGT-3'
66mer
0.55 dollars/base * 66mer = 36.3

pgb-3
5'-CGCTGAGAAGGTTTTCAAGCAATACGCAAACGATAACGGTGTAGACGGTGAATGGACCTATGATGA-3'
66mer
0.55 dollars/base * 66mer = 36.3

pgb-4
5'-AGCTACCGGGATCCTTATTCAGTTACGGTGAAGGTCTTGGTTGCGTCATCATAGGTCCATTCACCGT-3'
67mer
0.55 dollars/base * 67mer = 36.85

total cost without PAGE purification = 145.75
total cost with PAGE purification = 325.75
primers generated with random number seed:1234
```

Two-step Recursive PCR

The primers designed with PRIDE should give good results when used with the following two-step recursive PCR method. Since all the primers have the same overlap length (22 bp) and the same GC content, a single annealing temperature can be used during gene synthesis:

$$\begin{aligned} T_{anneal} &= T_m \text{ (overlap)} - 5\ ^\circ\text{C} \\ &= 64\ ^\circ\text{C} - 5\ ^\circ\text{C} \\ &= 59\ ^\circ\text{C} \end{aligned}$$

Step 1: Gene Synthesis

- 4 pmol of each oligo
- 0.25 mM dNTPs
- 2.5 U Pfu
- Pfu buffer
- 100 µl total volume
- 25 cycles:

  30 sec at 95 ˚C

  30 sec at T_m (overlap) – 5 ˚ C= 59 ˚C

  2 min at 72 ˚C

  final 10 min extension at 72 ˚C

Step 2: Amplification of Full Length Product

- ♦ 10 ml of the first reaction
- ♦ 50 pmol of each amplification oligo
- ♦ 0.25 mM dNTPs
- ♦ 2.5 U Pfu
- ♦ 100 ml total volume
- ♦ 30 cycles:

30 sec at 95 ˚C

30 sec at 45-50 ˚ C

2 min at 72 ˚C

final 10 min extension at 72 ˚C

References

1. Prodromou, C.; Pearl, L. H. Recursive PCR: A novel technique for total gene synthesis. *Protein Engin*. 5:827-829 (1992).

Table A-1. DNA Codons Used for *E. coli*

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	Arg	Asn	Asp	Cys	Gln	Gly	Glu	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
GCA	CGC	AAC	GAC	TGC	CAA	GAA	GGC	CAC	ATC	CTG	AAA	ATG	TTC	CCA	TCC	ACC	TGG	TAC	GTA
G	T		T	T	G	G	T	T	T		G		T	G	T	T	T	T	G
T																			