

The Pennsylvania State University

The Graduate School

College of Engineering

**DEVELOPMENT OF COMPUTATIONAL TOOLS FOR THE DESIGN AND
OPTIMIZATION OF COMBINATORIAL PROTEIN LIBRARIES**

A Thesis in

Chemical Engineering

by

Manish C. Saraf

© 2006 Manish C. Saraf

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2006

The thesis of Manish C. Saraf has been reviewed and approved* by the following:

Costas D. Maranas

Donald B. Broughton Professor of Chemical Engineering
Thesis Advisor
Chair of Committee

Stephen J. Benkovic

Evan Pugh Professor and Eberly Chair in Chemistry

Michael Pishko

Professor of Chemical Engineering
Associate Professor of Materials Science and Engineering

Patrick Cirino

Assistant Professor of Chemical Engineering

Andrew L. Zydney

Department Head
Walter L. Robb Chair and Professor of Chemical Engineering

*Signatures are on file in the Graduate School.

Abstract

Combinatorial protein library generation and screening has emerged as a powerful strategy for protein engineering. In commonly used recombination protocols (e.g., DNA shuffling [1, 2], StEP [3], ITCHY [4], SCRATCHY [5], SHIPREC [6]), the primary diversity generation mechanism entails the exchange of parental DNA fragments in the reassembled sequences. One of the key challenges in the use of such directed evolution techniques is that many of the reassembled hybrid proteins do not fold properly and thus are non-functional. Clearly, certain combinations of mutations/recombinations are incompatible leading to residue-residue clashes that contribute to the inability of the hybrids to fold correctly. In response to the problem at hand, this thesis presents both bioinformatics-based approaches and detailed atomistic structural calculations aimed at (i) identifying residue-residue clashes for prescreening purposes as well as targets for design and (ii) designing proteins and protein libraries with enhanced functionalities. The application of these tools will allow for appropriate allocation of diversity in the library while correcting problematic residue combinations. Therefore, experimental resources can now be focused towards the most promising regions of sequence space, thus increasing the chances of identifying novel engineered proteins.

Table of Contents

List of Figures	vii
List of Tables	x
Acknowledgements	xi
1 Introduction	1
1.1 Motivation and Objective	
1.2 Background	
1.3 Thesis Overview	
2 Using Multiple Sequence Correlation Analysis to Characterize Functionally Important Protein Regions	9
2.1 Background and Motivation	
2.2 Residue Correlation Analysis (RCA)	
2.2.1 Residue Correlation Coefficients	
2.2.2 Correlation Tendencies	
2.2.3 Site Entropy	
2.3 Computational Results and Comparisons	
2.3.1 Dihydrofolate Reductase (DHFR)	
2.3.2 Cyclophilin	
2.3.3 Formyl-Transferase	
2.3.4 Transmembrane amino acid transporter protein	
2.4 Summary and Discussion	
3 Using a Residue Clash Map to Functionally Characterize Protein Recombination Hybrids	41
3.1 Background	
3.2 Method for Generating Clash Maps	

3.2.1	Repulsive residue pairs	
3.2.2	Steric hindrance or cavity formation in the hybrids	
3.2.3	Hydrogen Bond Disruption	
3.3	Comparison with Experimental Results	
3.3.1	Glycinamide ribonucleotide transformylase (GART)	
3.3.2	Glutathione S-transferase (GST)	
3.3.3	β -lactamases	
3.3.4	Catechol-2,3-oxygenase (C23O)	
3.3.5	Dioxygenases	
3.4	Summary	
4	FamClash: A Method for Ranking the Activity of Engineered Enzymes	68
4.1	Background	
4.2	Hybrid construction and functional screening	
4.2.1	Plasmid constructions	
4.2.2	Construction of DHFR hybrid libraries	
4.2.3	Selection of DHFR hybrids	
4.2.4	DHFR Assays	
4.3	FamClash Method	
4.4	Results and Discussion	
4.4.1	Library construction and hybrid isolation	
4.4.2	FamClash analysis of EB library	
4.4.3	DHFR hybrid characterization and analysis	
4.5:	Summary	
5	Design of Combinatorial Protein Libraries of Optimal Size	88
5.1	Background	
5.2	OPTCOMB Modeling Framework	
5.3	Results and Discussion	
5.4	Summary	

6	IPRO: An Iterative Computational Protein Library Redesign and Optimization Procedure	112
6.1	Background	
6.2	The IPRO Modeling Framework	
6.2.1	Generating a set of sequences representative of the combinatorial library	
6.2.2	Generation of starting hybrid protein structures	
6.2.3	Selecting design positions	
6.2.4	Iterative protein optimization step	
6.3	Hybrid Construction and Functional Screening	
6.3.1	Construction of DHFR hybrid libraries	
6.3.2	Selection and determination of specific activities of active DHFR hybrids	
6.4	Application Example	
6.4.1	DHFR Library Characterization and Analysis	
6.4.2	IPRO Analysis of DHFR Libraries	
6.5	Summary and Discussion	
7	Conclusions	153
7.1	Future Perspectives	
	Bibliography	160

List of Figures

1.1	Schematic representation of the key steps of directed evolution experiments.	8
2.1	Formation of a repulsive ion pair in a recombinant hybrid that may disrupt contacting pairs as well as essential motions.	29
2.2	Description for calculation of correlation coefficient from multiple sequence alignment.	30
2.3	Comparison of experimentally determined functional regions of DHFR enzyme with those predicted using Residue Correlation Analysis (RCA).	31
2.4	Plot of residue correlation coefficients (r_{ij}) versus C_{β} - C_{β} distances for pairs of residues in the Cyclophilin protein family.	33
2.5	Comparison of experimentally determined functional regions of Cyclophilin enzyme with those predicted using Residue Correlation Analysis (RCA).	34
2.6	Plot of correlation coefficient versus C_{β} - C_{β} distance for pairs of residues in the formyl-transferase protein family.	35
2.7	Comparison of experimentally determined functional regions of formyl transferase enzyme with those predicted using Residue Correlation Analysis (RCA).	36
2.8	The correlation tendency plot (at $r_c = 0.4$) for different segments of the transmembrane amino acid transporter protein.	38
3.1	Residue-residue clashes may arise in protein hybrids due to different directionality in the parental sequences of a charged pair, residue sizes or hydrogen bond donor-acceptor pair.	58
3.2	Graphical representation of clashes using arcs	59
3.3	Different types of clashes for (a) purN/hGART and (b) hGART/purN shown as arcs linking the two positions involved.	60
3.4	Different types of clashes between residues in GST M1-1 and GST M2-2 hybrids are shown as bicolored arcs.	62

3.5	The identified residue clashes are shown against the ten active β -lactamase (TEM-1 (black), PSE-4(gray)) hybrids identified experimentally.	63
3.6	Residue-residue clashes in seven different thermally stable C23O hybrids obtained by shuffling ssDNA are shown.	64
3.7	Eight toluene-active members of the hybrid library obtained by shuffling genes encoding the α and β subunits of three dioxygenases (<i>tecA1A2</i> , <i>todC1C2</i> and <i>bhpA1A2</i>) are shown as horizontal bars.	65
4.1	Property values for the residue pair in the hybrid that are significantly different than those observed in the protein family denote a clash.	82
4.2	Predicted clashes in EB hybrids are shown for all single crossover EB hybrids. A clash between any two residue positions is shown as an arc.	83
4.3	The number of clashes in each of the single crossover EB (—) and BE (---) DHFR hybrids are plotted against crossover position.	84
4.4	Plot of specific activities and number of identified clashes of the 13 EB DHFR hybrids against crossover position. The specific activity and number of clashes for hybrid 62 is shown separately.	85
5.1	The combinatorial libraries are designed using two different design rules: (a) all parental sequences contribute fragments at each of the junction points, and (b) selective restrictions are imposed on the set of oligomers being contributed by the parents.	104
5.2	Clash maps determined using the FamClash procedure corresponding to the three different sequence combinations (<i>E. coli</i> - <i>L. casei</i> , <i>B. subtilis</i> - <i>L. casei</i> , and <i>B. subtilis</i> - <i>E. coli</i>)	105
5.3	Plot of the number of clashes between optimally designed oligomers (♦) using models (a) M1 and (b) M2 against library size. The average numbers of clashes between randomly generated designs (▲) for various library sizes are also shown.	106
5.4	Results obtained using model M2 for minimum and maximum fragment lengths of 15 and 30 residues respectively and $N = 21$.	107
5.5	Plot of (a) number of clashes, (b) percent of clash-free hybrids, and average number of clashes per hybrid in the optimal tiling pattern determined using OPTCOMB.	108
5.6	Plot of the optimal library size for different ranges (10-25, 15-30, 20-35,	

	25-40, and 30-45) of fragment lengths.	110
5.7	The tiling choices and the clash distributions for the hybrids for $N=15, 18, 21, 24, 27$, and 30 .	111
6.1	Illustration of (a) downstream redesign of promising hybrids and (b) upstream parental sequence redesign.	140
6.2	The four key steps involved in the IPRO procedure. Details of each of these steps are described separately in the text.	141
6.3	This figure highlights the key steps for constructing the initial structure of a hybrid protein from a set of parental structures with known crossover position(s).	142
6.4	Illustration of the key steps involved in each iteration of IPRO procedure.	143
6.5	The design positions within the perturbation region are permitted to change amino acid type while the flanking residue positions can only change rotamers but not the residue type.	145
6.6	Plot of the natural log of the specific activities against the binding scores for two different types of DHFR hybrids (a) <i>E. coli/B. subtilis</i> and (b) <i>B. subtilis/L. casei</i> .	146
6.7	Binding score profile before and after redesign of the <i>E. coli/B. subtilis</i> DHFR hybrids using the SSDP framework when (a) only clashing residue positions are considered and (b) only binding pocket residues are considered for redesign.	147
6.8	Binding score profile before and after redesign of parental <i>E. coli</i> and <i>B. subtilis</i> DHFR sequences using the HLDP framework.	148
6.9	(a) Substitution of serine with an arginine at position 64 stabilizes the binding with the cofactor NADPH due to formation of a new salt bridge. (b) Substitution of tyrosine and tryptophan at position 30 with a smaller aromatic residue phenylalanine perhaps reduces steric hindrance with the substrate DHF.	149

List of Tables

2.1	Summary of statistical analyses for the three protein families.	39
2.2	Residue pairs, their role in catalysis and correlation coefficients.	40
3.1	Summary of statistical analysis for the five protein families.	66
3.2	Clash map based analysis for the dioxygenase system.	67
4.1	Summary of the FamClash procedure.	86
4.2	Positions, residue pairs and nature of clashes in the hybrids.	87
6.1	Crossover positions for the <i>E.coli/B.subtilis</i> and <i>B.subtilis/L. casei</i> DHFR hybrids and their specific activities (mmol/min/mg).	150
6.2	Individual redesigns of the (a) clashing positions and (b) binding site residues for the <i>E.coli/B.subtilis</i> hybrid DHFR sequences.	151
6.3	Redesign of parental <i>E. coli</i> and <i>B. subtilis</i> DHFR sequences.	152

Acknowledgements

I am grateful to many people who have contributed significantly towards shaping this thesis.

It is difficult to overstate my gratitude to my Ph.D. thesis advisor, Costas Maranas. With his enthusiasm, his inspiration, and his great efforts to explain things, he made my learning experience at Penn State very interesting. Throughout my research period, he provided encouragement, sound advice, good teaching, and lots of good ideas. This thesis would not have been possible without him.

A lot of the work was done in collaboration with Dr. Stephen Benkovic. I would like to thank him and his group for giving me the opportunity to participate in their group meetings and for graciously sharing essential experimental data used in some of my publications.

It was always a pleasure to discuss my work with my thesis committee members, Stephen Benkovic, Michael Pishko, and Partick Cirino, and I thank them both for their insightful comments and enthusiastic support for this work.

I would like to thank all the past and present members of the Maranas group for making it such a great place to work, especially Greg and Anshuman from whom I have had the opportunity to learn a great deal and Tony who did an outstanding job of making me an enthusiastic Body Work's (fitness center) member. I'm proud to see that they all have continued on to successful careers.

Finally, I owe everything to my family. The memories of my father and his dreams for me has been a great source of inspiration in my life, Mom who has always been very caring and loving, brothers and sister who have always been at my side as great friends, and Manisha, my wife, for supporting me in so many ways. To them I dedicate this thesis.

Chapter 1: Introduction

Section 1.1: Motivation and Objective

The ability to proactively modify protein structure and function through a series of targeted mutations is an open challenge that is central in many different applications. These include, among others, enhanced catalytic activity [7-9] and stability [10, 11], creation of gene switches for the control of gene expression for use in gene therapy and metabolic engineering [12, 13], signal transduction [14, 15], genetic recombination [16], motor protein function, and regulation of cellular processes (see ref. [17] for a review). This task is complicated by the fact that proteins rely on complex networks of subtle interactions to enable function [18-20]. Therefore, the effect of a mutation is difficult to assess *a priori* requiring the capture of its direct or indirect effects on many neighboring amino acids. As a result, most protein engineering paradigms involve the synthesis and screening of multiple protein candidates (protein library) as a way to enhance the odds of identifying proteins with the desired functionality level.

These directed evolution experiments [1, 3, 21-24] (see Figure 1.1), for generating protein variants, typically involve repeated cycles of mutagenesis and/or recombination. Over the years an impressive array of such successful experiments [25-32] has been carried through inspection of protein structure and function. Nevertheless, it has become clear that as soon as more than a few mutations are required, the design problem rapidly becomes too complex to solve by inspection. In addition, the combinatorial explosion of possibilities and some experimental biases (e.g., codon usage) allows exploration of only a small region of the vast sequence space and, therefore, cause many opportunities to be missed. Therefore, it is desirable to be able to (i) *a priori* prescreen

protein hybrids for their potential of being stably folded [33] and functional, and (ii) identify what sequence permutations are the most promising in terms of improving/preserving protein structure and function. To this end, the underlying objective of this research is to develop a set of computational tools aimed at modeling and subsequently optimizing the stability and functionality of combinatorial protein libraries. I believe this research will have a broad impact on the general area of protein engineering by introducing formal systems information technology and molecular modeling approaches to the field.

Section 1.2: Background

Recent advances in protein engineering [5, 6, 34-36] have allowed researchers to go beyond the limitations of homology-dependent directed evolution methods. The success or failure of a directed evolution study is ultimately intertwined with the average activity/functionality of the generated combinatorial library. Typically only a small portion of the library is active, and experimental evidence [37] suggests that this average activity is likely to be even lower when low sequence identity pairs are recombined. While *a priori* determination with certainty of the activity level or lack thereof of a given protein sequence may be difficult, identifying (i) “sequence features” such as conserved amino acids and correlated amino acid pairs with a statistically measurable activity signature, (ii) clashing residue-residue combinations brought together through recombination, and (iii) residue redesign candidates for restoring lost functionality in hybrids provides strategies to be explored experimentally for library activity improvement. It is worthwhile to note that the magnitude of these effects does not have to be large to be effective given that directed evolution cycles over multiple generations will

lead to substantial amplification. A number of interesting efforts designed to pinpoint residues and regions critical to a protein's structure have recently been published in the literature. Voigt *et al.*, [38] examined two approaches for fitness level prediction of hybrid proteins in the context of directed evolution. First, they randomly generated a hypothetical fitness landscape including both one- and two-body interactions and subjected it to simulated rounds of mutagenesis and screening. As the fitness level increased, changing a highly coupled residue became more and more unlikely to provide improvement. Second, they used the concept of sequence entropy (calculated with mean-field theory [39, 40]) to predict a protein's structural tolerance for point mutation at each sequence position. It was shown that in random mutagenesis directed evolution experiments involving subtilisin E and T4 lysozyme, the majority of mutations responsible for improved hybrids occurred in positions with high structural tolerance for change. Bogarad and Deem [41] utilized a generalized block NK model [42, 43] for protein fitness and found that new folds can possibly be developed with the exchange of nonhomologous domains.

In the context of ITCHY and SCRATCHY this occurrence may disrupt protein structure and/or function within the hybrid library. Specific structural requirements involve detailed molecular modeling approaches for designing a protein or collection of proteins with a given structure. This typically involves finding the amino acid sequence that best fits the given protein fold. The protein fold is represented by the Cartesian coordinates of its backbone atoms, which are usually fixed in space so that the degrees of freedom associated with backbone movement are neglected (some notable exceptions to the "fixed backbone" design paradigm include the work of [44-49]). Candidate protein designs are

generated by selecting amino acid side chains (at atomistic detail) along the backbone design scaffold. For simplicity, side chains are usually only permitted to assume a discrete set of statistically preferred conformations called *rotamers* (see [50] for a review of current rotamer libraries). Thus, a protein design consists of both a residue *and* a rotamer assignment for each amino acid position. To evaluate how well a possible design fits a given fold, rotamer/backbone and rotamer/rotamer interaction energies for all of the rotamers in the chosen library are tabulated. These potential energies can then be approximated using any of many standard force fields (*e.g.*, CHARMM [51], DREIDING [52], AMBER [53], GROMOS [54]). Alternatively, energy/scoring functions that have been customized for protein design [55-57] can be used. Protein design potentials (see ref. [58] for a review) typically include van der Waals interactions, hydrogen bonding, electrostatics, solvation, and even entropy-based penalties for flexible side-chains (*e.g.*, arginine) [30, 59-61]. Both deterministic and stochastic methods have been used to solve this problem (see [62, 63] for reviews). Successful designs include a correctly folding zinc finger protein without zinc [64], alternate core packing for phage 434 cro [65] and ubiquitin [66], metal binding proteins [67, 68], a more stable version of an integrin I domain [69], α -helical bundle proteins with a right-handed twist [45], and the recent significant contributions of Hellenga's lab on binding proteins and receptors [70, 71].

This research builds on modeling and computational works outlined above and develops new ones for modeling and subsequently optimizing the stability and functionality of combinatorial protein libraries. By appropriately allocating diversity in the library while correcting problematic residue combinations, experimental resources can be focused towards the most promising regions of the sequence space, thus increasing the chances of

identifying novel engineered proteins. In the next section, I outline the techniques we have developed in this research to guide the design of combinatorial protein libraries. Both sequence and structure information encoded in the parental/family sequences as well as detailed molecular modeling techniques have been extensively utilized.

1.3 Thesis Overview

The subsequent chapters can broadly be divided into two parts. The first part, including Chapters 2, 3, and 4, outlines approaches developed to *a priori* prescreen protein hybrids for their potential of being stably folded [33] and functional. Specifically, these approaches identify (a) which protein regions are likely to tolerate mutations and which ones do not and (b) what amino acid pairs form residue–residue clashes in hybrid proteins. The next part, Chapters 5 and 6, outlines two computational approaches for systematic design of combinatorial libraries corresponding to two separate experimental paradigms for library generation: (1) parental/hybrid sequence redesign through point mutations and/or (2) recombination of parental segments.

In Chapter 2, a bioinformatics inspired approach is presented for identifying which protein regions are likely to tolerate mutations and which ones do not. The approach is based on only sequence information and examines correlation of residue substitutions occurring in the members of the protein family to be engineered. A high correlation between two sites implies the presence of interaction between these two positions. The proposed computational framework identifies regions with residues that interact with an uncommonly high number of other residues. I show that strongly correlated residue positions will be those that affect the dynamics of protein function or are involved in favorable interactions. Hence, uncoordinated changes in these regions are likely to have

adverse effects on the functionality.

Chapter 3 describes a rapid, protein sequence data-based approach to characterize all possible residue pairs present in protein hybrids for inconsistency with protein family structural features. This approach is based on examining contacting residue pairs with different parental origins for different types of potentially unfavorable interactions (i.e., electrostatic repulsion, steric hindrance, cavity formation, and hydrogen bond disruption). I contrast the identified clashing residue pairs between members of a protein family against functionally characterized hybrid libraries to demonstrate that residue clash maps can provide quantitative guidelines for the placement of crossovers in the design of protein recombination experiments.

Chapter 4 introduces the computational procedure FamClash for analyzing incompatibilities in engineered protein hybrids by using protein family sequence data. All pairs of residue positions in the sequence alignment that conserve the property triplet of charge, volume, and hydrophobicity are first identified and significant deviations are denoted as residue-residue clashes. This approach moves beyond earlier efforts aimed at solely classifying hybrids as functional or nonfunctional by correlating the rank ordering of these hybrids based on their activity levels. Experimental testing of this approach was performed in parallel to assess the predictive ability of FamClash.

Chapter 5 presents a computational procedure, OPTCOMB (Optimal Pattern of Tiling for COMBinatorial library design), for designing protein hybrid libraries that optimally balance library size with quality. The proposed procedure is directly applicable to oligonucleotide ligation-based protocols such as GeneReassembly, DHR, SISDC, and many more. Given a set of parental sequences and the size ranges of the parental

sequence fragments, OPTCOMB determines the optimal junction points (i.e., crossover positions) and the fragment contributing parental sequences at each one of the junction points. By rationally selecting the junction points and the contributing parental sequences, the number of clashes (i.e., unfavorable interactions identified using approaches highlighted in Chapters 2, 3, and 4) in the library is systematically minimized with the aim of improving the overall library quality.

Chapter 6 describes the computer program and algorithm IPRO (Iterative Protein Redesign and Optimization) devised for redesigning parental sequences so that their recombination will yield combinatorial libraries with an enhanced fraction of active members. The goal here is to find point mutations in the parental sequences that will propagate into the combinatorial library and ameliorate clashes (i.e., unfavorable interactions) in the resulting hybrids. IPRO also allows downstream redesign of promising hybrids. These models involve detailed atomistic structural calculations and factor important aspects such as backbone flexibility and protein-protein and protein-ligand docking.

Chapter 7 concludes by providing a summary of the key contributions of the proceeding chapters as well as some perspective on the future of computational protein design.

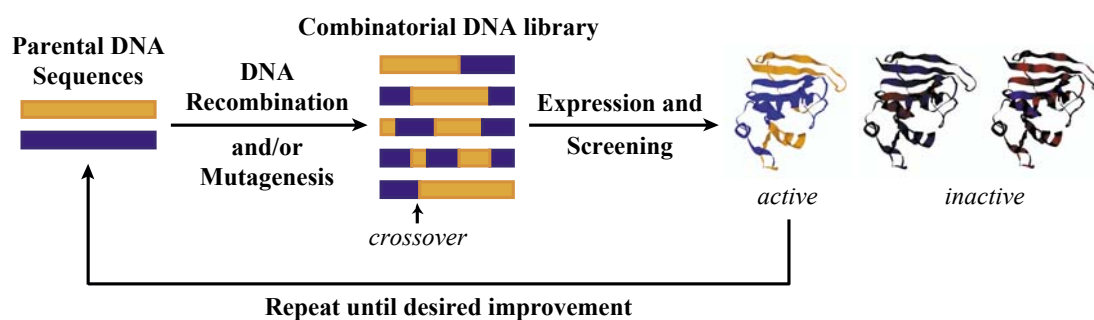


Figure 1.1: Schematic representation of the key steps of directed evolution experiments. Note that only a few of the library members pass through the screening step. Crossovers are defined as the junction points between segments from different parental sequences.

Chapter 2: Using Multiple Sequence Correlation Analysis to Characterize Functionally Important Protein Regions

Section 2.1: Background and Motivation

The contribution of different protein regions to function is determined by the interactions formed with substrates, cofactors and other residues. Considerable effort has been devoted to identifying functional protein regions from known amino acid sequences [72-74]. When families of sequences with similar structure and function are aligned, it is possible to glean conserved patterns that encapsulate important functional domains [75]. However, in many cases residues are not conserved but rather co-evolve with their interacting partners to retain important interactions and hence function. Through evolution, many families have diverged substantially, so that it is typically difficult to directly observe any distinct patterns. The inability to identify these functionally important regions (especially those that co-evolve and hence vary in a coordinated manner) poses a long-standing challenge in protein engineering efforts, particularly in the context of directed evolution experiments.

Directed evolution experiments aim at creating diverse sequences with novel properties (e.g., enhanced catalytic activity, stereoselectivity, thermostability, etc.) in the form of combinatorial libraries generated by (i) creating sequence permutations of parent sequences through recombination [1, 3, 21, 22]; and/or (ii) introducing point mutations at either specific positions [23] or randomly [24]. Moore *et al.* proposed modeling frameworks [76, 77] for quantifying the statistics of crossover allocations and mutations in the recombinant sequences. The key problem here is that usually such sequence modifications are not coordinated and, therefore, disrupt functional domains or introduce incompatible interactions (see, Figure 2.1) that frequently results in the loss of their

function. It is therefore desirable to be able to predict these functional sites that are less likely to tolerate uncompensated mutation/crossover, so as to guide these combinatorial libraries towards retaining a higher level of functionality. This objective defines the scope of this study.

Multiple sequence alignment (MSA) and entropy calculations provide some insight into identifying protein building blocks preserved through evolution. However, there are currently no reliable techniques for identifying regions that may subtly affect functionality by taking part in large number of interactions that co-evolve under functional or structural constraints. There have been a few studies in literature towards developing methods to identify these functional sites. Studies conducted by Voigt and coworkers [72, 78] propose that contact between residue pairs can be considered to represent interaction between them. Therefore, residues that are in contact with many other residues participate in numerous interactions and hence are unlikely to tolerate uncompensated mutations. While this is an interesting hypothesis, many studies suggest that even distant residues may interact strongly through a network of related interactions and/or through electrostatic forces [79-82]. Lichtarge and Sowa [83] proposed an alternative approach in which they identify functional sites by mapping tertiary structures of sequences that form the nodes near the root of the evolutionary tree. Spatially clustered residues are assumed to be functionally important since changes in the amino acid composition of these regions are linked with evolutionary divergences and, hence, functional specificity. Similarly, work by Landgraf and coworkers [84] identified residues with conserved structural neighbors and residue clusters that have high sequence similarity. Both studies, however, provide little insight into which residue pairs are

interacting. Moreover, many of the variable, functional loop regions may not map onto the corresponding loop of other members of the family on the static molecule, but may have related motion during the catalytic cycle thereby playing important role in ligand binding and dissociation.

Studies have shown [85, 86] that changes in protein properties are brought about by cumulative effects of many small adjustments, many of which are propagated over significant distances in the three-dimensional structure. Trace evidence of such coordinated mutations brought about by evolution are present in the protein sequence data of the members of a protein family. It has been postulated that a substitution at one position is compensated by a substitution elsewhere in the sequence to ensure that structural features essential for the functioning of the protein are conserved [85, 87-89]. In fact, studies have revealed that residues distant in sequence but near in three-dimensional space undergo simultaneous compensatory variation to conserve the overall physicochemical properties [86, 87, 89-92]. This hypothesis has been used to predict residue contact maps by identifying correlated mutations [88, 93, 94] whose signals are strengthened by comparing neighboring nodes in the phylogenetic tree [95]. Despite the compelling logic behind this hypothesis, these studies have met with only a limited success. The estimated statistical contact prediction has at best been only 15-20% accurate [96].

We hypothesize that strongly correlated residue pairs do not necessarily have to be in contact; rather, they may affect the dynamics of protein function by participating in a network of distal motions involved in catalysis or by participating in important interactions. Generally, the motions of protein regions are associated with ligand

binding/dissociation involved in catalysis. Our hypothesis is based on the assumption that uncompensated mutations in these regions will disrupt their motion/interaction and therefore attenuate important chemical steps in catalysis, affecting reaction rates by several orders of magnitude [97, 98]. This leads us to postulate that regions that are involved in the dynamics of the reaction or in an uncommonly high number of interactions with other residues are likely to tolerate only coordinated mutations. For example, if a lysine in a loop region is replaced by a glutamine, it may be necessary to substitute a glutamine by a lysine at a position elsewhere so that the net charge remains the same, ensuring that no essential motions are disrupted due to charge repulsion (Figure 2.1). To detect these functionally important regions, we introduce the use of the *correlation tendency* metric to quantify the average number of other residues to which a particular residue/region is correlated. In this work we demonstrate that highly mobile regions of the protein exhibit high correlation tendency values. This occurs mainly due to the many additional physical and functional contacts (i.e., through hydrogen bond, van der Waals interactions and long-ranged electrostatic interactions) these regions make during their motion associated with catalysis [99].

To test our hypothesis, we performed residue correlation analyses (RCA) on three protein families: (i) dihydrofolate reductase (DHFR), (ii) cyclophilin and (iii) formyl-transferase. These families were chosen based on the differences in the degree of sequence alignment and conservation and availability of structural and functional data. In addition, we use our approach in a predictive fashion to identify important regions of a transmembrane amino acid transporter protein for which there is little structural and functional information available. The Pfam database [100] was accessed to download

protein family sequence data. The RCA is comprised of two major steps: (i) identifying pairs of positions whose mutations occur in a coordinated manner and (ii) using these results to identify protein regions that interact with an uncommonly high number of other residues.

Section 2.2: Residue Correlation Analysis (RCA)

Protein chemists discovered early on that certain residue substitutions commonly occur in homologous proteins from different species [101]. Because the protein retains its functionality after these substitutions, the substituted residues are either compatible with the protein structure and function, or else the effects of these substitutions are compensated by some other changes [86, 89-92, 102]. Since these substitutions are coordinated, there exists a measurable correlation between these mutation patterns [88, 93, 94]. However, measuring amino acid variability requires the use of a metric of similarity that will reflect how likely one residue is to be substituted by another. Numerous methods have been suggested such as utilizing physicochemical vectors describing residue physical properties (e.g., side chain volume [103], charge [94], hydrophobicity [104]) and similarity matrices that codify empirical information from phylogenetic trees (such as BLOSUM [105] and PAM [101]). PAM250, BLOSUM62 and McLachlan [106] scoring matrices were used in our study to compute the correlation coefficient. Results obtained for the three scoring systems were very similar (data not shown) and the one selected here is the McLachlan scoring matrix. Alternatively, for identifying the functional coupling of two positions of the MSA, Lockless and Ranganathan [107] used vectors of 20 binomial probabilities of individual amino acid frequencies. These probabilities are determined by the distribution of residues in each

column of the alignment. Clearly, the sequence alignment obtained by randomly shuffling the residues in each column of the original alignment would yield identical results even though the resulting sequences may be significantly different from the parental sequences. Furthermore, since no metric of similarity between residues has been utilized in the above method, less frequent but conservative substitution patterns will not be recognized. In another recent study by Larson *et al.* [48], correlation signals are identified based on the probability of occurrences of residue pairs rather than by use of scoring matrices. Clearly, some conservative substitutions (i.e., substitution of a residue with another that is very similar in physical and chemical properties) cannot be detected by this method. Hence, the use of similarity matrices has the advantage that it can detect conservative substitution patterns more accurately as compared to other methods.

Highly correlated pairs may arise due to: (i) physical contact between them, (ii) distal interaction, (iii) interaction through conformational changes or (iv) occurrence by chance. In our analysis, predictions made based on RCA are assumed to be correct if the residues of the correlated pair are interacting through conformational changes or through distal interactions. Inter-residue distances are calculated for identifying contacting residues. Various cutoffs have been proposed in the literature to define contacting pairs. Two residues are said to be in contact if the distance between them is below a given arbitrary threshold. These distances could be the distance between the two beta-carbons (C_β) [108-111] or the two alpha-carbons (C_α) [112] of the corresponding residues. The average of the distances between all the atoms of the two residues or the distance between the nearest atoms belonging to the side chain or the backbone of the two residues [113] have also been used in these definitions. Here, we consider a pair of

residues to be contacting if the C_β - C_β (or C_α - C_α in the absence of C_β) distance is less than 8 Å.

Section 2.2.1: Residue Correlation Coefficients

The family of aligned sequences obtained from the Pfam database is assumed to be a randomly chosen sample of a population of all functional protein sequences. Correlation coefficients between any two positions (Figure 2.2) are calculated similar to the method proposed by Gobel *et al.* [88]. For a given pair of sequences (k, l), each substitution at a position (i or j) is associated with a similarity score (X_{ikl} and X_{jkl} respectively) obtained from the McLachlan scoring matrix. The expression used for computing the correlation coefficient (r_{ij}) between two sequence positions (i, j) in the alignment is:

$$r_{ij} = \frac{2}{N(N-1)} \sum_{k=1}^{N-1} \sum_{l=k+1}^N \left(\frac{X_{ikl} - \langle X_i \rangle}{\sigma_i} \right) \left(\frac{X_{jkl} - \langle X_j \rangle}{\sigma_j} \right) \quad (1)$$

where σ_i and σ_j are the standard deviations of the scores X_{ikl} and X_{jkl} at positions i and j about their means $\langle X_i \rangle$ and $\langle X_j \rangle$ respectively. The use of weights in computing the correlation coefficient has, however, been avoided since they not only penalize genuinely correlated signal in a group of similar sequences but often cannot be quantified in a universal fashion. Studies also indicate that sequence weighing is not an important factor in achieving high accuracy in covariation signal [48]. To prevent the correlation coefficient from being biased by over represented groups of similar sequences, we eliminate combinations of pairs of sequences that have repeated patterns in the corresponding columns. For example, in computing the correlation coefficient between two positions i and j of the alignment shown in Figure 2.2, the k - m sequence combination is not considered. Note that the denominator of the equation 1 (i.e. $N(N-1)/2$) = the

number of combinations of sequences) is adjusted accordingly. Positions that have a high percentage of gaps (>70%) are omitted to avoid misleading results due to the small amount of data available at these alignment positions. Furthermore, the results depend on how well the sequences align, and hence a few lengthy (or too short) sequences are deleted from the alignment to avoid introduction of excessive gaps. The remaining sequences are then realigned with CLUSTALW [114, 115] using BLOSUM/PAM matrices.

Two positions are considered to be correlated if the absolute value of the correlation coefficient between the two is above a threshold value ($r_c = 0.4$) [88]. The significance of the cutoff value is tested by performing a correlation analysis on sequence alignment obtained by (i) randomly shuffling the residues of each row of the alignment (i.e., each sequence still retains the same residues) thereby destroying the existing secondary structure elements and (ii) randomly shuffling the residues in each column of the alignment. In either case it was observed that no residue pair yielded correlation coefficient even close let alone above the threshold value. Residue pairs with correlation coefficient above the threshold value are identified, that are then used in computing the correlation tendencies of protein regions as outlined below. These pairs are also investigated for contacts to estimate the percent of the correlated pairs that are contacting. Inter-residue distances less than 8 Å are considered to be contacting. For this purpose, the Euclidean distance between two residues ($|\mathbf{r}_i - \mathbf{r}_j|$) is calculated from the coordinates obtained from the PDB database.

Section 2.2.2: Correlation Tendencies

The residue correlation coefficient (r_{ij}) is a measure of the relationship between the

scores of two sequence alignment positions. However, it is more informative to identify protein regions (i.e., a contiguous string of residues) that show strong correlation with a relatively high number of other residues. Residues adjacent in the sequence are contacting, and therefore, identifying these signals provides no additional information. Thus, the correlation of the position under consideration with the adjacent three residues in the sequence is not taken into account. In general, the correlation signals are quite noisy and therefore it is difficult to glean useful information from it. The noise level in the correlation data is reduced by averaging out these effects over secondary structure elements to calculate the correlation tendencies. The correlation tendency (t_m) of a segment m is defined as the ratio (x_m) of the number of correlated pairs with at least one of the residues in region m to the total number of correlated pairs, that is scaled by the ratio of its length l_m to the total sequence length L :

$$t_m = \frac{x_m}{l_m / L} \quad (2)$$

Because each residue position of the segment is weighted by its frequency of occurrence in the correlated set (a set including all correlated residue pairs), the correlation tendency reflects the frequency of interactions these residues are engaged in.

The protein segments, for the purpose of calculating correlation tendency values, are determined based on its secondary structure. Sequence alignment with no existing secondary structure elements (i.e., alignment obtained by randomly shuffling the columns of the original MSA without disrupting the relative order of the residues in the column) yielded t_m values close to 1. Hence, correlation tendency value greater than one is considered to be significant. Regions with t_m values higher than 1 are identified and located on the three-dimensional structure of the corresponding protein molecule. These

are then investigated for functional roles known from experimental and molecular dynamics studies.

Section 2.2.3: Site Entropy

Highly conserved positions do not carry correlation information and are not considered in the correlation analysis, but they do contain useful information with respect to functionality. Hence, in addition to identifying strongly correlated pairs it is important to measure variability at residue sites to identify conserved regions. A widely used measure of site variability is the site entropy (S_i) that is calculated using the expression:

$$S_i = -\sum_{a=1}^{20} p_a \log_2 p_a \quad (3)$$

where p_a is the probability of occurrence of an amino acid a in the column i of the aligned sequences. Domain entropy, S_m is derived by averaging positional entropy of residues in the domain m .

Entropy and correlation capture different statistical properties of family sequence data; therefore, we investigated whether correlation analysis captures information not accessible by simple residue variability measures such as entropy, and whether the two can be used in conjunction to better predict functional domains. Prediction of functional sites are made based on RCA ($t_m > 1$) and entropic measures ($S_m < \text{average sequence entropy}$). Functional sites are also identified using contacting residues as a representation all interacting residues [72]. For this purpose, a similar analysis was performed using contacting residue as was done for correlated residues in calculating correlation tendency. The accuracy of prediction of functional sites by these methods is expressed in terms of *sensitivity* (the fraction of true functional sites identified by prediction) and *specificity* (fraction of the predicted domains that form the functional sites) that are presented in

Table 2.1. These are compared to the specificity of identifying a functional site by random choice of a protein domain (the ratio of the number of functional domains (included as secondary structure elements) to the total number of secondary structure elements present in the protein sequence) that are also included in the same table. A detailed description of the results for the three protein families is presented next.

Section 2.3: Computational Results and Comparisons

Correlation analyses are performed on the protein families of dihydrofolate reductase (DHFR), cyclophilin and formyl-transferase. These families include protein members that have different levels of alignment and conservation. Cyclophilin and DHFR families include sequences that are closely related (maximum average tree distance of 16.54 and 25.8 respectively) and therefore align fairly well, whereas the formyl-transferase family (maximum average tree distance of 30.89) contains distant sequences that do not align well. In addition, unlike the DHFR (average entropy = 2.300) and formyl-transferase (average entropy = 2.563) families, the cyclophilin family is much more conserved (average entropy = 1.908). In comparison to the cyclophilin and formyl-transferase families, much more is known about DHFR and its functional domains.

Section 2.3.1: Dihydrofolate Reductase (DHFR)

Dihydrofolate reductase (DHFR) is an enzyme that is necessary for maintaining intracellular levels of tetrahydrofolate, an active form of the vitamin folic acid, and an essential cofactor in the synthetic pathway of purines, pyrimidines and several amino acids. It catalyzes the reduction of 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF) using nicotinamide adenine dinucleotide phosphate (NADPH) as a cofactor. X-ray crystallographic studies indicate that the members of the DHFR enzyme contain an eight-

stranded β -sheet and four α -helices interspersed with loop regions that connect these secondary structures (Figure 2.4b). Analysis of the DHFR complex with folate has revealed that isolated residues exhibit diverse backbone fluctuations on the nanosecond to picoseconds time scale. The most significant motions are observed in the M20 loop (residues 7-24), the neighboring FG loop (residues 116-132), the GH loop (residues 142-150) [82, 97, 116, 117], the distant CD loop (residues 64-71) and the hinge region connecting the two subdomains (residues 87-91) [82, 118, 119] (see Figure 2.4b). Motions detected in the region between residues 40 and 80 are strongly anti-correlated to the fluctuations in the M20 and FG loops [82]. Fluctuations in these loops play crucial roles in the catalytic pathway; for example, conformational changes in the M20 loop may limit the rate of THF dissociation. Mutational studies reveal that only specific residue substitutions are permitted in these loops. The replacement of four M20 loop residues with a glycine results in a 500-fold decrease in the rate of hydride transfer, and similar effects are observed for mutations in the FG loop approximately 17 Å from the active site [97, 120]. In addition, mutations in the NADPH (residues 42, 60) and DHF/THF (27, 113) binding regions have drastic effects.

The RCA analysis of the DHFR family includes 122 sequences accessed from the Pfam database. The scatter plot (Figure 2.3a) between r_{ij} values and C_β - C_β distances in Angstroms outlines the proximity of the correlated pairs in three-dimensional space. *Clearly most of the correlated pairs are not contacting and many of the contacting pairs are not correlated.* Of the 105 correlated pairs identified ($r_c = 0.4$) only 9.52% of them are contacting whereas contact by random choice of pairs has a likelihood of 3.4% alone. Figure 2.3b shows the average correlation coefficient between the residues of various

secondary structure elements. It has been observed that strong correlation exists between functionally important regions. Particularly the strongest correlation signals (shown in circles in Figure 2.3b) are detected between the two hinge regions, the CD loop and the GH loop. Correlation tendency values for different segments (based on secondary structure) were calculated as described earlier. A cutoff value (r_c) of 0.4 resulted in a high specificity of 90% whereas the corresponding sensitivity is 81.8%. Of the 20 segments into which the DHFR enzyme is divided, 11 are functionally important resulting in the likelihood of only 55% that a randomly chosen region (secondary structure) is functionally important. Our study shows that t_m values are greater than one for almost all of the mobile loop regions (except for the FG loop) while the converse holds for regions outside these loops (Figure 2.4(a) and (c)). High positional entropy is observed for two important hinge regions (36-38 and 87-91) and for the CD and GH loops indicating that these regions are not conserved. Correlation tendency values, however, show that residue changes at these positions are highly coordinated. The entropy calculations (Figure 2.4a) result in a specificity of 71.4% and a sensitivity of only 45%. Evidently, entropy alone does not capture functionality information clearly discernible with the correlation analysis. Furthermore, most of the t_m values in the region between residues 40 and 80 are greater than one with an overall average of 1.22. Agreement with these results suggests that the proposed correlation analysis is indeed capturing information related to distal motions during catalysis. Low entropy and high correlation tendency values are observed for residues 91-96 and residues 40-60 indicating that these regions are fairly conserved and limited changes at these positions are coordinated (Figure 2.4a). Interestingly, even though functional roles of the residues 91-96 are unknown, residues 40-60 (a subset of

the region 40-80) have been observed to fluctuate during catalysis [82]. A comparison of prediction results obtained by RCA, entropy measures, contacting pairs and random choice are summarized in Table 2.1.

A number of DHFR studies have delineated the role that specific residue pairs play during catalysis [82, 92, 97, 117]. Table 2.2 contrasts these results with a few of the pairs that are identified with the correlation analysis. Remarkably, for most of these pairs we find exceptionally high correlation coefficients, further strengthening the hypothesis that important information regarding function can be recovered from protein family sequence data through residue correlation analysis.

Section 2.3.2: Cyclophilin

Cyclophilin is a binding protein for the immuno-suppressive drug cyclosporin and also an enzyme with cis-trans isomerase activity. It catalyzes the interconversion between cis and trans conformations of X-Pro peptide bonds, where “X” could be any amino acid. Studies have indicated that internal protein dynamics are intimately connected to enzyme catalysis that influences the substrate turnover [121]. As in the case of DHFR, rapid fluctuations are observed in the loop regions of cyclophilin during catalysis. Significant conformational exchange dynamics were observed in the residue regions 54-56 and the loops 65-80 and 101-110 [121] as shown in Figure 2.6b. Furthermore, a narrow pass separates the two loops that provide a possible location of the extended substrate binding [122]. Studies indicate that residues L98 and S99, during the catalytic cycle, interact with the trans peptide while the cis isomer binds near residues 55, 82, 101-103 and 109 [121, 123].

For the RCA analysis of cyclophilin, 304 sequences were downloaded from the Pfam database. Correlation coefficients (r_{ij}) for all pairs are calculated as described

earlier and are plotted against C_β - C_β distances as shown in Figure 2.5. A cutoff of 0.4 is chosen to identify strongly correlated residues, resulting in the selection of 310 pairs as members of the correlated set of which only 12.16% are contacting. The correlation tendency plot (Figure 2.6a) indicates that of the three mobile regions mentioned above, two have t_m values significantly higher than one. This results in a high specificity of 33% and a sensitivity of 50%, whereas the likelihood of identifying important regions based on random choice is only 17.4%. These predictions are in good agreement (see Figure 2.6 (b) and (c)) with the results obtained through NMR relaxation experiments conducted by Eisenmesser and coworkers [121]. The motions in the loops are associated with cis isomer binding and therefore also include residues necessary for interacting with the cis isomer. The average entropy of the loop 65-95, where the most prominent motion is observed, is very close to the overall average clearly indicating that the loop region is not highly conserved. Residues 98 and 99, even though they are functionally important, did not show high t_m values. However, these positions are well conserved as low entropy values are observed at these positions (Figure 2.6a). In addition, regions 4-11, 42-45 and 143-146 show coordinated mutations resulting in correlation tendency values greater than one suggesting that they may be functionally important and hence require further investigation. Table 2.1 summarizes the statistical analyses carried out for the cyclophilin family.

Section 2.3.3: Formyl-Transferase

Glycinamide ribonucleotide transformylase (GART) catalyzes the transfer of a formyl group from 10-formyltetrahydrofolate to glycineamide ribonucleotide (GAR), a reaction in the purine biosynthetic pathway. The GART structure can be subdivided into

two sub-domains with the N-terminal domain consisting of a central core of smoothly twisted, parallel β -sheet of four strands surrounded on both sides by two pairs of α -helices (Figure 2.8b). The C-terminal consists of the remaining six β -sheets with a long α -helix [124, 125]. X-ray structure analysis of a ternary complex with the substrate GAR has confirmed that the phosphate group of GAR is tightly bound by the loop consisting of residues 10-13 [125, 126]. The terminal three oxygen atoms of the phosphate interact primarily with the main-chain NH groups of residues 11, 12 and 13, while the fourth phosphate oxygen is within hydrogen-binding distance of the NH of the Gln-170 side chain [124, 125]. The ribose hydroxyl group interacts with residues in the α -helix (162-185), while the rest of the sugar lies in a hydrophobic pocket formed by residues 86, 88, 107, 121, 166 and 171. The key residues in the active site are 103, 106, 108 and 144, and these are highly conserved as reflected by their average entropy shown in Figure 2.8a. Study of a high-resolution structure of GART with multisubstrate adduct has revealed that residues 112-132, 140-145 and 155-175 are highly mobile [126] (see Figure 2.8b). Mutations at position 144, a part of the loop with highest mobility and also a part of the active site, resulted in an enzyme that was 10^4 times less active than the wild type [127]. The folate derivative binds in the hydrophobic pocket formed by the residues 85, 88, 92, 197, 104 and 139 where it forms six hydrogen bonds interacting with residues 90, 92, 140, 141 and 144 [126, 128].

The formyl-transferase family includes the following members: (i) GART, (ii) formyltetrahydrofolate deformylase and (iii) methionyl-tRNA formyl-transferase. In total, 169 sequences were downloaded from the Pfam database of which seven sequences had large insertions and, hence, were removed. The MSA includes the first 181 residues

of the reference sequence (glycinamide ribonucleotide transformylase from *Escherichia coli*; PDB id: 1GRC). As observed in the previous two cases, most of the correlated pairs in formyl-transferase are not contacting (Figure 2.7). A cutoff value of 0.4 captures 207 pairs as members of the correlated set that results in a high specificity of 55.6%. Note that the corresponding specificity for entropy measures and contacting pairs are only 50% and 37.5% respectively. Figure 2.8a clearly shows that all the regions mentioned above including residues 41-51 and 132-138, exhibit high correlation tendency values, with the only exception being residues 112-132. However, entropy values show that the loop 112-132 includes residues that are highly conserved, but low correlation tendency indicates that most mutations observed in this region are not coordinated. Residues 132-138, similar to the loop 112-132, are fairly conserved but are also involved in numerous interactions as suggested by the observed high correlation tendency. Comparisons between prediction results obtained by various methods are shown in Table 2.1. The sensitivity value obtained by domain entropy is 100% compared to 83.3% of that of RCA, while contacting pairs perform poorly with a value of only 50%. Though most of the functionally important regions (especially the residues 8-13, 86-93, 103-108, 139-147 and 162-181) show low entropy, they are also discernible by RCA further strengthening the point that correlation analysis does identify regions with mutations that are highly coordinated.

Section 2.3.4: Transmembrane amino acid transporter protein

The common elements of the three protein families for which we carried out the RCA analyses is that they all have substantial structural and functional information available. Here we consider a protein family for which there is very limited structural and

functional data. The transmembrane amino acid transporter protein family (further referred to as the permease family) mainly includes proline and amino acid permeases that are integral membrane proteins involved in the transport of amino acids into the cell. The multiple sequence alignment obtained from the Pfam database consisted of 167 sequences with the reference sequence (in our study) as the amino acid permease (PID: g15237493). The total length of the sequence is 446 residues, however, the sequence alignment part includes only residues 34-437.

Correlation tendency values are calculated for segments based on predicted secondary structure. The widely used methods for protein structure prediction based on neural network methods (HNN and PROF [129], accessible at: <http://www.expasy.ch/>) are utilized. The correlation tendency values along with the average entropy for different regions are shown in Figure 2.9. A high degree of conservation is observed at positions 51-65, 233-239, 283-291 and 370-381. Residue 64, which is in the conserved region and adjacent to the completely conserved glycine (residue 65), shows a relatively high entropy. Extremely high correlation tendency and correlation coefficient values corresponding to residue 64 suggests that mutation at this position is largely coordinated with other mutations. Similar to residue 64, positions 96-98, 114, 154, 194, 195, 225, 226, 247-267 and 304-307 include regions that are strongly correlated to a fairly large number of other residues resulting in high correlation tendency values. Nevertheless positional entropy values show that these regions are not highly conserved, implying that possible mutations need to be coordinated thus requiring other adjustments. It has been observed that almost all of the regions with high correlation tendency, as in the other three cases, lie in the predicted loop regions that are likely to be crucial to protein

dynamics during catalysis.

Section 2.4: Summary and Discussion

In this work, a computational framework for identifying protein regions that are correlated to a large number of other residues was proposed. Residue correlation analyses (RCA) were performed on three protein families: (i) DHFR, (ii) cyclophilin and (iii) formyl-transferase. It was shown that residues/regions that have a high correlation tendency are either involved in important interactions or participate in conformational changes necessary for retaining function. RCA and entropy calculations were used to identify the protein building blocks that co-evolve under structural and functional constraints. Predictions of these methods were tested against experimental and molecular dynamics data available in the literature. Table 2.1 summarizes prediction results obtained by RCA, entropic measures, contacting pairs and random choice. The DHFR case study demonstrated that entropy alone cannot capture functionality information clearly discernible with RCA. However, in the other two cases, sensitivity values achieved by the entropy measures are 100%. A close examination of these two cases reveals that functionally important regions not found by RCA are those that have relatively low entropy, leading to a weakened correlation signal not detected by RCA. No clear relationship between entropy and correlation tendency was detected. It was observed that many of the conserved functional sites have high correlation tendency values reflecting the conservative nature of mutation patterns in these regions. However, we also found that a number of regions, particularly the functional loops that are highly variable, with high correlation tendency values. Contacting pairs captured less information about important interactions than the RCA approach and entropic measures

for all three protein families. This confirms that strongly correlated pairs and *not* contacting pairs is a better descriptor for identifying interacting residues. Combined results of the RCA and entropy measures identified all of the functionally important regions (except for the FG loop in the DHFR family) in the three families. Hence, RCA and entropic measures collectively form a better technique to identify functionally important regions.

In the current implementation, the correlation analysis examines the relationship among the leaves of the phylogenetic tree. However, one would expect that correlation signals would be stronger between the nodes of the tree at shorter evolutionary distances, and a number of groups have observed this while attempting to detect contacting residues [85, 95]. Future work in this area will involve developing methods for further refining the correlation calculations presented here by measuring signals at the tree nodes. Specifically, from the phylogenetic trees one can measure evolutionary distances between various sequences. From the comparison of the correlation signals between sequences of equal evolutionary distances, one can more precisely determine the effect of these distances on the compensatory covariation signals. This approach might be used to ‘filter’ correlation signals and increase our ability to detect it above the ‘noise’ observed in the divergent evolution of protein sequences.

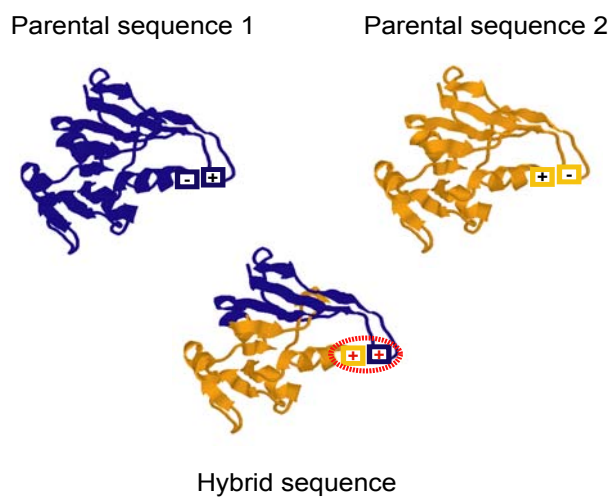


Figure 2.1: Formation of a repulsive ion pair in a recombinant hybrid that may disrupt contacting pairs as well as essential motions.

	i	j
1	MISAAALAV--D---RVMAMP---WN--LPADTLNKP-	
2	NISLANELI-T---RAGKLP---WQF-IKEDMENSV-	
:	SLNMAVVK--T---GGNQIP---WH--EPEDTMNSV-	
k	KLSLAISK--N---GVIDIP---WS--AKGETYNQW-	
m	KISLATSE--N---GVIDIP---WS--AKGETYNQW-	
:	KLSLAISK--N---GVIDIP---WS--AKGETYNQW-	
:	RIYLMGA--N---RVIDIP---WK--IPGETESKV-	
l	ELHAATA--N---GCIALP---WPP-LKGD TMGKV-	
:	KVSLMKAK--N---GVIHIP---WS--AKGENQW---	
:	-TAFLQDR--D---GLIHLP---WH--LPDQTVGKI-	
N	RFVLVVAD--N---RVI TMP---WH--LPETTTGHP-	

Figure 2.2: The scores X_{ikl} and X_{jkl} are obtained from one of the similarity matrices (PAM250, BLOSUM62 or McLachlan [106]) for positions i, j corresponding to the residues at these positions in the sequences k and l . These residues (i, j) are reported to be correlated if correlation coefficient value (r_{ij}) is above a threshold value ($r_c = 0.4$).

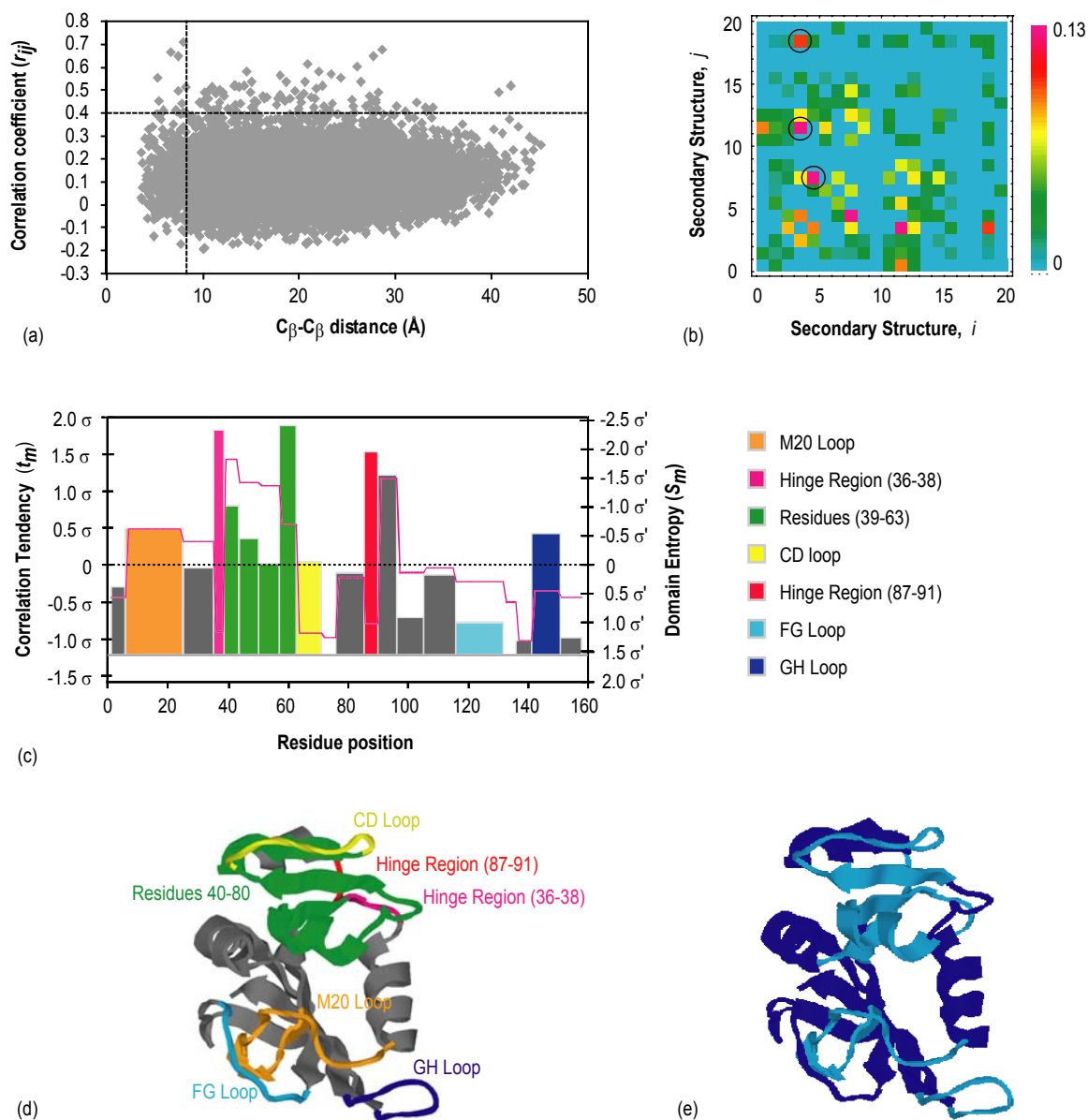


Figure 2.3: (a) Plot of residue correlation coefficients (r_{ij}) versus C_β - C_β distances (calculated from the crystal structure of 1DDR) for pairs of residues in the DHFR enzyme. The vertical broken line partitions pairs that are in physical contact (i.e., inter-residue distance less than 8 Å), while the horizontal broken line indicates the cutoff value ($r_c = 0.4$) above which the pairs are considered to be strongly correlated. Note many of the correlated pairs are not contacting and many of the contacting pairs are not correlated.

(b) Plot of the average correlation coefficient between residues of various secondary structure elements of DHFR. Strongest correlation signals are detected between the two hinge regions, the CD loop and the GH loop (shown as circled regions). **(c)** The correlation tendency (at $r_c = 0.4$) for different segments of the DHFR enzyme based on its secondary structure is plotted against residue position. The zero line in the figure indicates the average correlation tendency (1) and the entropy (2.3). The colored bars represent correlation tendency values for different regions that include residues involved in motions during catalysis or important interactions as found through experimental and molecular dynamics simulation studies. The graph also shows the average entropy (red line) for these domains on the secondary axis. Note the values are reversed on the secondary axis (i.e., peaks are conserved regions) for easy comparison with results obtained through RCA. The two axes are scaled based on their standard deviations ($\sigma = 0.8$, $\sigma' = 0.62$) about their means (1, 2.3 respectively). **(d)** Regions in the DHFR enzyme that are functionally important (as known from experimental and simulation studies) are highlighted in color. **(e)** Light blue regions correspond to regions having t_m values > 1 , whereas dark blue regions imply lower t_m values indicating less than average participation in correlated set. (PDB id: 1DDR).

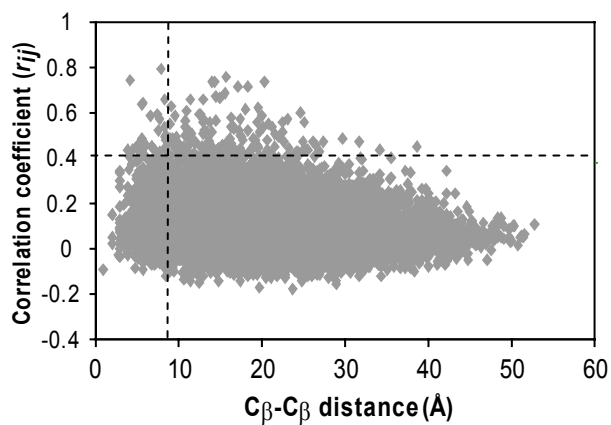


Figure 2.4: Plot of residue correlation coefficients (r_{ij}) versus C_β - C_β distances (calculated from the crystal structure of 1RMH) for pairs of residues in the cyclophilin protein. The vertical broken line partition pairs that are in physical contact (i.e., C_β - C_β distances less than 8 Å), while the horizontal broken line indicates the cutoff value ($r_c = 0.4$) above which the pairs are considered to be strongly correlated.

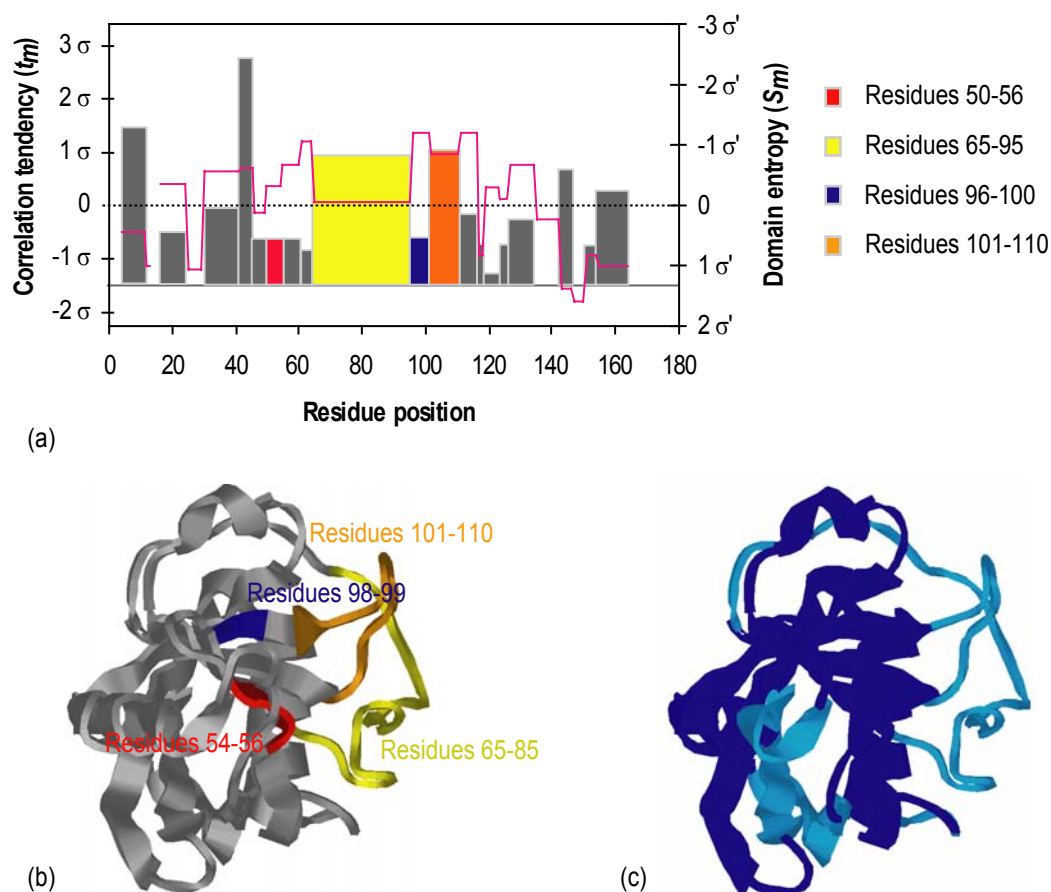


Figure 2.5: (a) The correlation tendency plot (at $r_c = 0.4$) for the cyclophilin enzyme based on its secondary structure. The zero line in the figure indicates the average value for both the correlation tendency (1) as well as the domain entropy (1.908). The correlation tendencies of functionally important regions are represented by colored bars. The red line corresponds to the domain entropy shown on the secondary axis. As in the case of DHFR, the secondary axis is reversed to represent highly conserved regions as peaks. The two axes are scaled based on their standard deviations ($\sigma = 0.68$, $\sigma' = 0.73$). (b) Loop regions in cyclophilin protein that are in motion during catalysis (as known from experimental studies) are highlighted in color. (c) Light blue regions identify domains with t_m values > 1 , whereas dark blue regions imply lower t_m values. (PDB id: 1RMH).

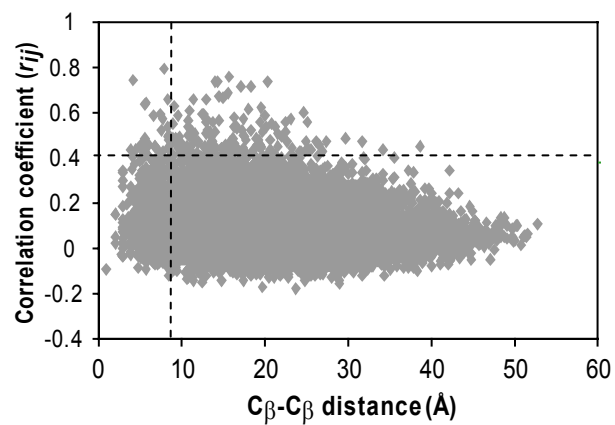
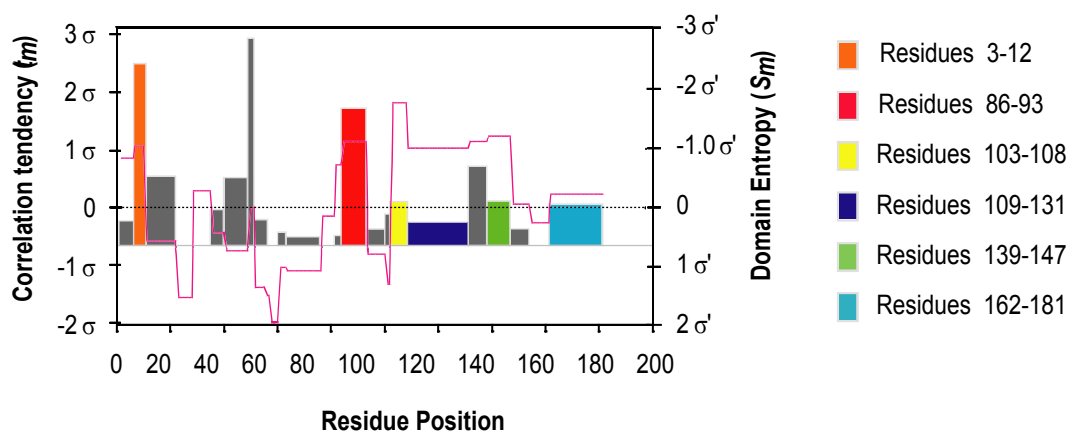
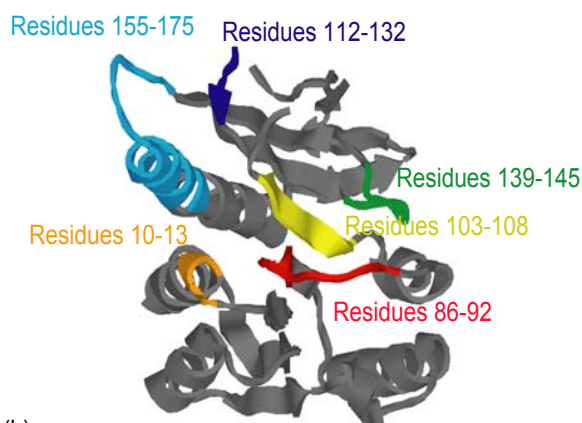


Figure 2.6: Plot of correlation coefficient versus C_β - C_β distance (calculated from the crystal structure of 1GRC) for pairs of residues in the formyl-transferase protein family. The vertical broken line partitions pairs that are in physical contact, while the horizontal broken line indicates the cutoff value ($r_c = 0.4$).



(a)



(b)



(c)

Figure 2.7: (a) The correlation tendency plot (at $r_c = 0.4$) of the formyl transferase family based on its secondary structure. The zero line in the figure represents both the average correlation tendency (1) and the average entropy (2.563). The correlation tendencies of functionally important regions (identified through experimental and simulation data) are shown as colored bars, while the red line corresponds to the domain entropy shown on the secondary axis. Values on the secondary axis are in reverse order for easy comparison between the results obtained from RCA and entropy measurements. The two axes are scaled based on their standard deviations ($\sigma = 1.42$, $\sigma' = 0.54$). (b) Functionally

important regions of formyl-transferase (known from experimental and simulation studies) are highlighted in color. **(c)** Light blue regions correspond to regions with t_m values > 1 , whereas dark blue regions have lower t_m values. (PDB id: 1GRC).

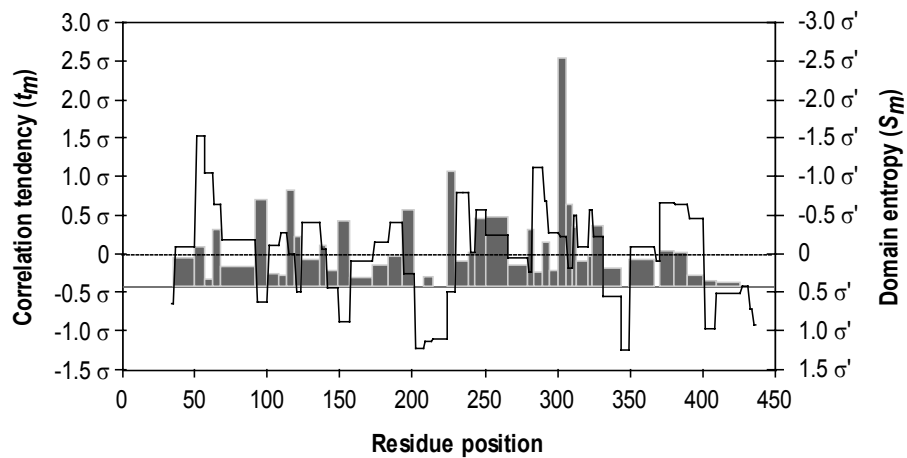


Figure 2.8: The correlation tendency plot (at $r_c = 0.4$) for different segments of the transmembrane amino acid transporter protein based on the predicted secondary structure. The zero line in the figure corresponds to the average value for correlation tendency (1) as well as for the domain entropy (2.843). The black line corresponds to the domain entropy shown on the secondary axis. As in the other cases, values on the secondary axis are in reverse order. The two axes are scaled based on their standard deviations ($\sigma = 1.26$, $\sigma' = 0.61$).

Table 2.1: Summary of statistical analyses for the three protein families.

Protein Family	Specificity				Sensitivity		
	RCA ^a	Entropy ^b	Contacting Pairs ^c	Random ^d	RCA	Entropy	Contacting Pairs
DHFR	90	71	50	55	82	45	36
Cyclophilin	33	31	25	17	50	100	50
Formyl-transferase	56	50	38	24	83	100	50

Results are shown based on ^aRCA: $r_{ij} > r_c$, $t_m > 1$; ^bdomain entropy; ^ccontacting pairs: average number of contacts per residue for a domain > overall residue average and ^drandom choice of domains. All values are shown as percentages.

Table 2.2: Residue pairs, their role in catalysis and correlation coefficients.

13-121	The residues in this pair belong to the M20 and the FG loops which are hydrogen bonded to each other. Mutation in these loops affects catalytic rate by 400 fold.	0.710
53-104	Mutation of this pair diminishes the rate by a factor of six or more. It lines the active site, implying that mutation alters the binding site geometry.	0.649
60-42	These amino acids are involved in hydrogen bonding with NADPH.	0.646
42-113	Residues 113 and 27 are hydrogen bonded with DHF, and residues 42 and 60 are hydrogen bonded to NADPH.	0.616
59-113		0.605
60-113		0.535
28-42		0.494
21-122	A hydrogen bond between them stabilizes the closed conformation of the M20 loop. Conformation changes of the M20 loop regulate ligand binding.	0.518

Chapter 3: Using a Residue Clash Map to Functionally Characterize Protein Recombination Hybrids

Section 3.1: Background

Directed evolution is a strategy for improving a specific biological function (thermostability, stereoselectivity, catalytic activity, expanded substrate specificity) through genetic diversification and selection, emulating natural evolution in an accelerated and guided fashion [130]. The diversity generating mechanism commonly entails the exchange of parental DNA fragments in the reassembled sequences through recombination and/or involves altered residue sites through random mutagenesis. One of the key challenges in the use of such directed evolution techniques for protein engineering is that in some cases, particularly when the parental sequences share low sequence identity, the reassembled sequences do not even fold properly and thus are non-functional. Moreover, it has been observed experimentally that the lower the sequence identity between the recombined parental sequences, the larger the proportion of the library that is not functional [131]. The majority of the DNA-shuffling methods can only recombine closely related sequences and generate crossovers only within regions of high (i.e., > 60%) sequence identity. However, with the advent of more versatile techniques such as ITCHY [4], SCRATCHY [5], SHIPREC [6] and SISDC [36] greater diversity can be created by recombining distant homologs. This unfortunately often leads to an increasingly large proportion of the combinatorial library being non-functional. In an earlier work [132], we showed that functionally important protein regions are not necessarily conserved and instead found that they are more likely to exhibit strongly correlated substitution patterns with other regions. Moore and Maranas (2003) utilized

the energetics of molecular interactions to identify residue-residue clashes using second-order mean field calculations and found that most of these clashes could be attributed to steric, charge or hydrogen bond disruptions in the hybrids. In this work, we directly look for these types of clashes based on protein sequence data and compare these predictions against sequence data obtained for functional recombinant libraries.

A number of hypotheses have been proposed [133, 134] to explain why functional crossovers are not randomly distributed along the sequence but rather form distinct patterns. One of the most recent methods, the SCHEMA algorithm [135], postulates that crossover patterns resulting in hybrids with a large number of contacting residue pairs originating from the same parental sequences are more likely to retain their functionality. The key idea here is that each contact is a representation of favorable interaction between the two residues. Thus by retaining these contacting residues in the hybrids, one retains the favorable interactions that exist in the parental sequences. This interesting approach has led to a number of successful predictions [36, 136]. One potential shortcoming, however, is that it cannot differentiate between hybrids with different directionality (i.e., an A-B versus a B-A crossover), which often have substantially different functionalities [5, 137]. Here, we rethink the effect of having contacting residue pairs with different parental origins. Instead of always counting them as unfavorable, we view such pairs as places where clashes may or may not occur between the contacting residues. This view allows us to reestablish “context” in the interaction between the residue pair and thus capture the effect of crossover directionality (e.g., A-B versus a B-A crossover) on function. Specifically, motivated by the results of Moore and Maranas (2003), we explore three out of the many different mechanisms that

may render a contacting residue pair detrimental to the ability of the hybrid to fold properly (i.e., stability) and thus retain its functionality: (i) introduction of repulsive residue pairs such as ++ or --, (ii) disruption of hydrogen bonds due to the formation of donor/donor or acceptor/acceptor pairs and (iii) generation of steric clashes or cavities. It is quite straightforward to show that upon recombination residue clashes such as the repulsive residue pairs, disrupted hydrogen bonds and steric clashes can be introduced due to reversed orientation of charged, acceptor/donor or bulky residue pairs (see Figure 3.1). Other forms of clashes, not considered here, include the disruption of important protein-specific interactions [138] such as metal binding motifs [139], the catalytic triad [140, 141] and a number of ligand binding sites [142, 143].

The proposed procedure extends the concept of a residue contact map [135] by relying on the construction of residue clash map (i.e., a plot representing all possible clashing residue pairs in the reassembled sequences) based on the properties of the pair of residues that are in contact and have different parental origins. Notably, we find that the pattern of clashing residue pairs is greatly dependent on the crossover directionality. By superimposing these predicted clashing residue pairs against functional crossover statistics available in the literature we find that these clashes are preferentially avoided in the hybrids with %ACC (percent of avoided calculated clashes) ranging from 61% to 100%. Note that here we define %ACC as the percentage of predicted clashes that are avoided by all the functional hybrids available in the data set. In contrast, results obtained based on the residue contact map (i.e., a plot representing all non-conserved contacting residue pairs that have different parental origins) yielded %ACC ranging from 30% to 71% while the results from randomly generated clashes yielded %ACC ranging from 9%

to 54%.

Section 3.2: Method for Generating Clash Maps

Parental sequences participating in directed evolution, though sometimes highly divergent at the sequence level, share very similar structural traits. This implies that the basic structural characteristics have to be largely preserved in at least among the functional protein hybrids. These structural constraints enable us to construct the contact maps of the hybrids by simply querying the inter-residue distances calculated from the coordinates of the parental sequences obtained from the Protein Data Bank (PDB) [144]. Note that the contact map of a parental sequence is the list of all residue pairs whose β -carbons (C_β), or α -carbons in the absence of C_β , are within a cutoff distance of 8 Å [145]. These contacting residue positions are adjusted according to the structural alignment between the two parental sequences using the Combinatorial Extension (CE) method [146]. Next, the contact map of the hybrid is generated by retaining only those contacting residue positions that are common to the contact maps of both parental sequences. Pairs of contacting residue positions with at least one residue conserved in both parental sequences are excluded since the corresponding residue pair in the hybrid will always be present at these positions in at least one of the parental sequences. In cases where there is no structural data for a particular parental sequence, a predicted structure is used for identifying contacting residue pairs. The predicted structure is inferred using Swiss-Model [147] and a homologous structure as the template. This homologous structure is obtained either from the ExpDB database (http://www.expasy.org/swissmod/SM_Check_ExpDB.html) or using a BLAST search on the PDB [148] to find the nearest match. In all cases described in this study, the template and the parental sequence whose

structure is modeled share a relatively high sequence identity (>60%). It has been reported that predicted structures modeled using templates with such a high sequence identity are fairly reliable [147]. The Swiss-Model protein modeling server uses the template as an initial structure and replaces the template structure side-chains with side-chain conformations selected from a backbone-dependant rotamer library. These selections are made using a scoring function trading off favorable interactions such as hydrogen bonds, disulfide bridges and unfavorable close contacts. Side-chain placement in the protein structure is fine-tuned through a steepest descent energy minimization algorithm using the GROMOS96 force field [149]. Next, the contact maps of the hybrids generated as described above are investigated for clashes based on the three mechanisms (i.e., electrostatic repulsion, steric clashes and hydrogen bond disruptions).

Section 3.2.1: Repulsive residue pairs

Residue pairs found in the contact map of the hybrids are screened for $+/+$ or $-/-$ charge contacts that may be brought about by recombination (see Figure 3.1a). A contacting pair that has a repulsive residue pair ($+/+$ or $-/-$) at these positions in either of the parental sequences is not counted since they evidently do not seem to disrupt functionality. Note that the crossover directionality is automatically accounted since charge repulsion may be generated between residue pairs in one hybrid but not necessarily in the hybrid that has the reverse directionality (see Figure 3.1a). For example, parental contacting residue pairs with a single charged residue ($n/+$ and $+/n$) may form upon recombination either a neutral pair (n/n) or a repulsive residue pair ($+/+$) depending on the directionality of the crossover. Also, lysine and arginine are considered to be positively charged and glutamate and aspartate as negatively charged.

Section 3.2.2: Steric hindrance or cavity formation in the hybrids

A significant reduction in the total volume of a contacting residue pair is likely to give rise to a cavity formation whereas a corresponding increase may cause a steric hindrance. Figure 3.1b illustrates the effect of such volume changes as a consequence of the reversed orientation of large (residues A, D) and small (residues B, C) side-chains in the parental sequences. Cavity formation or steric hindrance is detected by observing whether the combined volume of the contacting residue pair in the resultant hybrid is much lower or higher than the mean combined volume (M) of the same contacting residue pairs in the parental sequences ($A+B, C+D$).

$$M = \frac{1}{2}[(V_A + V_B) + (V_C + V_D)] \quad (1)$$

Here V_k is the side-chain volume of residue k ($k = A, B, C, D$) in \AA^3 . Specifically, the scores S_{AD} and S_{CB} (for hybrids 1 and 2 shown in Figure 3.1b) are defined separately for hybrids with different crossover directionality as a measure of the deviation from M :

$$S_{AD} = \begin{cases} |(V_A + V_D) - (M + \Delta)|, & \text{if } V_A + V_D \geq M + \Delta \text{ (steric hindrance)} \\ |(V_A + V_D) - (M - \Delta)|, & \text{if } V_A + V_D < M - \Delta \text{ (cavity formation)} \end{cases} \quad (2)$$

A parameter ($\Delta = |(V_A + V_B) - (V_C + V_D)|$), which quantifies the extent of difference between the combined volumes of the two parental contacting residue pairs, is introduced into these scores to account for the tolerance of such volume changes. If the contacting residue pairs in both parental sequences are of similar size, they could lead to a small (even zero) value of Δ , thus, resulting in artificially inflated scores particularly in cases where the large and small residues have reversed orientation. Therefore, a lower bound is set on Δ equal to 10% of the mean (M):

$$\Delta = \begin{cases} |(V_A + V_B) - (V_C + V_D)| & \text{if } |(V_A + V_B) - (V_C + V_D)| \geq \frac{M}{10} \\ \frac{M}{10} & \text{if } |(V_A + V_B) - (V_C + V_D)| < \frac{M}{10} \end{cases} \quad (3)$$

In general, the core of most proteins have a higher packing fraction as compared to the surface [150]. This suggests that steric clashes are less likely to be tolerated in the protein core [151] as they often lead to packing defects [152, 153]. To account for the difference in the tolerance level for steric clashes at the protein surface and in the core, we set different cutoff scores S_c for contacting pairs. Cavity formation and steric hindrance in the core of the protein (i.e., accessible surface area of side chain $< 8 \text{ \AA}^2$) are considered to be significant if they score above a cutoff value, $S_c = 15 \text{ \AA}^3$; whereas only steric hindrance is considered with a cutoff value of 30 \AA^3 at the surface. The accessible surface area of a side chain is obtained by rolling a water probe of radius 1.4 \AA over the exposed surface. These calculations are performed using the WHATIF software package [154].

Section 3.2.3: Hydrogen Bond Disruption

Protein family members share many common hydrogen bonds, particularly those that are essential for functionality [155, 156]. Swapping the positions of the donor and acceptor groups of a hydrogen bond within a sequence preserves the hydrogen bond. However, similarly to volume and charge clashes, orientation reversals of the donor and acceptor groups in parental sequences leads to hybrids with donor-donor or acceptor-acceptor contacting pairs, thus disrupting the hydrogen bond between the two residues (Figure 3.1c). Note that hydrogen bonds between two backbone atoms are not of interest here since both the acceptor (CO) and donor (NH) groups are retained upon recombination. Here, we consider all possible cases (i.e., side-chain/backbone and side-

chain/side-chain) to identify potentially disrupted hydrogen bonds. The WHATIF software package [154] is used to detect common hydrogen bonds and identify the donor and acceptor groups of the parental sequences.

Contacting residue pairs identified for hybrids that violate at least one of the above three criteria (i.e., charge repulsion, steric hindrance and hydrogen bond disruption) are denoted as arcs (see Figure 3.2) linking the two residue positions. A crossover occurring between these two positions result in differing parental origins for the two contacting residues, connected by the arc, in the resulting hybrid. This representation of clashes is generalized for hybrids with multiple crossovers by using bicolored arcs to encode the specific directionality of the parental combination leading to a clash. We next examine the effectiveness of the proposed residue clash maps at explaining known functional crossover combinations for a number of protein systems.

Section 3.3: Comparison with Experimental Results

Residue clash maps are generated for the following five systems:

- (i) glycinamide ribonucleotide transformylase (GART) hybrids from *Escherichia coli* (purN) and human (hGART),
- (ii) human Mu class glutathione S-transferase (GST) M1-1 and M2-2,
- (iii) β -lactamase TEM-1 and PSE-4,
- (iv) catechol-2,3-oxygenase (C23O) *xylE* and *nahH* and
- (v) dioxygenases *todC1C2* (toluene dioxygenase), *tecA1A2* (tetrachlorobenzene dioxygenase) and *bhpA1A2* (biphenyl dioxygenase).

These systems vary considerably not only in terms of pairwise sequence identity and number of functional hybrids, but also in the directed evolution protocol used for

generating crossovers. All possible residue pairs with different parental origin that are brought in contact in one (or more) of the resultant hybrids are screened for all three forms of clashes. These clashes are then shown as arcs composing the residue clash map (see Figure 3.2). This representation is used for hybrids with a single crossover (GART) while a generalized representation (i.e., bicolored arcs) is used for hybrids with multiple crossovers (GST, β -lactamase, C23O and dioxygenases). A detailed comparison of the available experimental data using the proposed (i) residue clash map, (ii) residue contact map and (iii) randomly generated clashes is presented. Randomly generated clash map is constructed by randomly choosing an arbitrary number of pairs of non-conserved residue positions from the structural alignment. Note that conserved residue positions are not of interest here since they are also conserved in the hybrids and therefore will not form a clash. These results are examined in terms of %ACC (percent of avoided calculated clashes), defined as the percentage of the predicted clashes avoided by the functional hybrids present in the data set, and %CFC (percent of clash free crossovers), defined as the percentage of the observed functional crossovers that do not lead to any of the identified clashes. The %ACC of the randomly generated clash map is obtained by averaging these values over 100,000 such randomly generated samples. Alternatively, these values can be calculated as the ratio of all pairs of non-conserved residue positions that have residues at these positions in the functional hybrids that are both simultaneously retained from either one of the parental sequences to the total number of combinations of such residue pairs.

Section 3.3.1: Glycinamide ribonucleotide transformylase (GART)

In this case study we identify all clashing residue pairs for the two single-

crossover incremental truncation libraries encoding purN/hGART and hGART/purN hybrids. These hybrids are constructed using purN (209 residues) and hGART (201 residues) sequences whose structures (PDB id: 1GAR, 1MEO respectively) are obtained from the Protein Data Bank (PDB). Structural alignment of the two structures using the Combinatorial Extension method results in an RMSD (root mean square distance) value of 1.30 Å and a sequence identity of 38.20%. The residue clash map is constructed after identifying all common contacting residues based on the structural alignment. The purN/hGART library includes eight steric clashes (shown as gray arcs in Figure 3.3a) and five repulsive residue pairs (shown as black arcs), while the hGART/purN library exhibits nine steric clashes, three cases of charge repulsion and one hydrogen bond disruption (shown as a broken arc in Figure 3.3b).

Lutz *et al.* (2001) generated incremental truncation libraries with crossovers in the sequence window from residue 53 to 144. The functional characterization results are superimposed onto the residue clash map (see Figure 3.3) along with experimental count of each one of these hybrids. The purN/hGART library includes 68 functional members and as seen in Figure 3.3a most of the functional crossover positions avoid disrupting any arcs. Note that most functional crossovers fall in the regions between residues 79-114 and 120-138 that are free of any type of clashes. Out of 68 functional members present in the library, only four involve crossovers (i.e., positions 70 and 144) that disrupt any arcs (i.e., (4, 31)-80 and 140-145 respectively) resulting in 94.12% of functional members being free of predicted clashes (see Table 3.1). The hGART/purN library, on the other hand, includes 56 functional members with only one (i.e., crossover position 83) disrupting an arc (i.e., 81-84). Interestingly, most of the crossover positions (82%) in hGART/purN

library are found in the region 53-65 whereas none are observed in this region for the purN/hGART library (see Figure 3.3) alluding to the strong effect of crossover directionality. Note that crossovers generated using ITCHY are uniformly distributed over the desired truncation range [157] without exhibiting any directionality bias. Therefore, we believe that the crossover directionality in hGART/purN vs. purN/hGART in region 53-65 is not likely to be due to bias in library generation but rather an outcome of the selection pressure. By superimposing the residue clash maps on the corresponding functional hybrid libraries (i.e., purN/hGART and hGART/purN), we find that 61.54% of the predicted clashes are absent in the set of functional hybrids (%ACC) and 90.91% of the functional hybrids included none of our predicted residue clashes (%CFC). In contrast, comparison of the residue contact map and randomly distributed clashes against the functional library yield much smaller %ACC's of only 30.20% and 9.74% respectively.

Section 3.3.2: Glutathione S-transferase (GST)

The two GST parental sequences (i.e., human Mu class glutathione S-transferases, GST M1-1 and M2-2) share a relatively high sequence identity of 84% and align well both at the sequence and structural level. Both sequences are 217 residues in length, and have available structures (PDB id. 1GTU, 2GTU). Even though they share only a 16% difference in the sequence at the protein level, their specific activities with the substrate aminochrome and 2-cyano-1,3-dimethyl-1-nitrosoguanidine (cyanoDMNG) differ by more than 100-fold [158]. The chimeric GSTs in the experimental study were modified so that the first 32 bp (~10 amino acids) of each were from GST M1-1 (see Figure 3.4). The two segments vary only at two positions (i.e., 3 and 8) implying that the modified

DNA shuffled parental sequences have a slightly increased sequence identity of 85.25% at the protein level. The twenty functional hybrid sequences involving multiple crossovers [159] are shown in Figure 3.4 with gray denoting fragments retained from GST M1-1 and black denoting fragments from GST M2-2. All recombinant sequences have a number of identical stretches of undetermined parental origin, shown in white. The hybrids are listed in decreasing order of activities with respect to aminochrome and CDNB.

The residue clash map for the GST hybrids is modified to account for multiple crossovers (Figure 3.4). Each arc in Figure 3.4 is bicolored to encode the origin of the clashing residues. Therefore, only if the residues joined by an arc originate from the parental sequences with the same color designation as the arc, a clashing interaction is introduced. As shown in Figure 3.4, we find five cases of charge repulsion corresponding to pairs 91-96 (-/-), 93-91 (+/+), 128-125 (-/-), 129-125 (-/-) and 167-165 (+/+), with the first position retained from GST M1-1 and the second position from GST M2-2. The signs within the brackets indicate the type of interaction that is present in the hybrid. Steric clashes are found between residues 106-107 and 159-103 with the first entry of each pair originating from GST M1-1 and the second from GST M2-2. Comparison of our results with the 20 functional hybrids [159] show that most crossover positions in the functional hybrids lie outside the range where clashes are found (i.e., regions 1-90 and 170-217) (Figure 3.4). Interestingly, even though some crossovers exist between these arcs, their directionality is such that no clash is formed. None of the 20 hybrids contain any predicted clashing pairs resulting in a %ACC of 100%. Residue contact map based

and randomly distributed clashes yielded much lower %ACC values of 56.33% and 13.10% respectively (see Table 3.1).

Section 3.3.3: β -lactamases

Surprisingly, even though the sequence identity between the two β -lactamase parental sequences (PDB id: 1G68 (PSE-4) and 1BTL (TEM-1)) is 43.17%, slightly more than the GART system, the number of identified clashes is significantly higher. The total number of clashes in the TEM-1/PSE-4 directionality is found to be 27 while the reverse directionality involved 30 clashes (see Figure 3.5). Hybrids for both directions contained 14 cases of charge repulsion while the remaining clashes resulted from steric clashes. Crossover sequence data for functional hybrids are taken from the *in vitro* recombination experiments conducted by Voigt and coworkers (2002) where 10 functional hybrids (see Figure 3.5) are reported. These crossovers were generated between residue positions 26-290. Notably, by superimposing the residue clash map against the crossover distribution, we find that 80.70% of the predicted clashes share such directionalities so that they are not found in any of the functional members of the library. Figure 3.5 shows that most of the predicted clashes fall in the range between positions 25-125 and are present in only 4 out of the 19 functional crossovers. On the other hand, residues contact map and random clash distributions yielded much lower %ACC values of only 65.00% and 14.68% respectively (see Table 3.1). Recently, Hiraga and Arnold (2003) published additional crossover results for functional β -lactamase hybrids constructed using SISDC (sequence-independent site-directed chimeragenesis). These new data were also compared to the predicted clash map shown in Figure 3.5 and the results of these comparisons are summarized in Table 3.1.

Section 3.3.4: Catechol-2,3-oxygenase (C23O)

Kikuchi and coworkers [160] obtained seven thermally stable hybrids using single-stranded DNA shuffling on the parental sequences *xylE* (Catechol-2,3-dioxygenase from *Pseudomonas putida*, PDB id: 1MPY) and *nahH* (synthetic construct). Because no structure is currently available for *nahH*, we used an estimated structure obtained using Swiss-Model [161] with the structure of *nahH* (IMPY) as the template. This was subsequently used to obtain the structural alignment using the combinatorial extension method [146]. The two sequences share 84.7% sequence identity at the protein level. A total of 6 clashes are identified for both directions, all of which resulted from electrostatic repulsion (see Figure 3.6). Five of these have *xylE/nahH* directionality (79-80 (+/+), 82-83 (-/-), 183-184 (-/-), 183-286 (-/-) and 285-286 (-/-)) and only one with *nahH/xylE* directionality (80-83 (+/+)). The residue clash map identified three clashes located in the region around residue 80 which is the region retained from the same parental sequence in all of the hybrids, thus, preventing the formation of clashes. Interestingly, all the functional hybrids in the library have different parental origins for the contacting residue pair 183-286, however, *none* have *xylE/nahH* directionality, thus avoiding the charge clash that could be formed in the hybrids with reverse (*xylE/nahH*) directionality (see Table 3.1).

Section 3.3.5: Dioxygenases

All four protein systems analyzed so far included hybrids constructed from two parental sequences. The dioxygenase hybrids involve three parental sequences and have a relatively higher number of crossovers per sequence. The active library was created [162] by recombining the α and β subunits of toluene dioxygenase (*todC1C2*),

tetrachlorobenzene dioxygenase (*tecA1A2*), and biphenyl dioxygenase (*bhpA1A2*). *tod* and *tec* are 89.16% identical at the protein level. The *bhp* sequence is less similar, exhibiting 62.30% and 61.85% pairwise sequence identities with *tec* and *tod*, respectively. No structures are available for any of these protein sequences, thus an estimated structure for each one of them is used. The dimeric state of the dioxygenases requires the use Swiss-Model in Optimize mode [147] for structure prediction. Naphthalene dioxygenase (PDB id: 1O7G), a distant homolog of the three dioxygenases was found using the ExPDB database [163] and was used as the template. Figure 3.7 shows the clash maps for the three different sequence combinations (i.e., *tec-tod*, *tod-bhp* and *bhp-tec*) contrasted against the eight active clones with one to eight crossovers per sequence. Comparisons of these results are summarized in Table 3.2. A total of 94 clashes are identified of which 94.68% result from the *tod-bhp* and *bhp-tec* combinations alone, a consequence of low sequence identity between these sequences. Notably, out of the 94 identified clashes only one clash is present in the hybrids (arising from charge repulsion (+/+)) between residues 13 and 385 with a *tec-bhp* directionality) resulting in a high %ACC of 98.9% and a %CFC of 96.8%. Alternatively, we calculated a total of 3,685 non-conserved contacting residues with different parental origins using the estimated structures out of which 84.42% result from the *tod-bhp* and *bhp-tec* combinations. Of these contacts, 1,063 are found to be present in the active hybrids, resulting in %ACC and %CFC values of 71.2% and 9.7% respectively (see Table 3.1).

Section 3.4: Summary

In this work, we introduced a rapid procedure for checking for three different types of clashes (i.e., electrostatic repulsion, steric hindrance, cavity formation and

hydrogen bond disruption) that could be introduced in protein hybrids. This approach was used to identify clashes between contacting residue pairs of the hybrids that have different parental origins for a number of experimental systems. The identified clashing residue pairs between pairs of parental proteins were then contrasted against functionally characterized hybrid libraries. Results of these comparisons, summarized in Table 3.1, show that the pattern of identified clashing residue pairs are consistent with experimentally found patterns of functional crossover combinations. The clash maps p-values (i.e., the fraction of randomly generated clash maps with %ACC greater than or equal to an observed value) were computed for some of the systems. A sample of 100,000 randomly generated clash maps was used with the average number of clashes in each sample equal to those predicted for that particular system. These p-values were found to be in the order of 10^{-2} - 10^{-3} , implying that the predictions are statistically meaningful.

Note also that we find that the residue clash maps are on average 1.55 times more specific (i.e., ratio of %ACC's) than residue contact maps and 5.03 times more specific than randomly generated clashes at explaining observed functional crossovers. While residue contact maps do capture some information on residue pairs that result in unfavorable interaction in the hybrids, not all disrupted contact pairs are detrimental to functionality. The proposed residue clash map improves prediction by filtering out many of the incorrectly predicted pairs. The clash map categorizes these clashes into three distinct types (i.e., electrostatic repulsion, steric clash and hydrogen bond disruption). By pinpointing the cause of these clashes one can then perform site-directed mutagenesis to ameliorate clashes by replacing problematic residues with ones that do not form any

clashes. Admittedly, the residue clash map does not account for the possibility of relieving some of the identified clashes through side-chain and/or backbone movement. This simplification is reflected in the results as the accuracy in crossover classification is reduced as the sequence identity and thus similarity between the parental sequences is reduced (see Table 3.1). Therefore, some of the residues that are in contact in the parental sequences may not necessarily remain in contact in the hybrid thus relieving some of the predicted clashes. Alternatively, new clashes may be introduced due to new contacts formed or altered side-chain conformations. Nevertheless, the proposed approach enables the rapid prescreening of an entire protein family for revealing favorable recombination partners that can subsequently be analyzed by more detailed molecular modeling methods that capture side-chain and backbone movement. So far the clash map based method can only classify hybrids as functional or non-functional but cannot rank hybrids with respect to their activity. In the next chapter I will discuss a computational framework FamClash that overcomes this limitation by ranking the hybrids with respect to their activity based on the identified clashes.

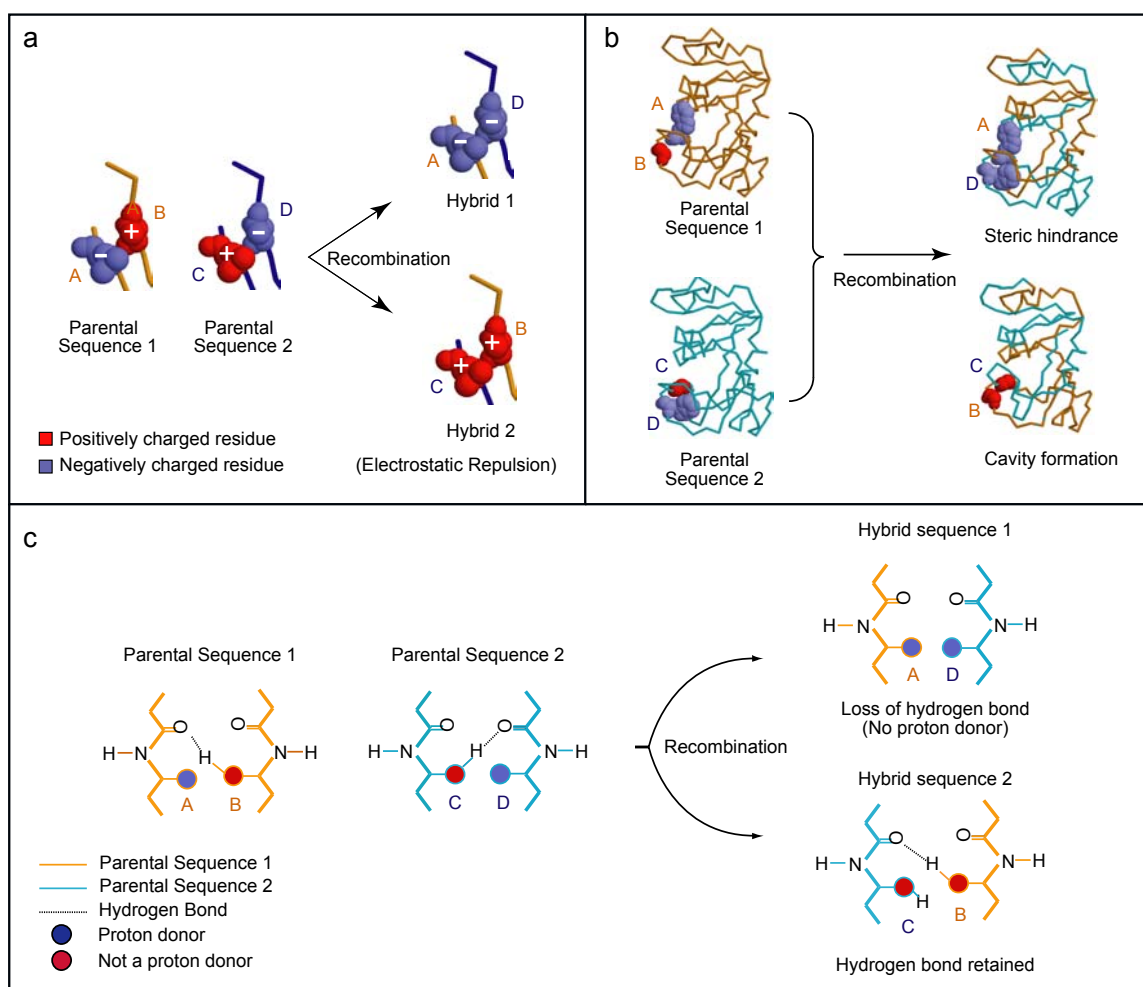


Figure 3.1: (a) The contacting residues A-B (parental sequence 1) and C-D (parental sequence 2) have opposite charges and different relative positions in the two parental sequences. Recombination results in electrostatic repulsion between residues A-D (-/-) in the first hybrid and B-C (+/+) in the second hybrid. (b) The first hybrid retains residues with large side chains from both parental sequences 1 and 2 (A-D) causing a steric hindrance. Pairing of the residues with small side-chains (C-B) in the second hybrid leads to a cavity formation. (c) Hybrid 2 retains proton donors (C, B) from both parental sequences and thus the hydrogen bond between the side-chain donor and the backbone acceptor is retained. Alternatively, hybrid 1 retains residues with side-chains that have no proton donors (A, D) resulting in the loss of the hydrogen bond between the two residues.

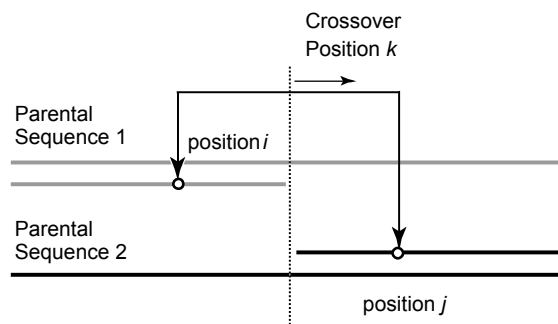
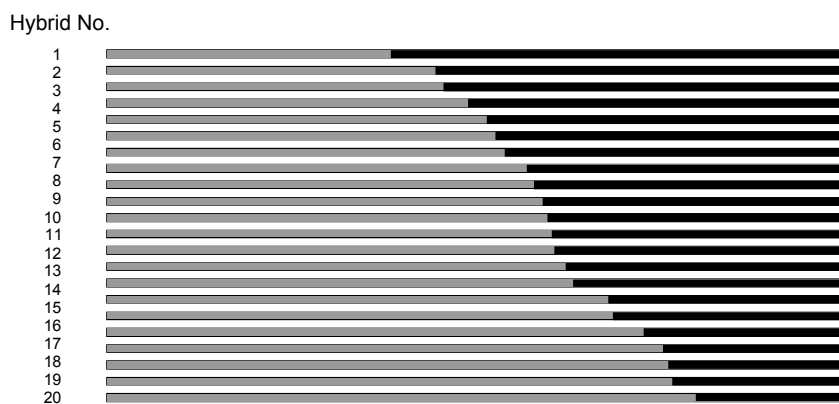
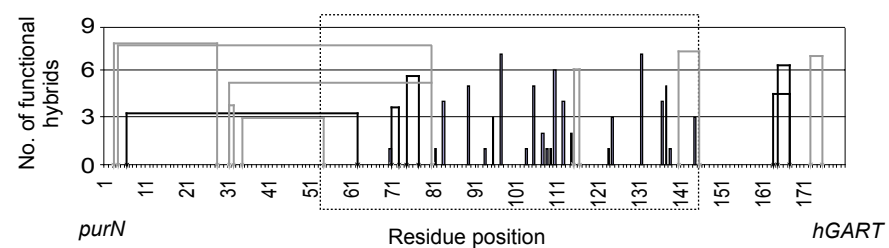
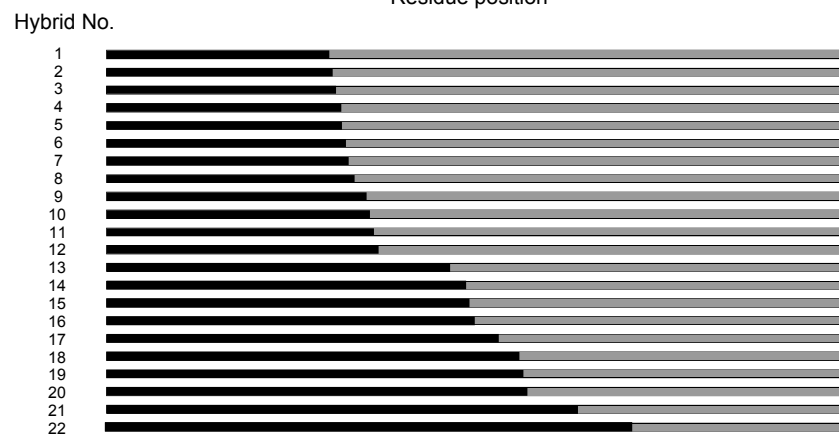
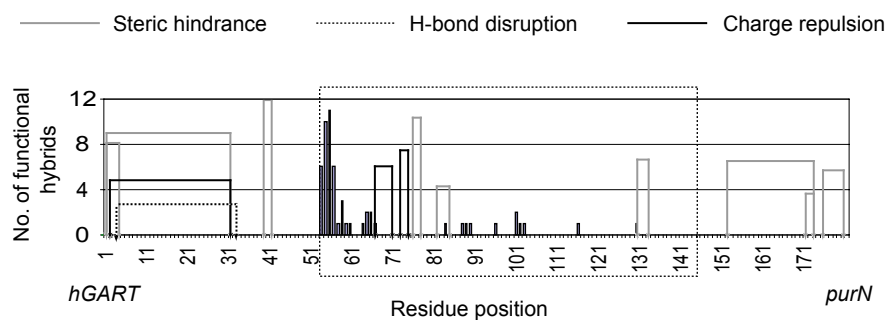


Figure 3.2: An unfavorable interaction between the two residues at positions i and j in the hybrid is represented by an arc between the two positions. The residue at position i is retained from parental sequence 1 and j from parental sequence 2. Arcs depict any one of the three forms of clashes: (i) electrostatic repulsion, (ii) steric clashes and (iii) hydrogen bond disruption. A crossover at position k ($i < k < j$) brings the two contacting residues with different parental origins together thus forming a clash.

(a) Incremental truncation library (*purN*/hGART)(b) Incremental truncation library (*hGART*/*purN*)**Figure 3.3:** Different types of clashes for (a) *purN*/hGART and (b) hGART/*purN* are

shown as arcs linking the two positions. Functional crossover positions [5] are shown as vertical bars whose heights represent their number. Shown below these clash maps are the functional hybrids with gray region corresponding to *purN* and black to *hGART*. Notably, the crossover distribution and directionality in both cases is such that most functional hybrids are free of the identified clashes.

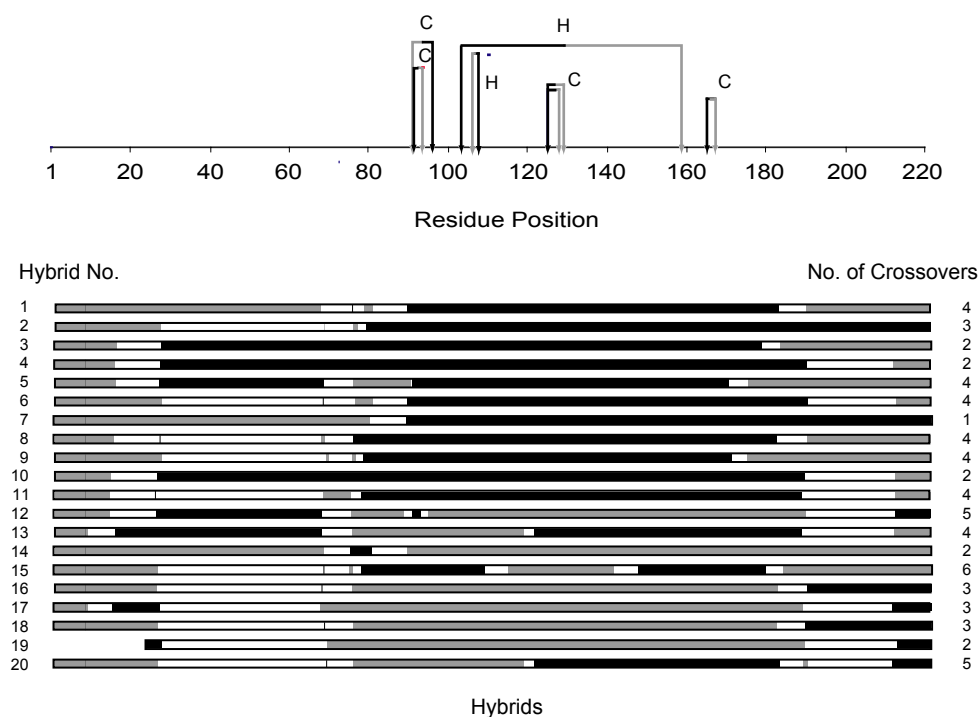


Figure 3.4: Residues in the hybrids retained from parental sequences with the same color (gray: GST M1-1, black: GST M2-2) as the arc connecting them, lead to an unfavorable interaction. The arcs indicate steric hindrance (H) or electrostatic repulsion (C) between the two residues. Shown below these arcs are the functional hybrids, constructed using DNA shuffling, of GST M1-1 (gray) and GST M2-2 (black). They are ordered in decreasing ratios of activities with respect to aminochrome and CDNB [159]. White segments represent conserved stretches of unknowable origin. Numbers to the right of each hybrid indicate the number of crossovers.

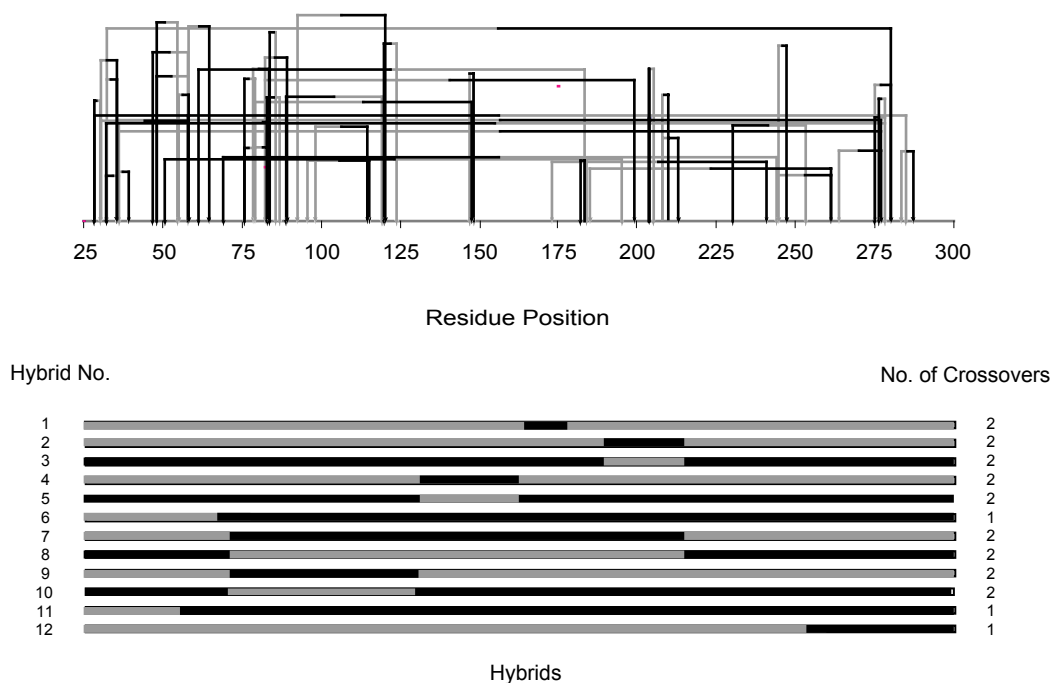


Figure 3.5: The identified residue clashes are shown against the ten active β -lactamase (TEM-1 (black), PSE-4(gray)) hybrids identified experimentally [72]. The total number of clashes in the TEM-1/PSE-4 directionality is found to be 27 while the reverse directionality has 30 clashes. Hybrids in either directionality contain 14 cases of charge repulsion while the remaining resulted from steric clashes.

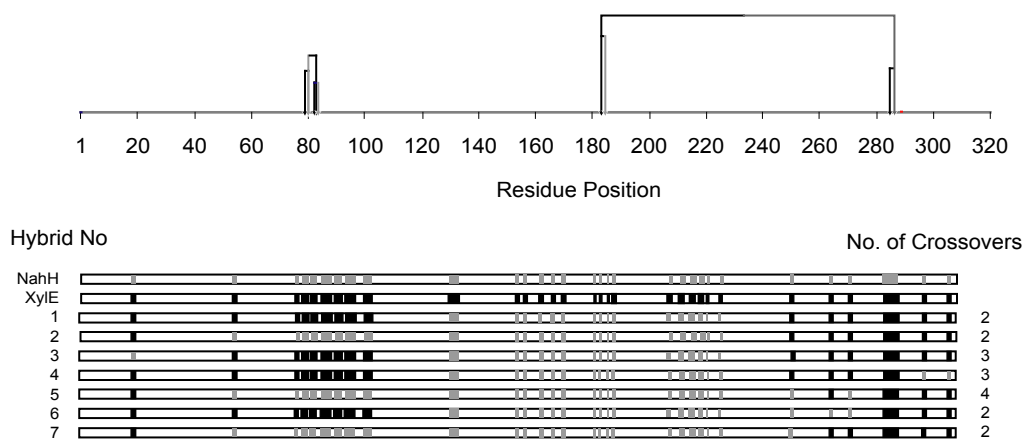


Figure 3.6: Seven different thermally stable C23O hybrids obtained by shuffling ssDNA are shown above [160]. The residues derived from *NahH* and *XylE* are shown in gray and black respectively, while conserved residue positions of ambiguous origin are colored white. Only six clashes all of which result from charge repulsion are identified.

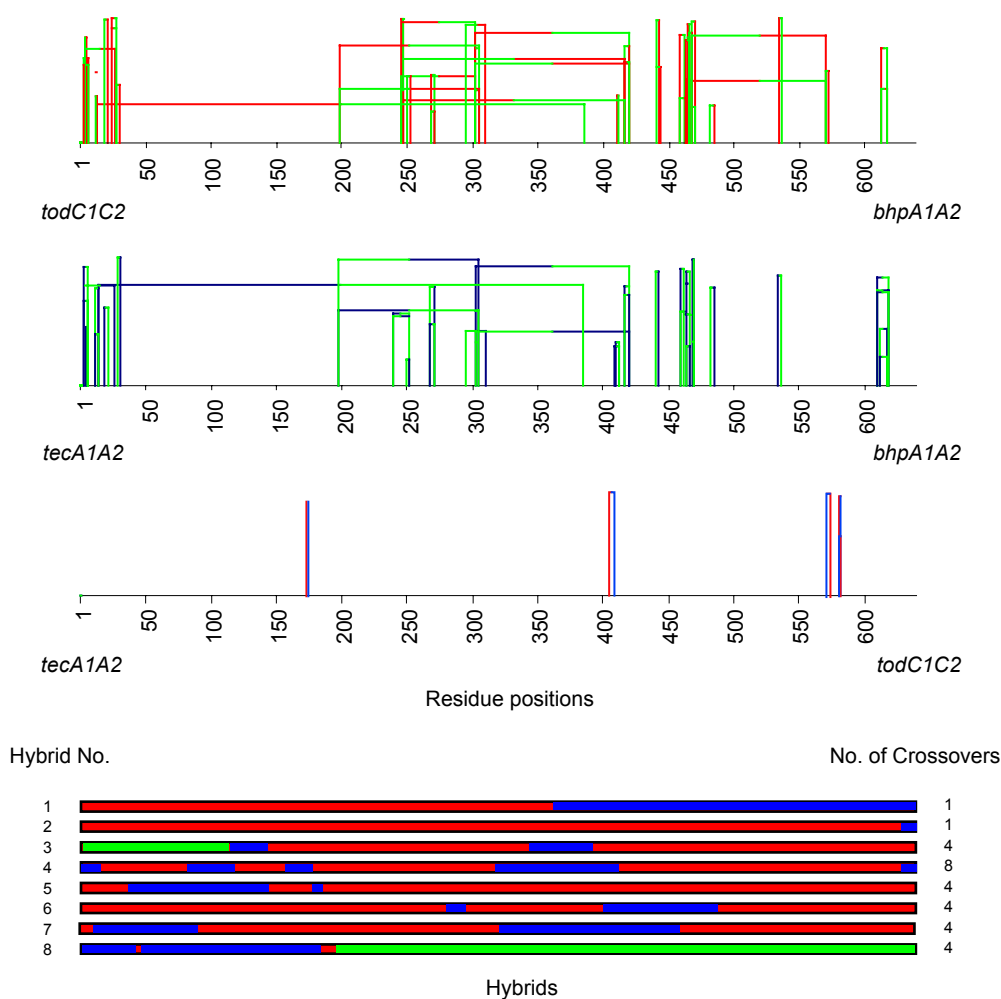


Figure 3.7: Eight toluene-active members of the hybrid library obtained by shuffling genes encoding the α and β subunits of three dioxygenases are shown as horizontal bars [162]. Sequence elements from *tecA1A2*, *todC1C2* and *bhpA1A2* are colored blue, red and green respectively. Shown above these are the clash maps corresponding to the three different sequence combinations (i.e., *tod-bhp*, *tec-bhp* and *tod-tec*) whose details are given in Table 3.2.

Table 3.1: Summary of statistical analysis for the five protein families.

Protein System	Sequence Identity (%)	RMSD ¹ (Å)	Total number of identified clashes	Correct ⁴ clashes	Residue clash map (RCM)		Residue contact map		Random clashes	[%ACC ^{RCM} / %ACC ^{Rnd}]
					%ACC ²	%CFC ³	%ACC	%CFC		
GART	38.20	1.30	13*	8*	61.54	90.91	30.20	0.00	9.74	6.31
GST	85.25	0.50	7	7	100.00	100.00	56.33	0.00	13.10	7.63
β-lactamase [#]	43.17	1.30	57	46	80.70	35.00	65.00	0.00	14.68	5.50
β-lactamase ⁺	43.17	1.30	57	44	77.19	31.03	62.31	0.00	13.07	5.90
C23O	84.70	0.10	6	6	100.00	100.00	70.86	0.00	25.86	3.87
Dioxygenases	~71.10	-	94	93	98.90	96.80	71.20	9.70	54.08	1.83

¹Root mean square distance (in Angstroms) between the crystal structures of the two parental sequences.

²%ACC is defined as the percentage of arcs representing clashes or contact pairs that are not disrupted by the crossover pattern found in the functional hybrid library.

³%CFC is defined the percentage of functional crossovers that do not disrupt any of the arcs representing clashes.







⁴Number of identified clashes absent in functional hybrid library.

*These values are based on clashes found in the region between residues 53-144.

[#]These results are based on the crossover data taken from results published by Voigt and coworkers (2002).

⁺These results are based on the crossover data published by Hiraga and Arnold (2003).

Table 3.2: Clash map based analysis for the dioxygenase system.

Crossover type	Total number of clashes (see Figure 7)	Clashes present in hybrids 1-8.
<i>bhp-tod</i> 	26	0
<i>bhp-tec</i> 	21	0
<i>tod-bhp</i> 	21	0
<i>tec-bhp</i> 	21	1
<i>tec-tod</i> 	2	0
<i>tod-tec</i> 	3	0

Chapter 4: FamClash: A Method for Ranking the Activity of Engineered Enzymes

Section 4.1: Background

Recent advances in protein engineering [5, 6, 34-36] have allowed researchers to go beyond the limitations of homology-dependent directed evolution methods. The ability to freely explore protein sequence space has revealed a number of troublesome trends. Firstly, the lower the sequence identity of the recombined parental sequences, the smaller the percentage of the combinatorial protein library that remains functional [5, 6]. This has been reported in several studies [164-166] using differing protocols, thus implicating the global nature of this effect. More troublesome is the finding that the remaining functional hybrids tend to have only residual activities. Therefore, it appears that exploring protein sequence space freely comes at the expense of severely degrading the average stability and functionality of the combinatorial library. This has motivated the development of computational methods to prescreen hybrids for their potential of being stably folded [33] and functional. These analyses then serve to direct the sampling of protein sequences by the combinatorial library towards desirable regions in sequence space. Specifically, favorable positions for junctions between fragments from different parental sequences can be identified, and restrictions can be imposed on sets of parental sequences that contribute fragments to a particular junction.

Therefore, further improvements in the stability and functionality of hybrid proteins may be attained by developing quantitative methods that identify deleterious interactions arising from residue pairs within the gene fragment combinations. To this end, Monte Carlo simulations by Bogarad and Deem [134] suggested that swapping of low-energy structures is least disruptive to protein structure. The SCHEMA algorithm

[135] postulates that contacting residue pairs in the hybrids that have different parental origins are unlikely to interact favorably, and thus are preferentially avoided in functional hybrids. This hypothesis has been successfully applied to a number of experimental studies [36, 135, 136] to explain the distribution of functional crossover positions. Moore and Maranas [137] proposed the second-order mean-field approach (SIRCH) to identify residue-residue clashes in hybrids that prohibit them from folding into the correct backbone structure. Interestingly, most of the clashes identified resulted from (i) electrostatic repulsion, (ii) steric hindrance or cavity formation, and (iii) disruption of hydrogen bonds. Subsequently, Saraf and Maranas [167] proposed a rapid method to identify directly such clashes between contacting residue pairs in the protein hybrids. Comparison with sequence data of functional clones derived from many studies [5, 135, 159, 160, 162] revealed that the method was capable of classifying hybrids (crossover combinations) as functional or non-functional accounting for mirror chimeras. However, neither this method nor any of those discussed earlier manage to *a priori* rank functional hybrids with respect to their level of activity. Given that the goal of directed evolution studies is not just to retain residual activity levels but rather to reach/improve upon the parental levels of activity, the ability to move beyond active/non-active classification and computationally rank-order hybrids defines the next key challenge.

Protein family sequence [132, 168-170] and structure [83, 84, 171] data have often been used as a basis for predicting the presence or absence of functionality. Saraf and coworkers [132] have shown that residue pairs, important for functionality, frequently exhibit a correlated mutation pattern, implying that the physicochemical properties of these residue pairs are also coupled. Correlation in sequence alignment has

also been inferred as structural constraints, translating to residue-residue contacts [145, 172]. These correlation signals are stronger when obtaining measurements using ancestral sequences inferred from phylogenetic data [173, 174]. In a similar effort, Govindarajan and coworkers [175] showed that for many pairs of positions in protein families certain residue combinations are highly preferred. It is reasonable to expect that the same correlation pattern may extend to the properties of specific residue pairs, e.g., size, hydrophobicity, charge [176]. For example, a lysine-lysine residue pair is often substituted for an arginine-arginine owing to the similarity in the charge, volume and hydrophobicity between these residue pairs [177-181].

In this work, we introduce the FamClash procedure for inferring the rank-ordering of the relative levels of activity of protein hybrids. FamClash is based on the method developed by Govindarajan and coworkers [175] that encompasses sequence information from not only the parental sequences but also from members composing the entire protein family to be engineered. In addition, since many studies have shown that the interactions of even distal residues can have a significant impact on the activity of the hybrids [182-184], we include such non-contacting pairs in our analysis. FamClash proceeds in three steps: (i) pairs of positions in the protein family sequence alignment are first identified where residue pairs within a protein having similar properties are preferentially retained; (ii) next, residue pairs at these positions in the hybrids are examined to check if they retain the properties observed in the protein family; and (iii) finally, ranking these hybrids with respect to their probable activity based on the extent of departure from the family sequences, measured in terms of number of clashes.

FamClash is experimentally tested by constructing single-crossover hybrids of *E.*

coli and *B. subtilis* DHFRs. Results demonstrate that the specific activities of the hybrids are qualitatively consistent with FamClash predictions. This combined experimental and computational study lays the groundwork for developing approaches to protein engineering using enzymes with low sequence identity. Furthermore, valuable information is derived as to what residue positions need to be redesigned.

Section 4.2: Hybrid construction and functional screening

Section 4.2.1: Plasmid constructions

Plasmid pAZE was designed for combinatorial construction and genetic selection of DHFR hybrids. To build this plasmid, the *lacI^Q* gene was PCR amplified from pMAL-c2x (New England Biolabs) with *NheI*-tailed primers, digested with *NheI*, and ligated into the *SpeI* site of pZE12-*luc* [185]. The ribosome-binding site and *luc* gene were removed with *EcoRI* (blunted) and *XbaI*, and this piece was replaced with a *SacII* (blunted), *XbaI* fragment from pDIM-N2 [186]. Residues 1-120 of *E. coli* DHFR were PCR amplified, digested with *NdeI* and *BamHI*, and ligated into pMAC [187] cut with the same enzymes. Residues 31-168 of *Bacillus subtilis* DHFR were PCR amplified, digested with *PstI* and *SpeI*, and ligated downstream of the *E. coli* fragment on pMAC. The *NdeI-SpeI* piece was removed from pMAC and ligated into pAZE, and the resulting plasmid was named pAZE-EB and confirmed by DNA sequencing. A complimentary plasmid for *B. subtilis* N-terminal DHFR hybrids, named pAZE-BE, was constructed from fragments 1-121 of *B. subtilis* and 31-159 of *E. coli* DHFRs. An additional plasmid with a fixed crossover at position 62 was constructed in vector pAZE by overlap extension [188]. Primer sequences can be provided upon request.

Section 4.2.2: Construction of DHFR hybrid libraries

Plasmids pAZE-EB and pAZE-BE were linearized at a unique *SalI* site between the *E. coli* and *B. subtilis* fragments. The THIO-ITCHY PCR technique was used to construct libraries of *E. coli/B. subtilis* DHFR hybrids in both orientations [189]. Libraries were initially constructed and frozen in *E. coli* strain DH5 α -E.

Section 4.2.3: Selection of DHFR hybrids.

E. coli strain MH829 has a deletion of the DHFR (*folA*) gene and was used for the *in vivo* selection of functional DHFR hybrids [190]. Library plasmid was purified and electroporated into strain MH829. Transformed cells were washed twice in MMA [188] and plated on 245 \times 245 mm library plates of MMA supplemented with 0.5% glycerol, 0.6 mM arginine, 50 μ g/ml thymidine, 25 μ g/ml kanamycin, 100 μ g/ml ampicillin, and 1 mM MgSO₄. Selections were performed at room temperature and IPTG was added to induce expression, usually at 250 μ M final concentration. Isolates were restreaked onto the same media, grown at 30°C, and plasmids were sequenced to identify crossover positions. All DNA sequencing was performed at the Nucleic Acids Facility of Penn State University.

Section 4.2.4: DHFR Assays

DHFR ligands were prepared as previously described [191]. The specific activities of wild-type and hybrid DHFRs were determined in cell-free lysates. The plasmid pAZE (described above) was used to express all DHFR proteins, and to increase expression, *lacI* was destroyed on all plasmids by *EcoRV* and *SfoI* digests. Plasmids were transformed into DHFR mutant strain MH829, and 50 ml of cultures were grown at 30°C in LB broth supplemented with 100 μ g/ml ampicillin, 50 μ g/ml thymidine, and 0.5

mM IPTG. Cultures were grown to an absorbance of 1.0 at 600 nm, centrifuged, and resuspended in 25 ml of 20 mM Tris-HCl, pH 7.7, with 2 mM DTT. Cells were centrifuged again, resuspended in 1 ml buffer, and broken by sonication. Insoluble material was removed, and lysates were assayed on a Cary 100 Bio UV-Vis spectrophotometer (Varian Inc., Palo Alto, CA), held at 25 °C with a water-jacketed cuvette holder. Cell-free lysate was preincubated 3 minutes in MTEN buffer, pH 7.0, containing 1 mM DTT and 100 μ M cofactor to avoid hysteresis [191], and the reaction was initiated by adding 100 μ M substrate. To follow the reaction, the decrease in absorbance was monitored at 340 nm ($\Delta\epsilon_{340} = 13.2 \text{ mM}^{-1}\text{cm}^{-1}$).

Section 4.3: FamClash Method

FamClash relies on identifying residue positions in the parental protein family sequences for which the sum of residue properties are conserved. Hybrids are then evaluated with respect to whether they confirm to the identified conserved properties. Any deviations are denoted as residue-residue clashes. This is accomplished by first analyzing the family sequence alignment obtained from the Pfam database [192] using scoring matrices. These scoring matrices encode physicochemical properties of amino acids such as charge [193], volume [194] and hydrophobicity [195, 196]. The additive charge (C_{ij}^m), volume (V_{ij}^m) and hydrophobicity (H_{ij}^m) for a pair of residues k, l at positions i and j in sequence m is defined as the sum of the charge (c), volume (v) and normalized average hydrophobicity metric (h) of residues k and l :

$$C_{ij}^m = c_{ik}^m + c_{jl}^m, \quad V_{ij}^m = v_{ik}^m + v_{jl}^m, \quad H_{ij}^m = h_{ik}^m + h_{jl}^m \quad (1)$$

All 20×20 pairwise residue combinations are partitioned into three-dimensional (3D)

property bins derived by subdividing the observed property ranges (see Figure 4.1). A residue pair populates a particular bin ϕ_{pqr} if all of its properties lie within the rectangle defined by: $\left[\left(C_{ij}^m = p \right), \left(q \leq V_{ij}^m < q + \Delta_V \right), \left(r \leq H_{ij}^m < r + \Delta_H \right) \right]$ as shown in Figure 4.1. Note that the total charge p of a residue pair can assume only one of five distinct values (i.e., -2, -1, 0, 1, 2). In contrast, volume (q) and hydrophobicity (r) values may vary continuously within ten equally sized bins ranging between 0-300 Å³ and -2.30-3.7 kcal/mol, resulting in Δ_V and Δ_H values of 30 Å³ and 0.6 kcal/mol, respectively.

A pair of positions in the sequence alignment is deemed “conserved” if at least 20% of its residue pairs, including the parental residue pairs, populate the same 3D property bin ϕ_{pqr} . Conservation of additive property values signify that any significant deviations from the observed ranges may lead to residue-residue clashes (see Figure 4.1). To safeguard against identifying conservation due to chance, the *mutual information index* (M_{ij}^{pqr}) between all pairs of positions in the alignment for the corresponding property bin ϕ_{pqr} is calculated. Chance occurrences are revealed when the occupancy frequencies of residues k, l at two positions i and j (a_{ik}, a_{jl}) are independent. In such a case, the joint probability $P(a_{ik}, a_{jl})$ of observing an amino acid k at position i and at the same time amino acid l at position j is equal to the product $P(a_{ik})P(a_{jl})$ of the individual probabilities of occupancy for these two residues and position. The M_{ij}^{pqr} score quantifies the degree of dependence (or independence) between the distributions of residues at the two positions:

$$M_{ij}^{pqr} = \sum_k \sum_l P(a_{ik}, a_{jl}) \cdot \log_2 \left\{ \frac{P(a_{ik}, a_{jl})}{P(a_{ik})P(a_{jl})} \right\} \quad \forall k, l : a_{ik} a_{jl} \in \phi_{pqr} \quad (2)$$

Note that completely independent residue positions will have a M_{ij}^{pqr} score exactly equal to zero. The higher the value of M_{ij}^{pqr} , the stronger the extent of covariance between positions i, j for property values within bin ϕ_{pqr} . A pair of positions i, j and bin ϕ_{pqr} is considered to be statistically significant if its M_{ij}^{pqr} score is greater than a cutoff value (M_c).

A bootstrap replicate analysis is used to determine the threshold value (M_c) for M_{ij}^{pqr} scores. This establishes how likely is a M_{ij}^{pqr} score greater than M_c to occur by chance alone. First, two vectors are extracted from the sequence alignment (i.e., columns of residues at positions i, j from the alignment). Next, multiple copies ($\sim 10^5$) of each of these vectors (bootstrap replicates) are generated by randomly choosing residues by permuting the original vector. Finally, M_{ij}^{pqr} scores for all 10^5 pairs of randomized vectors are computed for each property bin ϕ_{pqr} . This distribution of scores serves to elucidate the probability of having a M_{ij}^{pqr} score greater than a given cutoff score (M_c) by chance. This probability, also known as the p -value, is calculated as the ratio of the total number of pairs yielding scores above M_c divided by the total number of pairs in the distribution. The M_{ij}^{pqr} score corresponding to a p -value of 5×10^{-3} is chosen as the cutoff.

A clash is defined to occur between two statistically significant residue positions i, j in the hybrid (residue at position i retained parental sequence p_1 and j from p_2) if at least one of the following criteria is met:

$$c_i^{p_1} + c_j^{p_2} \neq \bar{C}_{ij} \quad (3)$$

$$\left(v_i^{p_1} + v_j^{p_2} \right) - \bar{V}_{ij} > \delta_{v_{ij}} \text{ (steric) or } \left(v_i^{p_1} + v_j^{p_2} \right) - \bar{V}_{ij} < -2\delta_{v_{ij}} \text{ (cavity)} \quad (4)$$

$$\left| (h_i^{p_1} + h_j^{p_2}) - \bar{H}_{ij} \right| > \delta_{h_{ij}} \quad (5)$$

Because cavity formation tends to be less problematic than steric hindrances (see [167]) a more relaxed cutoff for cavity formation is chosen. Here \bar{C}_{ij} , \bar{V}_{ij} and \bar{H}_{ij} are the mean charge, volume and hydrophobicity found to be conserved between positions i and j in the protein family members. Assessing the departure away from the mean property values for any pair of positions i, j , identified as conserved, requires the definition of cutoff ranges for volume (δ_{vij}) and hydrophobicity (δ_{hij}) as follows:

$$\delta_{vij} = \max \left\{ \left| V_{ij}^{p_1} - V_{ij}^{p_2} \right|, \frac{\bar{V}_{ij}}{10} \right\}, \quad \delta_{hij} = \max \left\{ \left| H_{ij}^{p_1} - H_{ij}^{p_2} \right|, \frac{\bar{H}_{ij}}{10} \right\} \quad (6)$$

A lower bound on the cutoff ranges is set to 10% of the mean values to prevent small deviations in the properties to be denoted as clashes. Table 4.1 summarizes the steps of FamClash procedure.

Section 4.4: Results and Discussion

Section 4.4.1: Library construction and hybrid isolation

Two ITCHY libraries were constructed from the *E. coli*/*B. subtilis* (EB) or the *B. subtilis*/*E. coli* (BE) DHFR pairs sharing a 44% sequence identity at the protein level. The naive library sizes were 1.9×10^6 and 2.0×10^6 members respectively, providing complete coverage of the minimum library size of 7.3×10^4 [(270 bp)²]. A genetic selection for functional hybrids was developed using an *E. coli* strain containing a complete deletion of DHFR [190]. The nature of the selection required the use of inactive DHFR fragments to make ITCHY libraries, which limited the crossover window to residues 31-120 (see Methods). Following selection, hybrids were picked at random

and sequenced, 55 from the EB (i.e., *E. coli*/*B. subtilis*) library and 10 from the BE (i.e., *B. subtilis*/*E. coli*) library. DNA sequencing showed that 30 of the EB library members had duplications of various sizes, and that all the BE library members had duplications.

The number of DHFR hybrids with duplications was somewhat unexpected, especially considering how rarely they were identified in ITCHY libraries of GAR transformylases [186, 189]. In the BE library, attempts were made to identify perfect crossovers (i.e., containing no duplications) by removing hybrids larger than wild-type DHFR through gel electrophoresis (data not shown). However, even after sorting, all BE hybrids contained at least one or two amino acid duplications, many with considerably larger ones. The stringency of the genetic selection was designed to be very low, accepting DHFR hybrids with k_{cat} values 10^3 fold lower than wild-type (data not shown), which may have contributed to the high number of duplications observed. To simplify the analysis, 13 perfect crossovers from the EB library were selected for further studies. These DHFR hybrids were chosen to provide the best distribution across the 90 amino acid crossover window (see Figure 4.2), and all hybrids containing duplications were not pursued further.

Section 4.4.2: FamClash analysis of EB library

Conserved pairs of positions for the two aligned DHFR sequences were identified by evaluating the M_{ij}^{pqr} scores as outlined in the Methods section. The DHFR protein family sequence alignment was obtained using the Pfam database [192] including a total of 265 DHFR sequences (as of Nov. 15, 2003). Statistically significant residue positions were identified by the bootstrap replicate analysis. Residue pairs in the EB and BE libraries corresponding to the statistically important residue positions (p -value $< 5 \times 10^{-3}$)

were identified and their properties were investigated for consistency with the protein family sequence data. Specifically, we found that 14 residue pairs for the EB hybrids showed significant deviations in the property triplet from what is found to be conserved among the corresponding residue positions in the protein family sequences (see Figure 4.2 and Table 4.2). Only six such pairs were identified for hybrids with a BE directionality (see Table 4.2). We observed that most of these clashes are due to large changes in the total volume of the residue pairs. In fact, many of the identified clashes in the hybrids are a direct consequence of reversed orientation of residue pairs in the two parental sequences. For example, the residue pair 36/135 in *E. coli* is a lysine and a serine while in *B. subtilis* the same pair involves the same residues but in a reversed order (see Table 4.2). This consequently results in a steric hindrance in the EB hybrid and a cavity formation in the BE hybrid. Both hydrophobicity and charge were found to be fairly conserved and thus very few clashes due to deviation from charge and hydrophobicity values were identified. Table 4.2 lists all the identified clashes between residue pairs in the hybrids (also see Figure 4.2). Notably we found that many of the predicted clashes are between distant residue pairs.

Figure 4.3 shows the total number of identified clashes for the single crossover incremental truncation EB and BE libraries. Notably, the BE hybrids have about half as many clashes as the EB hybrids. Also, five of the six clashes identified in BE hybrids are also present in the EB hybrids (see Table 4.2). Interestingly, in four out of five cases of volume clashes common to both libraries, EB hybrids retain residue pairs with larger side-chains presumably leading to steric hindrances, whereas in the BE library a corresponding volume reduction was observed. This suggests that BE hybrids, by

avoiding steric clashes, are more likely to retain functionality in comparison to their EB mirror chimeras. This is consistent with the experimental results where BE hybrids are found to be much more tolerant to insertions.

Section 4.4.3: DHFR hybrid characterization and analysis

Specific activities of the EB hybrids were determined in lysates of the *E. coli* DHFR mutant MH829. The hybrids with the lowest activity, crossovers 55-96, all reside in the adenosine binding subdomain. This region of DHFR is directly involved in NADPH binding [197], and splicing together residues in this subdomain from divergent DHFRs could have dramatically affected cofactor binding, implying the thermodynamic dissociation constants, K_D values, are significantly affected. Molecular dynamics simulations have identified anti-correlations between the 55-96 region and both the Met-20 loop (residues 14-24) and β F-G loop (residues 116-125), suggesting that the protein dynamics of these hybrids also might have been perturbed [198]. Further, functional connectivities between the cofactor and substrate binding sites have been observed for DHFR [199, 200], which could be affected by crossovers in the NADPH binding region.

The DHFR activity was plotted against crossover position and compared to the FamClash predictions (Figure 4.4). Log-log plots are frequently used to correlate activity versus mutational data. This implies that the change in free energy is proportional to the log of the total number of mutations alluding to a continuously diminishing effect of additional mutations. Also, SCHEMA results [136] have demonstrated that the logarithm of the fraction of functional recombinants is proportional to the negative of the logarithm of schema disruptions. In analogy with these results we decided to use a log-log plot to contrast activities and total number of clashes. As shown in Figure 4.4, the trend of

DHFR activities correlated surprisingly well with the number of clashes in each hybrid and appears to exhibit a “V” shape, although the small sample size could have contributed to this observation. It is possible that many perfect crossovers in the gaps shown in Figure 4.4 are active DHFR hybrids, and the activities of these potential hybrids may deviate from the observed trend. The stringency of the selection could be raised to enrich for only the most active hybrids. However, the results from previous ITCHY libraries suggested there would be valleys of low activity [186], and the goal of this work was to obtain the most complete crossover distribution possible for comparison to computational predictions.

Notably, as shown in Figure 4.4, EB hybrid 79 has fewer clashes than the neighboring hybrids. The FamClash method predicted that residue 62 from *E. coli* clashes with residue 78 from *B. subtilis* and residue 80 from *E. coli* clashes with residues 127 and 156 from *B. subtilis*. Both these clashes are absent in EB hybrid 79, and consistent with these predictions, this hybrid showed considerably better activity than flanking hybrids 73 and 81. In addition, crossover position 62 was predicted to have the maximum number of clashes. This hybrid was subsequently constructed and assayed, and the activity of this hybrid was poor, consistent with the downward trend observed in the plot. However, the activity of hybrid 62 was noticeably higher than hybrid 73, which was predicted to have fewer clashes. This is consistent with a diminishing effect of increasing number of clashes in analogy with the observation that increased number of mutations do not additively effect activity [201]. Also, increasing numbers of clashes may not have the predicted additive effect on enzyme activity, perhaps due to the

inability at this time to rank the importance of each clash and to capture higher order effects.

Section 4.5: Summary

In the current implementation of FamClash, all clashes are considered equally deleterious. One would expect that some clashes may be more severe than others and, therefore, may have significant impact on activity, sometimes even greater than the combined effect of more than one clash. Moreover, more than two residues may be involved in retaining a particular property that cannot be identified when analyzing just pairs of residues, alluding to the limitations of the FamClash procedure. Nevertheless, the results presented here show that FamClash is quite successful at qualitatively predicting the pattern of the specific activity of the hybrids. Similar trends have been observed for other systems not presented here. More importantly, by identifying these clashes, this method provides valuable insights for protein engineering interventions to remedy these clashes. Specifically, by appropriately substituting residues at the clashing positions, significant improvement in the activity of the hybrids can be achieved. In the next two chapters, I will describe two computational frameworks, OPTCOMB and IPRO, which identify optimal tiling pattern of parental sequence fragments or amino acid substitutions that systematically avoid/eliminate these clashes.

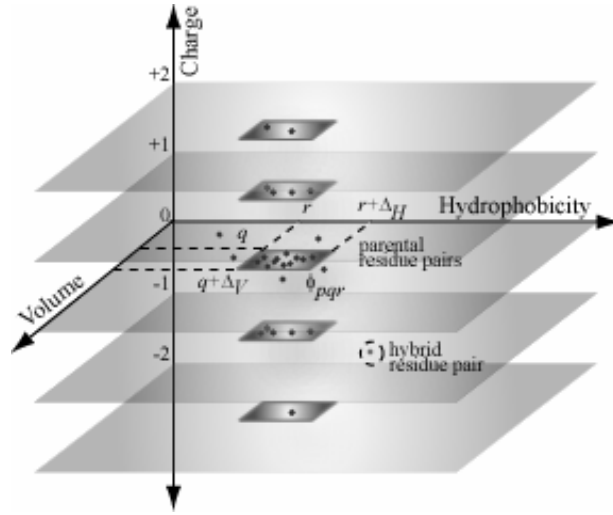


Figure 4.1: Residue pairs i, j whose properties are within a specified range in terms of charge (p), volume (q) and hydrophobicity (r) are said to belong to the same 3D property bin ϕ_{pqr} ($i, e., C_{ij} = p, q \leq V_{ij} < q + \Delta_V, r \leq H_{ij} < r + \Delta_H$). Property values for the residue pair in the hybrid that are significantly different than those observed in the protein family denote a clash.

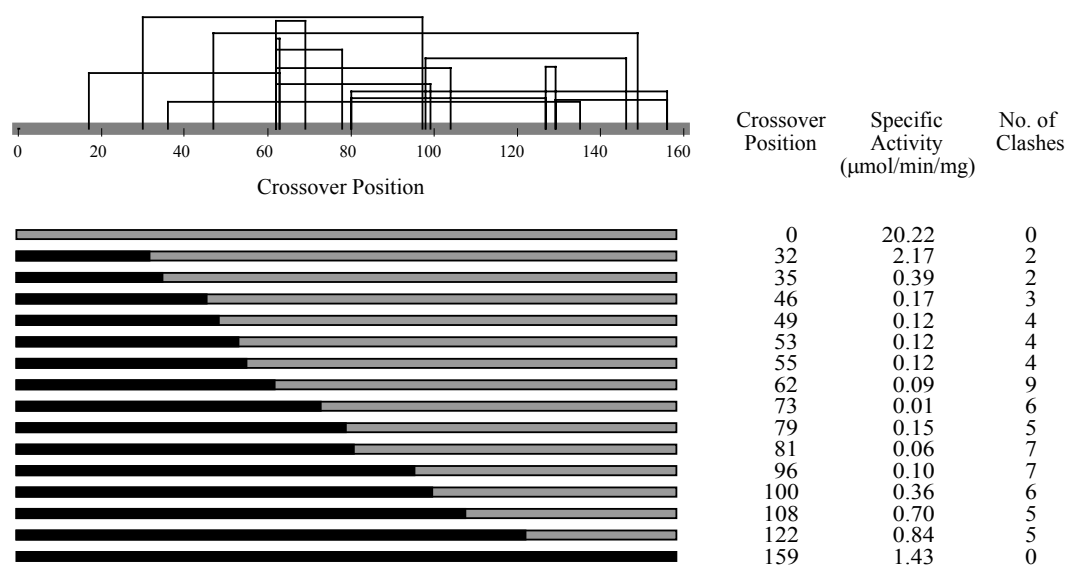


Figure 4.2: Predicted clashes in EB hybrids are shown for all single crossover EB hybrids. A clash between any two residue positions is shown as an arc. The specific activity ($\mu\text{mol}/\text{min}/\text{mg}$) and number of clashes in each hybrid are also shown. Note that the 0 and 159 crossover positions correspond to the parental *B. subtilis* and *E. coli* DHFR sequences, respectively.

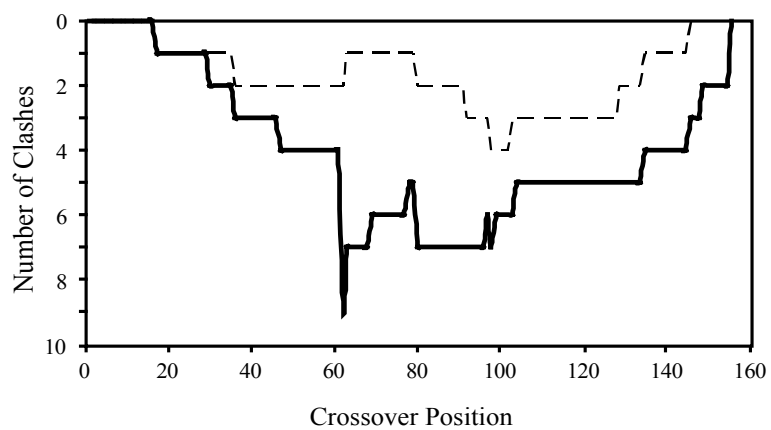


Figure 4.3: The number of clashes in each of the single crossover EB (—) and BE (---) DHFR hybrids are plotted against crossover position.

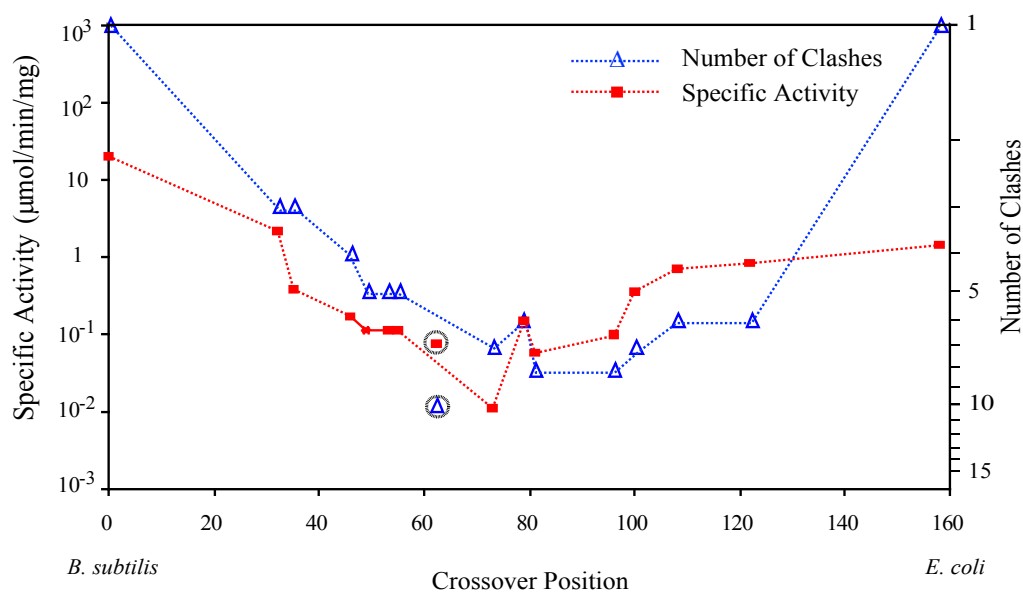


Figure 4.4: Plot of specific activities (■) of the 13 EB DHFR hybrids against crossover position. The total number of identified clashes (Δ), (i.e., $\log(1 + \text{number of clashes})$) for each one of these hybrids is also overlaid in the plot. Note that the 0 and 159 crossover positions correspond to the parental *B. subtilis* and *E. coli* DHFR sequences, respectively. The specific activity and number of clashes for hybrid 62 is shown separately.

Table 4.1: Summary of the FamClash procedure.**FamClash Procedure**

Step 1. Identify all pairs of positions i, j in the sequence alignment where at least 20% of the residue pairs have charge, volume and hydrophobicity that lie in the same 3D property bin ϕ_{pqr} .

Step 2. Evaluate mutual information (M_{ij}^{pqr}) score for all pairs of positions i, j denoted as conserved for the corresponding bin ϕ_{pqr} .

Step 3. Perform bootstrap replicate analysis and select positions i, j that meet the p -value cutoff of 5×10^{-3} .

Step 4. Investigate the selected residue positions in the hybrids for clashes based on the following criteria:

$$c_i^{p_1} + c_j^{p_2} \neq \bar{C}_{ij} \quad (\text{I})$$

$$(v_i^{p_1} + v_j^{p_2}) - \bar{V}_{ij} > \delta_{v_{ij}} \text{ (steric) or } (v_i^{p_1} + v_j^{p_2}) - \bar{V}_{ij} < -2\delta_{v_{ij}} \text{ (cavity)} \quad (\text{II})$$

$$\left| (h_i^{p_1} + h_j^{p_2}) - \bar{H}_{ij} \right| > \delta_{h_{ij}} \quad (\text{III})$$

Step 5. Hybrids are given a score equal to the number of clashes identified in **Step 4**.

Table 4.2: Positions, residue pairs and nature of clashes in the hybrids.

Hybrid	Residue positions	Residue pair (parent 1)	Residue pair (parent 2)	Residue pair (hybrid)	Nature of clash *
<i>E.coli</i> / <i>B.sub</i>	17/63	ES	DT	ET	steric
<i>E.coli</i> / <i>B.sub</i>	30/97	WG	YA	WA	steric/hyd
<i>E.coli</i> / <i>B.sub</i>	36/135	LS	SL	LL	steric/hyd
<i>E.coli</i> / <i>B.sub</i>	47/149	WH	FY	WY	steric/hyd
<i>E.coli</i> / <i>B.sub</i>	62/63	LS	VT	LT	steric
<i>E.coli</i> / <i>B.sub</i>	62/69	LD	VE	LE	steric
<i>E.coli</i> / <i>B.sub</i>	62/78	LV	VL	LL	steric
<i>E.coli</i> / <i>B.sub</i>	62/99	LV	VL	LL	steric
<i>E.coli</i> / <i>B.sub</i>	62/104	LL	VF	LF	steric
<i>E.coli</i> / <i>B.sub</i>	80/127	ED	DE	EE	steric
<i>E.coli</i> / <i>B.sub</i>	80/156	EL	DY	EY	steric
<i>E.coli</i> / <i>B.sub</i>	98/146	RQ	QK	RK	chg/steric/hyd
<i>E.coli</i> / <i>B.sub</i>	127/129	DE	ED	DD	cavity
<i>E.coli</i> / <i>B.sub</i>	129/156	EL	DY	EY	steric
<i>B.sub</i> / <i>E.coli</i>	17/63	DT	ES	DS	cavity
<i>B.sub</i> / <i>E.coli</i>	36/135	SL	LS	SS	cavity/hyd
<i>B.sub</i> / <i>E.coli</i>	80/127	DE	ED	DD	cavity
<i>B.sub</i> / <i>E.coli</i>	92/103	FL	MF	FF	hyd
<i>B.sub</i> / <i>E.coli</i>	98/146	QK	RQ	QQ	chg/steric/hyd
<i>B.sub</i> / <i>E.coli</i>	127/129	ED	DE	EL	steric

*Clashes formed may be due to departure from volume (steric hindrance-steric or cavity formation), charge (chg), or hydrophobicity (hyd) values observed in the protein family.

Chapter 5: Design of Combinatorial Protein Libraries of Optimal Size

Section 5.1: Background

The directed evolution of variants of a single gene [202] or a family of genes [162] coupled with a screening or selection step has emerged as a dominant strategy for creating proteins with improved or novel properties [203]. Recent developments in methods for directed evolution have led to new approaches [36, 165, 204] for creating diverse combinatorial libraries with tunable statistics irrespective of sequence homology. Two of these methods, GeneReassembly [165] and Degenerate Homoduplex Recombination (DHR) [204], use synthesized degenerate oligonucleotides for tailoring the diversity of a library. These oligonucleotides are designed to include coding information for the polymorphisms present in the parental set, while also including “customized” sequence identity at predetermined locations enabling annealing-based recombination. The “customized” sequence identity enables the targeted introduction of crossovers at only desired positions. Alternatively, in sequence-independent site-directed chimeragenesis (SISDC) [36], the exact location of crossovers is predetermined by the use of marker tags for endonuclease recognition. These are two examples out of many currently available protocols that are capable of creating the desired level and type of diversity in a combinatorial library [203].

Despite these developments, protein engineering remains a formidable task because it is still unclear what should the optimal level and type of diversity be for sampling the sequence space spanned by the parental sequence set [164, 205]. Most proteins in nature exhibit complex networks of dynamic interaction for function [155, 182, 197, 206]. Therefore, a large number of crossovers between parental sequences is

likely to disrupt vital interactions [135, 137, 207, 208] rendering most hybrids non-functional. In fact, it is commonly observed that the average activity of a library tends to drop off as parental sequence similarity decreases [164, 165]. On the other hand, a combinatorial library generated by introducing only a few crossovers will sample only a very small portion of sequence space by retaining many large contiguous parental sequence stretches. Therefore, a key open challenge is how to *a priori* identify the optimal design of a library. This entails the identification of (i) the optimal library size, (ii) number and location of junction points, and (iii) the parental sequences that contribute a fragment at each one of the junction points (see Figure 5.1).

A number of strategies have been developed to assess the quality of a library based on sequence and/or structural information encoded within the parental/family sequences to guide the design of combinatorial libraries [135, 137, 207, 208]. Typically this involves the definition of a scoring metric for evaluating the fitness of hybrid protein sequences against the parental sequences. This concept was pioneered with the development of SCHEMA algorithm [135] that hypothesizes that structural disruptions are introduced when a contacting residue pair in a hybrid has differing parental origins. Hybrids are scored for stability by counting the number of disruptions [136, 209]. Recently, a dynamic programming algorithm was proposed [209] that identifies the location of junction points that minimize SCHEMA disruption without allowing for parental fragment skipping. Alternatively, a number of methods have been developed in our group based on (i) mean-field energy calculations to infer correlations in substitution patterns (SIRCH [137]), (ii) pinpointing property value deviations (i.e., charge, volume, and hydrophobicity) from parental sequences [208], and (iii) family sequence statistics

for clash identification (FamClash[207]). Comparisons with experimental studies [136, 207, 208] have shown that crossovers are indeed preferentially allocated to avoid the predicted clashes among functional hybrids. Interestingly, using FamClash [207] we demonstrated in one case that hybrid activity levels were inversely proportional to the number of clashes in these hybrids.

These methods hint at a design strategy that forms the basis for the computational design procedure OPTCOMB introduced in this article. OPTCOMB pinpoints the location of junctions between fragments as well as their sizes and their parental origins such that the number of clashes between the fragments constituting the library is minimized. Two optimization models are considered abstracting two classes of experimental strategies for combinatorial library generation: (i) no restrictions are imposed on the contributing parental sequences (e.g., SISDC (see Figure 5.1a)) and (ii) restrictions can be imposed on the set of oligomers being contributed by the parental sequences in certain locations (e.g., DHR and GeneReassembly (see Figure 5.1b)). Both optimization models are tested on the computational design of a combinatorial library formed by three dihydrofolate reductase (DHFR) sequences from *Escherichia coli*, *Bacillus subtilis*, and *Lactobacillus casei*.

Section 5.2: OPTCOMB Modeling Framework

The design of a combinatorial library entails a number of discrete decisions such as (i) the placement and the number of junction points to be selected, (ii) whether or not a given position along the sequence is a junction point, and (iii) if a particular parental sequence contributes a fragment/oligomer at a given junction point. To model these decisions, the OPTCOMB optimization models draw upon mixed-integer linear

programming formulations that use binary variables to mathematically represent these discrete decisions. These binary variables act as on/off switches that encode, for instance, the presence/absence of a junction point. The OPTCOMB procedure makes use of models ***M1*** and ***M2*** corresponding to the experimental setups illustrated by Figures 1a and 1b respectively. Specifically, model ***M1*** abstracts experimental protocols where all parental sequences contribute a fragment at each one of the junction points. The design variables are binary variables that denote the presence or absence of a junction point along the sequence. On the other hand, model ***M2*** abstracts experimental protocols where “skipping” of parental fragments is permitted. Additional design variables are included in the model to account for whether or not a particular parental sequence contributes a fragment at a junction point. In both ***M1*** and ***M2***, the design variables are adjusted such that the total number of clashing residue pairs between fragments that constitute the library is minimized. These clashes can be identified using many available computational approaches [135, 137, 207, 208].

In addition to the constraints included in the two models that penalize the simultaneous selection of clash forming fragments, additional constraints can be added to impose additional requirements. For example, such requirements may include the preservation of two or more residues to ensure that crucial interactions for catalysis or binding with external molecules are retained [132, 210]. Constraints can also be included to guide the selection of junction points based on user-defined requirements. For example, constraints can be used to direct selection of junction points within loop regions [211] so that structural elements (i.e., α -helices, β -sheets, etc.) are not disrupted enabling the swapping of low energy secondary structures [134]. In addition, constraints can be

incorporated to minimize bias in family DNA shuffling so that each of the parental sequences contributes a similar number of fragments/oligomers to the library or alternatively to restrict crossover positions to regions of high sequence identity for proper ligation. The inclusion of such constraints in the current implementation, though not explicitly covered here, is quite straightforward.

The simpler model **MI** is applicable when no restrictions are imposed on the contributing parental sequences (Figure 5.1a). The only design variables whose values need to be determined are the locations of junction points. The sets, parameters, and variables used in model **MI** are described below.

Sets:

$k, k_1, k_2 \in \{1, 2, \dots, K\}$ = set of parental sequences

$i, i_1, i_2 \in \{1, 2, \dots, I\}$ = set of aligned positions

Parameters:

N = Number of oligomers

L_{min} = Length of shortest allowable oligomer

L_{max} = Length of longest allowable oligomer

$C_{i_1 i_2}^{k_1 k_2} = 1$ if a clash exists between residue i_1 of parental sequence k_1 and residue i_2 of parental sequence k_2 ; $i_1 < i_2$; $k_1 \neq k_2$
 $= 0$ otherwise

Variables:

$Y_i = 1$ if an oligomer starts at position i (i.e., a junction point)
 $= 0$ otherwise

$Z_{i_1 i_2} = 1$ if there exists at least one pair of parental sequences for which there is a clash between residues at positions i_1 and i_2 .
 $= 0$ otherwise

Note that here the values assigned to parameters $C_{i_1 i_2}^{k_1 k_2}$ are either 1 or 0 depending on whether there exists a clash between the two residues. Alternatively, continuous values (e.g., between zero and one) that quantify the severity of the clashes could also be used.

Based on the above defined sets, parameters, and variables, the model **MI** of OPTCOMB yields an optimization problem implemented as the following mixed-integer linear programming (MILP) formulation.

$$\underset{Y_i \in \{0,1\}}{\text{minimize}} \sum_{i_1=I}^I \sum_{\substack{i_2=I \\ i_2 > i_1}}^I \sum_{k_1=1}^K \sum_{k_2=1}^K Z_{i_1 i_2} \cdot C_{i_1 i_2}^{k_1 k_2} \quad (1)$$

$$\sum_{i=I}^I Y_i \geq N \quad (2)$$

$$\sum_{i'=i}^{i+L_{min}-1} Y_{i'} \leq 1, \quad i = 1, 2, \dots, I - L_{min} + 1 \quad (3)$$

$$\sum_{i'=i}^{i+L_{max}-1} Y_{i'} \geq 1, \quad i = 1, 2, \dots, I - L_{max} + 1 \quad (4)$$

$$Z_{i_1 i_2} \leq \sum_{i=i_1+1}^{i_2} Y_i, \quad \forall (i_1, i_2 > i_1, k_1, k_2) \text{ such that } C_{i_1 i_2}^{k_1 k_2} = 1 \quad (5)$$

$$Z_{i_1 i_2} \geq Y_i, \quad \forall (i_1, i_2 > i_1, k_1, k_2) \text{ and } i = i_1 + 1, i_1 + 2, \dots, i_2 \text{ such that } C_{i_1 i_2}^{k_1 k_2} = 1 \quad (6)$$

$$\sum_{i=L_{max}+1}^{I-L_{min}+1} Y_i = 1 \quad (7)$$

$$0 \leq Z_{i_1 i_2} \leq 1 \quad ; \quad Y_{i=I} = 1 \quad (8)$$

The objective function (Equation 1) of model **MI** entails the minimization of the number of clashes between oligomers/fragments selected for library design. Constraint 2 ensures that the number of oligomers present is greater than or equal to some specified target thus establishing the library size. The lower and upper bounds on the lengths of all oligomers is enforced by constraints 3 and 4 respectively. Typically these lengths are determined based on the specifics of the ligation protocol (e.g., GeneReassembly (39-60 nucleotides or 13-20 amino acids) [165], DHR (54-72 nucleotides or 18-24 amino acids) [204]). Note that the oligomer size ranges (L_{min} , L_{max}) determine the range of values that N can take and therefore indirectly determine the library size. For a given value of L_{min} and L_{max} , the values of N can range between $N_{min} = K \times \lfloor I / L_{max} \rfloor$ and $N_{max} = K \times \lfloor I / L_{min} \rfloor$, where $\lfloor \bullet \rfloor$

corresponds to the floor function. Therefore, the library size will range between $K^{N_{min}/K}$ and $K^{N_{max}/K}$. Clearly, as the oligomer sizes reduce, the parental sequences can be divided into larger number of fragments allowing a larger number of combinations of these fragments to be available for the construction of hybrids. Equation 5 in conjunction with equation 6 determines whether a clash is formed between any two positions (i_1, i_2) of the selected fragments from parents (k_1, k_2). Equation 7 ensures that the last fragment of each parental sequence falls within the allowable range of fragment lengths.

Note that in model **M1**, the included constraints ensure that all parental sequences must contribute a fragment at all junction points without skipping. Therefore, the only means of clash relief is the judicious selection of junction points such that the minimum number of clashes is formed while ensuring that minimum and maximum fragment size limits are satisfied. Alternatively, model **M2** allows for more flexibility as it accounts for the “skipping” of parental fragments. Clashes are relieved based on the selection of crossover positions and also on the choice of parental fragments at each one of the junction points (Figure 5.1b). This additional complexity requires additional variables and constraints to capture information on the selection/rejection of fragments of different parental sequences at each one of the junction points. Note that by restricting the contributing parents at each one of the junction points many more clashes can be relieved for the same number of junction points. Model **M2** retains all the variables defined for model **M1** in addition to the following new ones:

New Variables:

$$\begin{aligned}
 y_{ik} &= 1 && \text{if a new oligomer starts at position } i \text{ for parent } k \\
 &= 0 && \text{otherwise} \\
 Y_i &= 1 && \text{if a new oligomer starts at position } i \text{ for at least one parent} \\
 &= 0 && \text{otherwise}
 \end{aligned}$$

$$Z_{i_1 i_2}^{k_1 k_2} = 1 \quad \text{if residues } i_1 \text{ of parent } k_1 \text{ and } i_2 \text{ of parent } k_2 \text{ are selected and } C_{i_1 i_2}^{k_1 k_2} = 1$$

$$= 0 \quad \text{otherwise}$$

$$\text{minimize} \sum_{y_{ik}, Y_i \in \{0,1\}} \sum_{i_1=1}^I \sum_{i_2=1}^I \sum_{k_1=1}^K \sum_{k_2=1}^K Z_{i_1 i_2}^{k_1 k_2} \cdot C_{i_1 i_2}^{k_1 k_2} \quad (9)$$

$$\sum_{k=1}^K \sum_{i=1}^I y_{ik} \geq N \quad (10)$$

$$\sum_{i'=i}^{i+L_{\min}-1} Y_{i'} \leq 1, \quad i = 1, 2, \dots, I - L_{\min} + 1 \quad (11)$$

$$\sum_{i'=i}^{i+L_{\max}-1} Y_{i'} \geq 1, \quad i = 1, 2, \dots, I - L_{\max} + 1 \quad (12)$$

$$Y_i \geq y_{ik}, \quad i = 1, 2, \dots, I \text{ and } k = 1, 2, \dots, K \quad (13)$$

$$Y_i \leq \sum_{k=1}^K y_{ik}, \quad i = 1, 2, \dots, I \quad (14)$$

$$Z_{i_1 i_2}^{k_1 k_2} \leq \sum_{i=i_1+1}^{i_2} y_{ik_1} \cdot y_{ik_2}, \quad \forall (i_1, i_2 > i_1, k_1, k_2) \text{ such that } C_{i_1 i_2}^{k_1 k_2} = 1 \quad (15)$$

$$Z_{i_1 i_2}^{k_1 k_2} \geq y_{ik_1} \cdot y_{ik_2}, \quad \forall (i_1, i_2 > i_1, k_1, k_2) \text{ and } i = i_1 + 1, i_1 + 2, \dots, i_2 \text{ such that } C_{i_1 i_2}^{k_1 k_2} = 1 \quad (16)$$

$$\sum_{i=L_{\max}+1}^{I-L_{\min}+1} Y_i = 1 \quad (17)$$

$$0 \leq Z_{i_1 i_2}^{k_1 k_2} \leq 1; \quad y_{i=L,k} = 1 \quad (18)$$

Note that equations 15 and 16 involve the product of binary variables. This is linearized by introducing a new set of variables $w_{ik_1 k_2} = y_{ik_1} \cdot y_{ik_2}$ to exactly recast the product as a set of linear constraints [212]:

$$y_{ik_1} \cdot y_{ik_2} = w_{ik_1 k_2}$$

$$w_{ik_1 k_2} \leq y_{ik_1}; w_{ik_1 k_2} \leq y_{ik_2}; w_{ik_1 k_2} \geq y_{ik_1} + y_{ik_2} - 1; 0 \leq w_{ik_1 k_2} \leq 1$$

The objective function (Equation 9) entails the minimization of the number of clashes between fragments that constitute the library. Equation 10 ensures that the total number of oligomers selected for library design is greater than some specified lower bound. The lower and upper bounds on the lengths of all oligomers is enforced by constraints 11 and

12 respectively. Equation 13 ensures the presence of a junction point if a particular parent contributes a fragment starting at that position. Equation 14 ensures that at least one parental sequence contributes a fragment at any given junction point. Equation 15 in conjunction with Equation 16 determines whether a clash is formed between any two positions (i_1, i_2) of the selected fragments (k_1, k_2) . Finally, equation 17 ensures that the length of the last segment of each parental sequence falls between L_{min} and L_{max} .

The solution of the OPTCOMB models (**M1** or **M2**) provides the complete design of the combinatorial library of a given specified size that minimizes the presence of clashes. By successively varying the number of junction points or fragments (N), a trade-off curve between library size and percent of clash-free variants (or the average number of clashes per hybrid) can be generated. This curve provides a systematic way for determining the optimal library size given the set of parental sequences and the residue clash map. Note that in this study we have used the percent of clash-free hybrids in a library as a surrogate measure of library quality. However, the OPTCOMB model can also be used for cases where the metric of library quality is different. In such a case, the objective and scoring $\left(C_{i_1 i_2}^{k_1 k_2}\right)$ functions will need to be appropriately defined. For example, when the metric of quality is the average stability of the library, the scoring function (or the number of clashes here) should be a descriptor of stability [208] rather than of activity.

Section 5.3: Results

The optimal trade-off between library size and clashes is examined using OPTCOMB for combinatorial libraries composed of the well studied dihydrofolate reductase (DHFR) proteins from *Escherichia coli*, *Bacillus subtilis*, and *Lactobacillus*

casei. Clashes between residues of different parental sequences are first derived using the FamClash [207] procedure. According to the FamClash procedure clashes occur when a statistically significant deviation in the properties (such as charge, volume, and hydrophobicity) of pairs of residues in the hybrids are observed from the values observed in the protein family [207]. Similar results are observed when clashes are identified based on steric hindrance, charge repulsion, and hydrogen bond disruption [208]. The DHFR protein family sequence data required for clash prediction is downloaded from the PFAM [213] database including 300 sequences in total. Out of the total 50 clashes identified, 20 clashes are between *E. coli*-*B. subtilis* (sequence identity = 44.0%), 9 clashes are between *B. subtilis*-*L. casei* (sequence identity = 36.10%), and 21 clashes are between *L. casei*-*E. coli* (sequence identity = 28.4%) sequence pairs (see Figure 5.2). Notably, most of the clashes (41 out of 50) are associated with the *E. coli* sequence even though it is not the most divergent of the three sequences. These clashes are encoded using the $C_{i_1 i_2}^{k_1 k_2}$ parameters and imported into the OPTCOMB procedure to guide the design of the combinatorial library. The OPTCOMB optimization models (**M1** and **M2**) are solved using the CPLEX solver [214] accessed via the GAMS [215] modeling environment. This computational base enables us to explore the following questions:

1. How many clashes remain in the combinatorial library designs obtained using models **M1** and **M2** as a function of library size and how does this number compare with randomly generated libraries?
2. What are the oligomer/fragment tiling characteristics of the optimally designed libraries?

3. Is there an optimal library size that leads to a minimum of retained clashes per hybrid?
4. What is the effect of library size on the relative contribution of fragments by the three parental sequences, clash distribution, and the tiling combinations?

To answer the first question, model (*M1* and *M2*) driven designs are first contrasted against randomly generated libraries to assess whether the systematic selection of junction points affords significant gains over random choices. The optimal designs obtained using models *M1* and *M2* are also compared against each other to infer the extent of improvement achieved by disallowing fragments from participating in library design. Both OPTCOMB models (*M1* and *M2*) are solved for different values of N (number of oligomers) allowing for a minimum and maximum oligomer length of 15 and 30 residues respectively covering the range of length of oligonucleotides used in the GeneReassembly and DHR protocols [165, 204]. Library designs of increasing size are generated computationally for N equal to 15, 18, 21, 24, 27 and 30. In addition, random tiling combinations are generated for the same number and length of oligomers using the same design constraints outlined for models *M1* and *M2* (see Figure 5.1a and 1b) and the average number of clashes per hybrid are calculated for different library sizes. As expected, we find that in both cases the libraries designed using OPTCOMB include much fewer clashes than the randomly generated libraries. Figure 5.3 depicts the number of clashes (♦) retained between optimally designed oligomers using model *M1* (Figure 5.3a) and model *M2* (Figure 5.3b) against library size. These clashes are contrasted against the average number of clashes (▲) between oligomers for randomly generated tiling combinations for the two cases. These results clearly demonstrate that substantial

improvement in library design can be made by pro-actively minimizing clash retentions. Comparisons between optimal designs obtained with models **M1** and **M2** reveal that the additional flexibility of “skipping” of certain parental fragments at key junction positions reduces clash retention by approximately 50% (See Figure 5.3) for the same library size.

The second question focuses on the tiling characteristics of optimal library designs. We find that, in general, the optimal designs obtained using model **M2** involve fragments of roughly similar lengths with, however, widely varying contributions from different parental sequences. In contrast, optimal designs using model **M1** typically employ non-uniform fragment lengths. For example, Figure 5.4 shows the optimal tiling pattern obtained using model **M2** for $N=21$. Only a small portion of the *E. coli* sequence is present while most of *L. casei* and the entire *B. subtilis* sequence are participating in the optimal library design reflecting that OPTCOMB systematically disallows fragments from the *E. coli* sequence implicated in clash formation. The concatenation of the oligomers shown in Figure 5.4 yields a library composed of 1,536 hybrids that avoid 44 out of the 50 clashes identified using FamClash. The remaining six clashes are shown as arcs connecting the two implicated residues (see Figure 5.4). In contrast, libraries designed by random selection of junction points and sequence tiles involve on average 26 clashes. Notably, the designed crossover positions do not follow any easily discernable patterns in terms of the underlying secondary structure. Although many of the designed crossovers fall within the loop regions, many of them are found to be within α -helices and β -sheets. The crossover positions also seem to be equally distributed between conserved and non-conserved stretches of parental sequences.

The third question examines the optimal trade-off between library size and quality

exemplified by the number of clashes between fragments chosen for the library design, the percent of clash-free hybrids and the average number of remaining clashes per hybrid. Clearly, the number of both the clash-free and clash-containing hybrids increases with increasing library size. However, because there is a limit to the number of sequences that can be screened, we use the percent of clash-free hybrids as a metric of quality. Trade-off curves for these three different library quality metrics are generated using model **M2** to assess library quality (see Figure 5.5). Figure 5.5a shows the trade-off curve between library size and number of clashes between fragments that constitute the library for different values of N . The number of clashing residue pairs is, as expected, monotonically increasing with library size. Interestingly, we find that the rate of increase, beyond a library size of approximately 1.6×10^3 (shown as dashed line in figure 5.5a), is dramatically enhanced. It appears that beyond this size threshold OPTCOMB runs out of nearly clash-free fragment combinations and thus clash-forming fragments must be used to meet the increased library size requirements. The same behavior is observed for libraries designed using varying ranges of fragment length implying a global trend. This transition point also shows prominently in the trade-off curves between (i) the percent of clash-free hybrids and the library size (see Figure 5.5b) and (ii) the average number of clashes per hybrid versus the library size (Figure 5.5c). We find that the percent of clash-free hybrids increase up to this transition point and afterwards it begins to decline (Figure 5.5b). Accordingly, the average number of clashes per hybrid decreases up to this point and begins to rise again (Figure 5.5c). The reason for this trend is that for small library sizes the OPTCOMB model chooses the junction points and the contributing parental sequences such that most of the clash-forming fragments are avoided. However, there is

only a limited number of clash-free fragment combinations all of which are selected before the threshold library size. Therefore, to obtain library sizes beyond this threshold size, the model is forced to choose fragments involving increasingly higher number of clashes resulting in the decline in the percent of clash-free hybrids (or alternatively resulting in the increase in the percent of hybrids with clashes) in the library. It is noteworthy that this transition point is at approximately for the same library size (or value of N) for all library quality metrics (see Figure 5.5 a, b, and c). The *a priori* identification of this optimal library size is of considerable importance to the application of directed evolution protocols by answering the question of what is the appropriate library size that best balances diversity with quality for a given protein engineering task.

As expected the optimal library size is a strong function of the fragment/oligomer sizes and is found to increase substantially with decreasing fragment length ranges. Figure 5.6 depicts the optimal library size for different ranges of fragment sizes. Smaller fragment sizes afford more fragment choices for library design and significantly more tiling combinations to choose from. Because different experimental protocols for directed evolution have different requirements on fragment lengths, the trade-off curves such as the one shown in Figure 5.6 can aid in selecting the correct protocol based on library size or the sequence space to be explored.

The last question explores the effect of combinatorial library size (or N) on the tiling combination, the clash distribution, and the relative contribution of the three parental sequences towards the library. We find that the optimal tiling combination and the relative contribution of the parental sequences change significantly when N is varied (see Figure 5.7) and that there exists persistently “skipped” fragments (e.g., residues 80-

130 of the *E. coli* sequence) in the tiling combinations. For example, we observe that the contribution of the *B. subtilis* and *L. casei* sequences to the library increases with N . Interestingly, we find that while initially the *E. coli* sequence contribution to the library is equal to the one from *L. casei* (~40% each for $N = 15$), it rapidly drops to 10% (for $N = 18$) after which it increases to meet the increasingly higher required numbers of oligomers (see Figure 5.7). At the end, ($N = 30$) there are no skipped fragments thus recovering the solution of model **MI**. Although, the fragment sizes are allowed to vary from 15 to 30 residues, we find that the fragment size chosen in the library design are fairly uniform and range between 15-18 residues. Clearly, smaller fragments allow for more flexibility and therefore enhance the chances of avoiding the clashes. The largely non-varying fragment sizes imply that the location of the junction points as well do not change significantly (see Figure 5.7). The distribution of number of hybrids based on the number of clashes follow a lognormal distribution with the number of clashes in the hybrids varying from 0 to 10. The distribution of clashes is narrow for small values of N and broadens with increasing N . Note that the number of clashes present in the hybrids of a given library vary between 0-10 and is significantly lower than the number of clashing residue pairs that can be formed between the fragments that constitute the library (as many as 39 for $N = 30$).

Section 5.4: Summary

In this work, the OPTCOMB (Optimal Pattern of Tiling for Combinatorial library design) procedure was introduced for the optimal design of synthetic oligomer ligation based protocols [36, 165, 204]. The capabilities of OPTCOMB were demonstrated by computationally designing recombinant libraries composed of sequences from

Escherichia coli, *Bacillus subtilis*, and *Lactobacillus casei* dihydrofolate reductase (DHFR) proteins. The key result of this study is the computational verification of the existence of an optimal library size that best balances library diversity and quality. The optimal library size was found to be a strong function of fragments size and involved the coordinated skipping of certain parental fragments.

Clearly, the obtained results depend on the accuracy of the clash prediction frameworks [135, 137, 207, 208]. We expect that more accurate clash prediction methods will become available in the future that can capture backbone movement in the hybrids through the use of sophisticated potential energy/scoring functions [30, 33, 216]. Nevertheless, OPTCOMB provides a versatile framework that can handle the information generated by various clash prediction methods [135, 137, 207, 208].

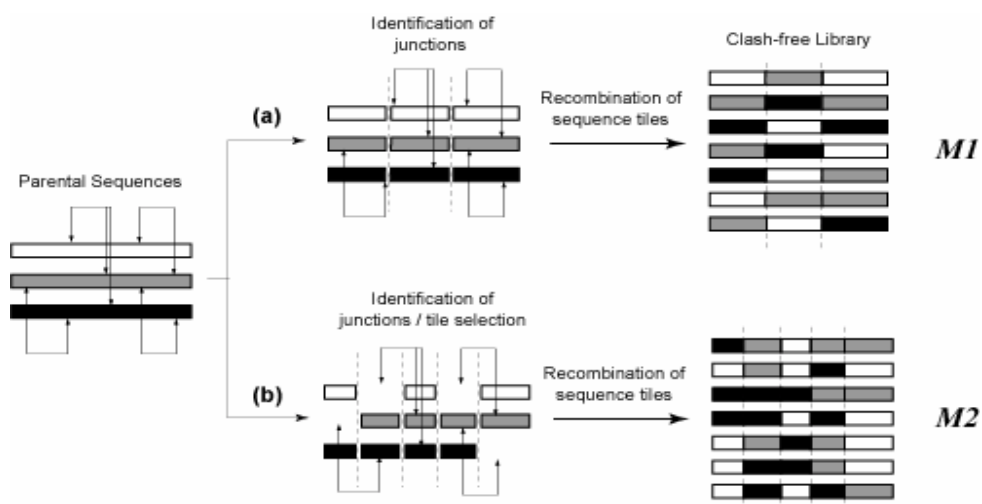


Figure 5.1: A pictorial representation of the example where three parental sequences form a combinatorial library through recombination. The clashes between different residues are shown as double-headed arrows. The junction points are shown as dashed lines. The combinatorial libraries are designed using two different design rules: **(a)** all parental sequences contribute fragments at each of the junction points, and **(b)** selective restrictions are imposed on the set of oligomers being contributed by the parents.

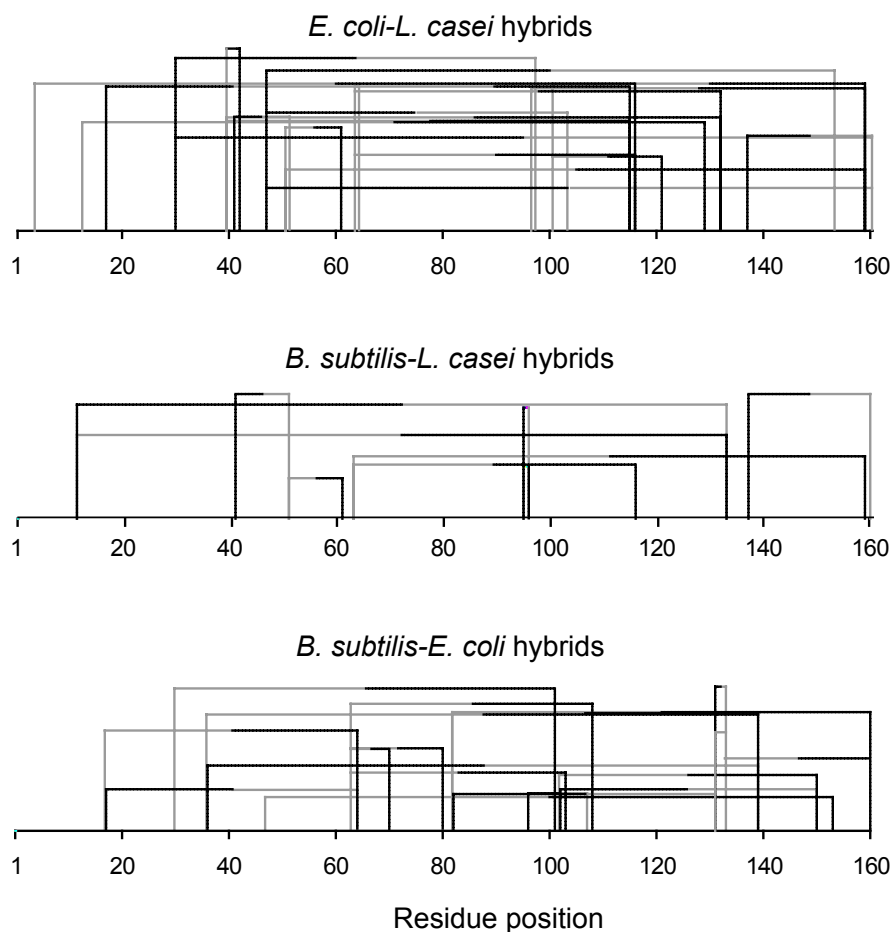


Figure 5.2: Clash maps determined using the FamClash procedure [207] corresponding to the three different sequence combinations (*E. coli-L. casei* (black-gray), *B. subtilis-L. casei* (black-gray), and *B. subtilis-E. coli* (black-gray)). Note that the color shown in the parentheses alongside each pair of sequences correspond to the corresponding pair of parental sequences. Residues in the hybrids retained from parental sequences with the same color as the arc connecting them lead to a clash.

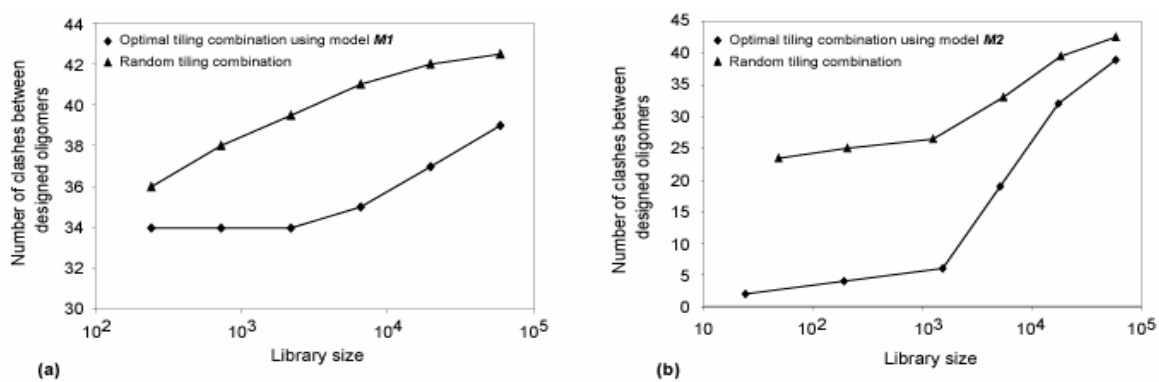


Figure 5.3: Plot of the number of clashes between optimally designed oligomers (♦) using models (a) *M1* and (b) *M2* against library size. The average numbers of clashes between randomly generated designs (▲) for various library sizes are also shown.

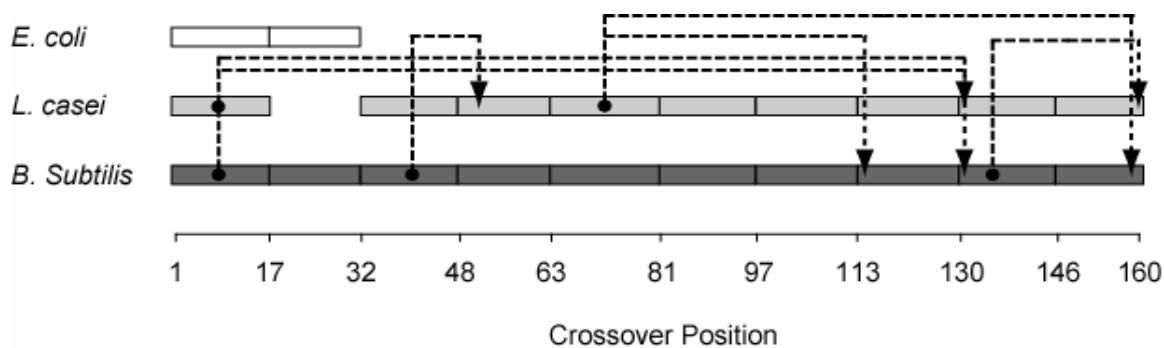


Figure 5.4: Results obtained using model *M2* for minimum and maximum fragment lengths of 15 and 30 residues respectively and $N = 21$. The clashes that are retained are shown as dashed arcs with the position of the first residue of a clashing pair in the hybrid being represented by a dot (●).

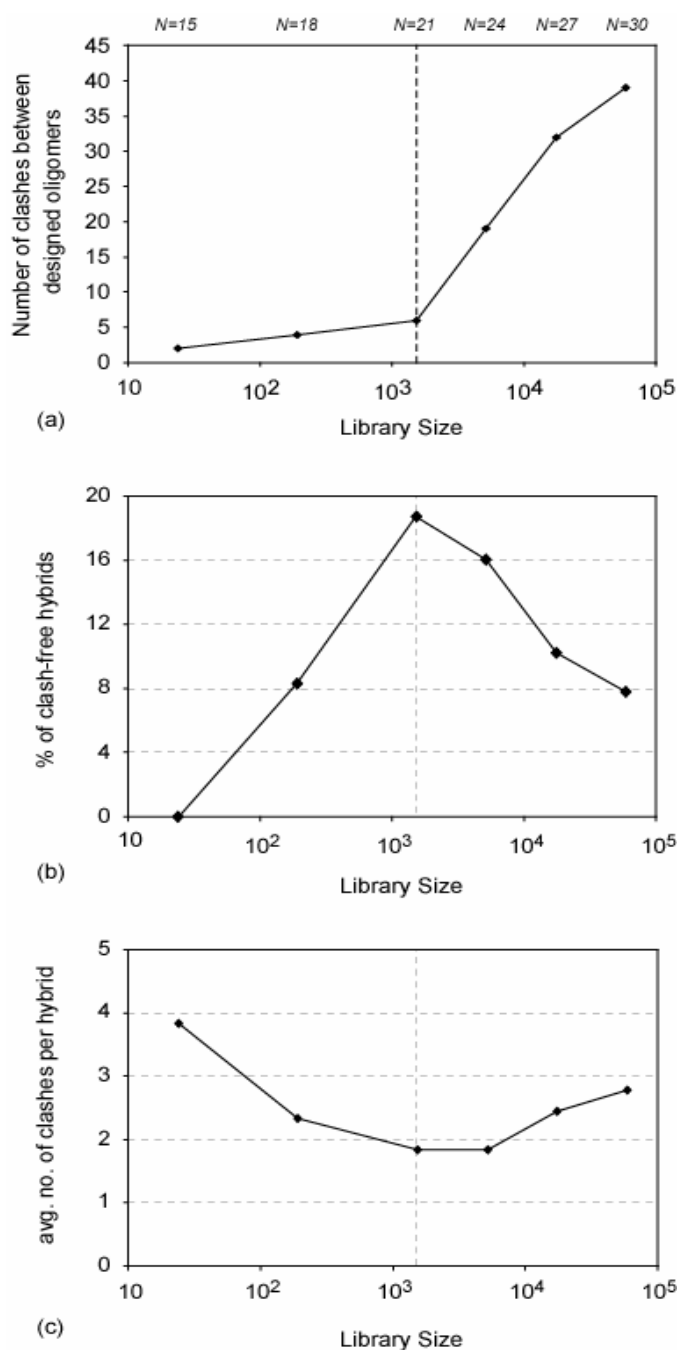


Figure 5.5: (a) Plot of the number of clashes between selected parental fragments (corresponding to $N = 15, 18, 21, 24, 27$, and 30 ; $L_{min} = 15$ and $L_{max} = 30$) forming the library against library size. There is an optimal library size $\sim 1.6 \times 10^3$ (shown with a dashed line) beyond which the number of clashes increases significantly. (b) Plot of the

percent of clash-free hybrids versus library size. Notably, at the transition point/optimal library size (1.6×10^3) the percent of clash-free hybrids is at a maximum. **(c)** Plot of the average number of clashes per hybrid versus library size. Again the minimum number of clashes is observed at the optimal library size (1.6×10^3).

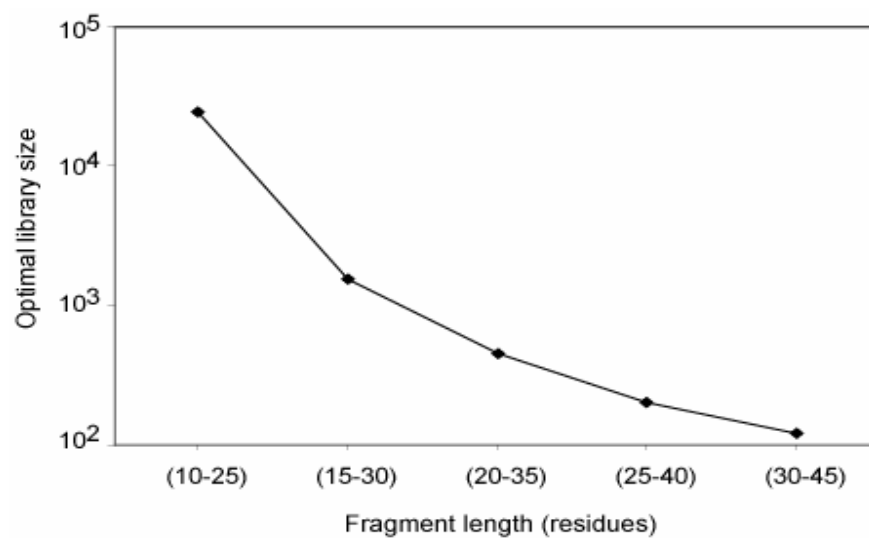


Figure 5.6: Plot of the optimal library size for different ranges (10-25, 15-30, 20-35, 25-40, and 30-45) of fragment lengths. The optimal library size decreases with increasing fragment sizes.

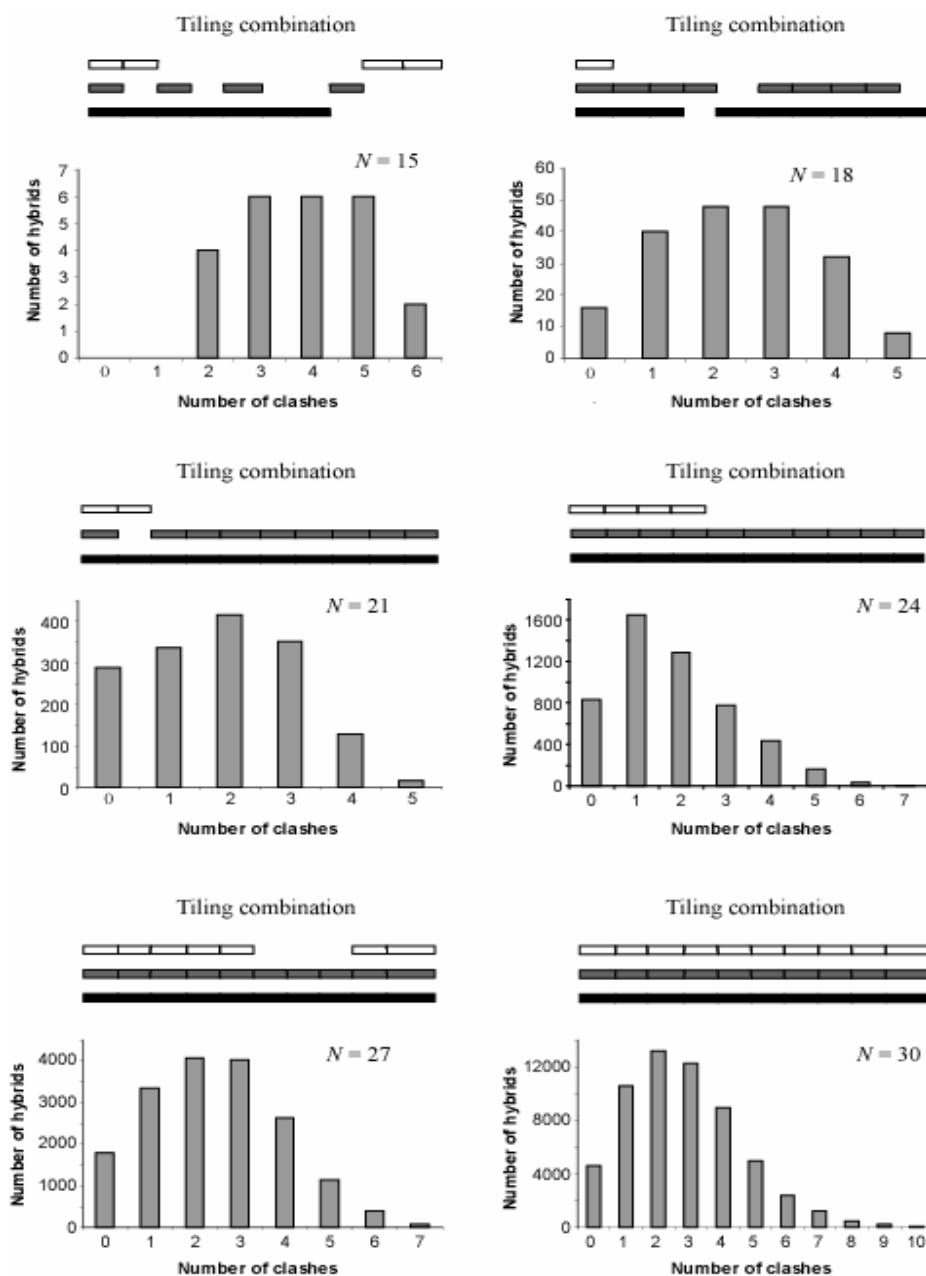


Figure 5.7: The tiling choices and the clash distributions for the hybrids for $N=15, 18, 21, 24, 27,$ and 30 .

Chapter 6: IPRO: An Iterative Computational Protein Library Redesign and Optimization Procedure

Section 6.1: Background

The ability to proactively modify protein structure and function through a series of targeted mutations is an open challenge that is central in many different applications. These include, among others, enhanced catalytic activity [7-9] and stability [10, 11], creation of gene switches for the control of gene expression for use in gene therapy and metabolic engineering [12, 13], signal transduction [14, 15], genetic recombination [16], motor protein function, and regulation of cellular processes (see ref. [17] for a review). This task is complicated by the fact that proteins rely on complex networks of subtle interactions to enable function [18-20]. Therefore, the effect of a mutation is difficult to assess *a priori* requiring the capture of its direct or indirect effects on many neighboring amino acids. As a result, most protein engineering paradigms involve the synthesis and screening of multiple protein candidates (protein library) as a way to enhance the odds of identifying proteins with the desired functionality level. These directed evolution design paradigms [1, 3, 21-24] typically involve juxtaposition of repeated library generation and screening (Figure 6.1). On the other hand most computational approaches for guiding protein design are focused on the downstream redesign of single parental sequences or promising hybrids (Figure 6.1).

A number of computational models and techniques have been developed (see ref. [217] for review) to aid in the *in silico* evaluation of protein redesign candidates. Typically these techniques attempt to find a single or multiple amino acid sequences that are compatible with a given three-dimensional structure specific to a targeted function (e.g., enzymatic activity). The protein fold is usually represented by the Cartesian

coordinates of its backbone atoms which are fixed in space so that the degrees of freedom associated with backbone movement are neglected. More recent approaches [44-49] allow for some backbone movement. Candidate protein designs are generated by selecting amino acid side chains (using atomistic detail) along the backbone design scaffold. For simplicity, side chains are usually only permitted to assume a discrete set of statistically preferred conformations referred to as *rotamers* (see [50]) for a review of current rotamer libraries). Thus, a protein design consists of both a residue *and* a rotamer assignment for each amino acid position. To evaluate how well a possible design fits a given fold, rotamer/backbone and rotamer/rotamer interaction energies for all the rotamers in the rotamer library are tabulated. These energies are approximated using standard force fields (*e.g.*, CHARMM [51], DREIDING [52], AMBER [53], GROMOS [54]). Scoring functions customized for protein design [55-57] (see ref. [58] for a review) typically include van der Waals interactions, hydrogen bonding, electrostatics, solvation, along with entropy-based penalty terms for flexible side-chains (*e.g.*, arginine) [30, 59-61]. Because activity level or other performance objectives are very difficult to compute directly, alternative surrogates of hybrid fitness, such as stability or binding affinity, are employed in most studies. The use of these indirect objectives further necessitates the need for designing a combinatorial library rather than a single hybrid to improve the chances of success.

Even for a small 50-residue protein, an enormous number (*i.e.*, $153^{50} \approx 10^{109}$ assuming a 153-rotamer library [218]) of designs is possible. Both stochastic and deterministic search strategies have been used to tackle the computational challenge of finding the globally optimum design within this vast search space. Despite these challenges, a number of success stories of combinatorial design for many different

applications has been reported [25-32] in the last few years demonstrating the feasibility of using computations to guide protein redesign. Briefly, successes include many-fold improvements in enzyme activity and thermostability [32, 219, 220], improved enantioselectivity [221-223], enhanced bioremediation [224-226], and even the design of genetic circuits [12, 13, 16] and vaccines [227-229]. It is increasingly becoming apparent, however, that instead of computationally generating a set of distinct protein redesigns, it is more promising to use computations to shape the statistics of an entire combinatorial library. This allows one to assess and then “steer” diversity toward the most promising regions of sequence space [230]. This paradigm is more likely to succeed compared to constructing, one at a time, protein designs. On the other end, construction of combinatorial libraries based on mutation and/or recombination without any guidance from models/computations is a daunting task because only an infinitesimally small fraction of the diversity afforded by DNA and protein sequences can be examined regardless of the efficiency of the screening procedure.

In response to these challenges, in this work we introduce a new computational procedure IPRO (Iterative Protein Redesign and Optimization) that allows for the upstream redesign of parental sequences (Figure 6.1). The key idea here is that the residue changes within the parental sequences will propagate in the combinatorial library; effectively introducing mutations within the hybrid sequences in the library (see Figure 6.1). Judicious selection of these mutations in the parental sequences can simultaneously relieve unfavorable interactions or clashes [231-233] within the hybrid sequences and therefore enhance the overall quality of the library in one step mirroring the experimental protocol design. Note that even though IPRO is geared towards parental sequence redesign, it can be used, as a

limiting case, for the redesign of a single or handful of individual sequences.

The key feature of the IPRO protocol is the cycling between sequence design, ligand re-docking, and backbone movement of a set of sequences representative of the combinatorial library. The goal of the sequence design here is to choose mutations within the parental sequences, and therefore in the hybrid sequences, that optimize the average binding energy/score (or alternative surrogates of design objectives) of the hybrid sequences in the library. The genetic algorithm of Desjarlais and Handel [65] and the Monte Carlo minimization protocol of Kuhlman and coworkers [61] involve similar sequence design and backbone perturbation moves. However, they only allow for the design of a single sequence at a time and involve full-scale optimization over rotamers for only a local backbone perturbation. On the other hand, IPRO allows for the design of the entire combinatorial library and involves optimization over the local perturbation region using a globally convergent Mixed-Integer Linear Programming (MILP) formulation. In addition, IPRO allows for the re-docking of the associated ligands (e.g., substrates, cofactors, solvent, etc.) after a pre-specified number of design iterations.

In the next section we describe in detail the IPRO procedure and introduce the globally convergent Mixed-Integer Linear Program that drives residue redesign. We also discuss the methods used for generating and identifying hybrid *E. coli/B. subtilis* dihydrofolate reductase (DHFR) and *B. subtilis/L. casei* DHFR enzymes containing single crossover positions and assays for DHFR activity. Next, we provide an example application of IPRO to highlight the features and type of output obtained with IPRO. The study involves the computational identification of parental redesigns that are likely to improve a single crossover *E. coli/B. subtilis* DHFR combinatorial library composed of 16 hybrids

[232]. We conclude by discussing the implications of our results and some of the modeling and algorithmic enhancements that we are currently incorporating to further improve the IPRO framework.

Section 6.2: The IPRO Modeling Framework

The IPRO procedure is composed of four parts (see Figure 6.2): (a) A set of hybrid sequences matching the members of the combinatorial library, if less than about one hundred, is generated. For larger libraries only a representative sample of the diversity of the combinatorial library is considered. (b) For each hybrid sequence an initial structure is computationally generated. This is a critical step as the efficacy of the identified redesigns depends heavily on the accuracy of the modeled structures. (c) A set of positions, ranging from a single residue position to the entire sequence length, to be targeted for redesign is compiled. Note that the larger the number of design positions, the larger is the search space and, therefore, the higher is the computational burden. Typically we only consider between 3-20 design positions that include residue positions within or in the neighborhood of the active site. In addition, restrictions on the type of allowable residue redesigns (e.g., hydrophobic, charged, etc.) can be imposed for each redesign position. (d) Next, a set of residue changes is identified in the parental sequences which upon propagation among the combinatorial library members lead to the optimization of the average library score (e.g., binding energy or stability [55-57]). This optimization step is carried out globally using a MILP model within a local perturbation window whereas simulated annealing is used to accept or reject the residue redesigns associated with each backbone perturbation step.

Section 6.2.1: Generating a set of sequences representative of the combinatorial library

A set of hybrid sequences is selected to exhaustively or statistically represent the

combinatorial library. This step begins with the sequence/structural alignment [146] of the parental sequences. A statistical description of the combinatorial library is obtained by considering the specifics of the combinatorialization protocol. For example, in case of DNA shuffling, models such as eShuffle [77] or those developed by Maheshri and Schaffer [234] can be used to estimate the library diversity. Alternatively, for an oligonucleotide ligation based protocol such as GeneReassembly [235], SISDC [236], and Degenerate Homoduplex Recombination (DHR) [237] a statistically unbiased sample of fragment concatenations is constructed that broadly captures the diversity of the resulting combinatorial library. In the limiting case when there is only a single starting sequence to be redesigned, IPRO reverts back to the traditional single protein sequence design procedure. Note, however, that the concept of designing for the optimum of the average of a library of sequences can also find utility in this case when not a unique but rather an ensemble of putative structures is available for the protein to be redesigned. The ensemble of modeled structures then plays the role of the combinatorial library when fed to IPRO. By optimizing with respect to the ensemble average of the putative structures a more robust redesign strategy is likely to be obtained.

Section 6.2.2: Generation of starting hybrid protein structures

The initial putative structures of the hybrid proteins forming the library are obtained by splicing fragments of the parental structures consistent with its sequence (see Figure 6.3). The coordinates of the fragment structures are taken from the structural alignment of the parental sequences. The fold at the junction point(s) typically involves a “kink” as a result of the “ad-hoc” concatenation of the parental structures which becomes even more prominent in case of insertions. This is “smoothened” by allowing the

backbone around the junction point to move. The backbone ϕ and ψ angles of seven residues on either side of the crossover position(s) are allowed to vary and their new positions are determined through energy minimization. In the current implementation of IPRO we use the CHARMM [238] energy function and molecular modeling environment. Note that during the energy minimization the bond lengths (b), bond angles (χ_1, χ_2 , etc.), and internal coordinates of the side-chains are *restrained* to their original values (b_o, χ_o) by penalizing any deviations (see equations (1) and (2)). The bond stretching is penalized using Hooke's law formula (equation (1)) and the distortions in the bond angles are penalized using the harmonic function (equation (2)). In addition, distances between certain key atoms can also be restrained using Equation (1). Note that because less energy is required to distort an angle than to stretch a bond, the force constant associated with bond angle distortion is accordingly smaller.

$$\Delta E_{bond_len_penalty} = \sum_{bonds} 1000(b - b_o)^2 \text{ kcal/mol } \text{\AA}^2 \quad (1)$$

$$\Delta E_{bond_angle_penalty} = \sum_{angles} 60(\chi - \chi_o)^2 \text{ kcal/mol rad}^2 \quad (2)$$

Alternative methods to parental fragment splicing and relaxation for modeling the hybrid structures include techniques such as homology modeling [147, 239] and *ab initio* structure prediction methods [239, 240]. After the structure of the hybrid protein is modeled, the missing hydrogen atoms are added to the hybrid protein in accordance with the standard procedure used in CHARMM [51]. Finally, the positions of the associated ligands are identified using crystallographic data (whenever available) in conjunction with the ZDOCK docking software [241, 242]. Notably the ZDOCK software allows for the user specified rough placement of the docked molecules thus significantly reducing

the computational expense of the docking calculations.

Section 6.2.3: Selecting design positions

The selection of the set of positions, that will be allowed to mutate (i.e., candidate redesign positions) for each of the parental sequences, is largely dependent on the design objective and associated surrogate criterion. Typically, design objectives involve one or more of the following: (i) protein stability, (ii) binding affinity, (iii) specific activity, and (iv) substrate specificity. Protein stability is associated with the ability of the protein to fold correctly under a set of conditions. Generally, unfavorable interactions present within the proteins such as the electrostatic repulsion, hydrogen bond disruptions, steric clashes, or a combination of these tend to prevent these proteins from folding correctly [231]. A number of structure or sequence data based (SCHEMA [135], SIRCH [233], and clash maps [231]) and functionality-based (FamClash [232]) scoring strategies can be used to quantify the extent of such unfavorable interactions in each hybrid. Residue positions that participate in a disproportionate number of such clashing interactions serve as design positions. On the other hand, when binding affinity, specificity, or specific activity is the design objective, residues within or in the neighborhood of the binding site are chosen as candidates for design. In general, the design positions are either the clashing residues, binding pocket residues, or a combination of both. In most cases the set of candidate design positions is subsequently revised (either upward or downward) by using information about the functional role of different residues often found in the literature.

Section 6.2.4: Iterative protein optimization step

The optimization procedure of IPRO involves iterating between sequence design,

backbone optimization, and ligand re-docking (see Figure 6.4). This iterative procedure involves six main steps as follows:

(i) *Backbone Perturbation*: Different backbone conformations are sampled by iteratively perturbing small regions of the backbone that are randomly chosen during each cycle along the length of the sequence (N). For this purpose, a segment (from one to five contiguous residues (k to k') excluding prolines) of the protein sequence is randomly chosen for perturbation. Because the special structure of proline makes the polypeptide backbone more rigid, prolines, whenever present, are considered part of the backbone. The ϕ and ψ angles of the positions within the perturbation window are perturbed by up to $\pm 5^\circ$ from their current values. The probability distribution of the perturbation (between -5° to $+5^\circ$) follows a Gaussian distribution with a mean of zero and a standard deviation of 1.65° . This ensures that smaller perturbations are chosen more often (64% chance that the perturbations are between $\pm 1.65^\circ$) compared to larger ones that in most cases are found to result in steric clashes. Note that the backbone conformations of both parental and hybrid sequences are perturbed during each cycle. While the perturbation positions are the same for every hybrid and parental sequences, the perturbation magnitude in the backbone angles may vary. This allows different parental and hybrid sequences to assume diverse backbone conformations to better accommodate the differing side chains.

(ii) *Rotamer-Rotamer/Rotamer-Backbone energy tabulations*: Given the backbone conformations determined in Step (i) and the rotamers and rotamer combinations permitted at each position, this step involves the calculation of the interaction energies of all rotamer-backbone and rotamer-rotamer combinations within an interaction-dependant cut-off distance (cut-off distance for van der Waals = 12 Å, hydrogen bond = 3 Å, and

solvation = 9 Å). This energy tabulation must be performed separately for each hybrid and parental structure. The computational expense is reduced by only updating the part of the tables that are affected by the current perturbation. These values are then fed as parameters to the side-chain/sequence optimization model.

(iii) *Side-chain/Sequence Optimization*: This step optimizes the amino acid choices and conformations (rotamers) for the given backbone structure over a 10-15 residue window that includes the perturbation positions and five residue positions flanking it on either side (see Figure 6.5). Specifically, the design positions within the perturbation region are permitted to change amino acid type while the flanking residue positions (5 residues on either side) can only change rotamers but not the residue type. This entails two discrete decisions: (1) identifying the choice of amino acid at any given position; and (2) selecting the rotamer of the chosen amino acid that minimizes the selected surrogate objective function. To model these discrete decisions, IPRO draws upon Mixed Integer Linear Programming (MILP) optimization model formulations that use binary variables to mathematically represent these discrete decisions.

For clarity of presentation, we will first describe the MILP formulation for the special case, i.e., redesign of a single parental sequence. This description will then serve as the starting point for the more general combinatorial library design optimization formulation. In both cases, the set of allowed side-chain conformations and amino acid choices at any position is encoded within sets (R_i and R_{ih} respectively) where i denotes the residue position and h denotes a hybrid sequence in the combinatorial library in case of parental sequence redesign. Positions within the perturbation window but outside the set of redesign candidates are restricted to the original amino acid type but can change

their rotamer state. All other residue positions outside the perturbation window are fixed and cannot change either residue type or rotamer. As expected, the parental sequence redesign problem is much more complex than the single hybrid design. This is because a substituted residue need not assume the same rotamer conformation in each library member. In other words, the hybrids are “tied together” at the sequence level, but not necessarily at the rotamer level. Starting with the simpler MILP formulation for the design of a single hybrid sequence, we first outline the sets, parameters, and variables used in the model as described below:

Sets:

$k, k' \in \{1, 2, \dots, N\}$ = set of starting and ending positions for perturbation; $k < k'$
 $i, j \in \{k-5, k-4, \dots, k, \dots, k', \dots, k' + 4, k' + 5\}$ = set of positions for perturbation
 $r, s \in \{1, 2, \dots, R\}$ = set of rotamers
 R_i = set of rotamers available at position i

Binary Variables:

$X_{ir} = \begin{cases} 1, & \text{if rotamer } r \text{ is selected at position } i \\ 0, & \text{otherwise} \end{cases}$

Continuous Variables:

$Z_{irjs} = \begin{cases} 1, & \text{if rotamers } r, s \text{ are selected simultaneously at positions } i, j, \text{ respectively} \\ 0, & \text{otherwise} \end{cases}$

Parameters:

E^{sb} = substrate - backbone energy
 E_{ir}^{rb} = rotamer - backbone energy of rotamer r at position i
 E_{ir}^{rs} = rotamer - substrate energy of rotamer r at position i
 E_{irjs}^{rr} = rotamer - rotamer energy of rotamers r, s at positions i, j respectively

Based on the above defined sets, variables, and parameters, the single sequence design problem (SSDP) is implemented as the following mixed-integer linear programming (MILP) formulation which is a special case of the quadratic assignment problem [243]:

$$\text{Minimize } \sum_i \sum_r X_{ir} (E_{ir}^{rs}) \quad (3)$$

$$\sum_i \sum_r X_{ir} (E_{ir}^{rb} + E_{ir}^{rs}) + \sum_i \sum_{j>i} \sum_r \sum_s Z_{irjs} \cdot E_{irjs}^{rr} + E^{sb} \leq E_{cutoff} \quad (4)$$

$$\sum_r X_{ir} = 1, \quad \forall i; r \in R_i \quad (5)$$

$$X_{ir} = 0 \quad \forall i, r \text{ such that } E_{ir}^{rs} > \delta_i; r \in R_i \quad (6)$$

$$Z_{irjs} = X_{ir} \cdot X_{js} \quad \forall i, r, j, s; r \in R_i, s \in R_j \quad (7)$$

The objective function (Relation (3)) here entails the minimization of the binding score between the substrate and the protein as an example. The objective function can be changed depending on the design requirements. In many cases, (e.g., binding score) the objective function does not encode information about the interactions in the entire protein. Therefore, the minimization step may lead to mutations or rotamer changes that adversely affect the overall stability of the protein. Constraint 4 is included to safeguard against this by requiring that the total energy of the protein be below a prespecified cutoff value E_{cutoff} . The versatility of the adopted MILP modeling description enables the incorporation of this explicit stability requirement that is absent in most other frameworks proposed for protein design/redesign. In the same spirit, additional energy-based requirements can be imposed to ensure, for instance, retention of important hydrogen bonds between a donor and an acceptor. Constraint 5 ensures that only one rotamer is selected at any given position i along the sequence. Note that the rotamers may be that of the original residue or of other residues, depending on whether or not position i is a design position. Constraint 6 prevents any rotamers from being selected at position i that have sufficiently high energy values ($> \delta_i$) that preclude them from the optimal solution. This rotamer elimination procedure formalizes the “background optimization” concept proposed by Looger and Hellinga [244] and allows for eliminating rotamers that are

guaranteed not to be part of the optimal solution (see ref. [244] for details) . This concept allows us to *a priori* trim down the search space and therefore reduces the computational time. Appendix A describes this procedure in detail. Constraint (7) determines which rotamers r and s are simultaneously selected at positions i and j respectively. This is encoded with variable Z_{irjs} which is equal to one only if both variables X_{ir} and X_{js} are equal to one. This implies that Z_{irjs} is equal to the product of the two binary variables. These nonlinear terms are then recast into an equivalent linear form by summing Z_{irjs} over s and r , respectively as shown below:

$$\sum_s Z_{irjs} = \sum_s [X_{ir} \cdot X_{js}] = X_{ir} \sum_s [X_{js}] = X_{ir} \quad \forall i, r, j > i; \quad r \in R_i, \quad s \in R_j \quad . \quad (8)$$

$$\sum_r Z_{irjs} = \sum_r [X_{ir} \cdot X_{js}] = X_{js} \sum_r [X_{ir}] = X_{js} \quad \forall i, j > i, s; \quad r \in R_i, \quad s \in R_j \quad (9)$$

$$0 \leq Z_{irjs} \leq 1 \quad \forall i, r, j > i, s; \quad r \in R_i, \quad s \in R_j \quad (10)$$

By replacing constraint (7) with constraints (8), (9), and (10), the linearity of the SSDF formulation is preserved. The complete MILP formulation for SSDP includes constraints (3)-(10) excluding constraint (7).

Unlike the single sequence protein design formulation SSDP, the hybrid library design problem (HLDP) involves the simultaneous optimization of the hybrids (h) comprising the combinatorial library. Because the hybrid sequences in the combinatorial library are derived from the parental sequences, their amino acid composition must be restricted to the amino acid type present in the corresponding parental sequences after the targeted mutations. To this end, we introduce parameters $(v_{i'ap}, aa_{irh})$ that link the amino acid type a selected at a given position i' in parental sequence p to those present in the

hybrid sequences at the corresponding position i . In case of insertions and deletions, the positions i and i' in the hybrid and parental sequences, respectively, may not be the same. Therefore, one needs to keep track of both the parental sequence p and what position i' in that sequence corresponds to a given position i in a hybrid sequence h . Specifically, parameter $v_{i'ap}$ is equal to one if amino acid a occurs at position i' in parental sequence p , while parameter aa_{irh} stores the amino acid type of rotamer r at position i in hybrid h . In addition, binary variable (Y_{iah}) is introduced and set to be equal to one if amino acid a is selected at position i in hybrid sequence h . Unlike amino acid type changes which are propagated throughout the entire library, rotamer choices can differ between hybrid and/or parental sequences. These new complexities give rise to the following new sets, parameters, and variables definition.

Sets:

$p \in \{1, 2, \dots, P\}$ = set of parental sequences

$h \in \{1, 2, \dots, H\}$ = set of hybrids

$i' \in \{1, 2, \dots, N_p\}$ = set of positions in parental sequence p

$k, k' \in \{1, 2, \dots, N_h\}$ = set of starting and ending positions for perturbation in hybrid h ; $k < k'$

$i, j \in \{k-5, k-4, \dots, k, \dots, k', \dots, k' + 4, k' + 5\}$ = set of positions for perturbation in hybrid h

$a \in \{1, 2, \dots, 19\}$ = set of amino acids excluding proline

$r, s \in \{1, 2, \dots, R\}$ = set of rotamers

R_{ih} = set of rotamers available at position i in hybrid h

Binary Variables:

$$X_{irh} = \begin{cases} 1, & \text{if rotamer } r \text{ is selected at position } i \text{ in hybrid } h \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{iah} = \begin{cases} 1, & \text{if amino acid } a \text{ is selected at position } i \text{ in hybrid } h \\ 0, & \text{otherwise} \end{cases}$$

Continuous Variables:

$$Z_{irjsh} = \begin{cases} 1, & \text{if rotamers } r, s \text{ are selected at positions } i, j \text{ in hybrid } h \\ 0, & \text{otherwise} \end{cases}$$

Parameters:

E_h^{sb} = substrate - backbone energy of hybrid h

E_{irh}^{rb} = rotamer - backbone energy of rotamer r at position i in hybrid h

E_{irh}^{rs} = rotamer - substrate energy of rotamer r at position i in hybrid h

E_{irjsh}^{rr} = rotamer - rotamer energy of rotamers r, s at positions i, j in hybrid h

aa_{irh} = amino acid type of rotamer r at position i in hybrid h

$v_{i'ap} = \begin{cases} 1, & \text{if amino acid } a \text{ occurs at position } i' \text{ in parental sequence } p \\ 0, & \text{otherwise} \end{cases}$

By building on the SSDP formulation using the new additional sets, variables, and parameters, the problem of parental sequence redesign and associated hybrid library design problem (HLDP) is modeled as the following MILP formulation:

$$\text{Minimize } 1/H \sum_h \sum_i \sum_r X_{irh} \cdot (E_{irh}^{rs}) \quad (11)$$

$$\sum_h \left\{ \sum_i \sum_r X_{irh} \cdot (E_{irh}^{rb} + E_{irh}^{rs}) + \sum_i \sum_{j>i} \sum_r \sum_s Z_{irjsh} \cdot E_{irjsh}^{rr} + E_h^{sb} \right\} \leq H \cdot E_{cutoff} \quad (12)$$

$$\sum_r X_{irh} = 1, \quad \forall i, h; r \in R_{ih} \quad (13)$$

$$X_{irh} = 0 \quad \forall i, r, h \text{ such that } E_{irh}^{rs} > \delta_{ih}; r \in R_{ih} \quad (14)$$

$$Z_{irjsh} = X_{irh} \cdot X_{jsh} \quad \forall i, r, j, s, h; r \in R_{ih}, s \in R_{jh} \quad (15)$$

$$\sum_a Y_{iah} = 1, \quad \forall i, h; r \in R_{ih} \quad (16)$$

$$Y_{iah} = \sum_r X_{irh} \quad \forall (i, a, h) \text{ such that } aa_{irh} = a; r \in R_{ih} \quad (17)$$

$$Y_{iah} = v_{i'ap} \quad \forall i, h, k, p \text{ such that position } i \text{ corresponds to position } i' \text{ in the parental sequence } p \quad (18)$$

Slightly modified versions of Constraints (11)-(15) were also present in the SSDP formulation. Briefly, Constraint (11) is the objective function of HLDP involving the minimization of the average surrogate score (e.g., binding energy) of the hybrids in the library. Constraint 12 ensures the stability of the hybrid sequences in the library by imposing an energy cutoff. Constraints (13) and (14) ensure selection of only one rotamer

r at any given position i in any hybrid sequence h while eliminating any rotamers with a high enough energy to preclude them from the optimal solution. Equation (15) is identical to equation (7) in SSDP. Constraint (16) ensures that only one amino acid type a is permitted at any given position i in a hybrid h . Constraint (17) determines the amino acid type (Y_{iah}) of the rotamer selected at position i in a hybrid h . Finally, equation (18) ensures that amino acid type a at position i in the hybrid sequence h is the same as the amino acid type at position i' in parental sequence p . This is in accordance with position i of hybrid h being retained from position i' of parental sequence p . Equation (15), as in the case of Equation (7), involves the product of two binary variables. It is exactly recast into a linear form in the same manner as shown below.

$$\sum_s Z_{irjsh} = \sum_s [X_{irh} \cdot X_{jsh}] = X_{irh} \sum_s [X_{jsh}] = X_{irh} \quad \forall i, r, j > i, h; \quad r \in R_{ih}, \quad s \in R_{jh} \quad (19)$$

$$\sum_r Z_{irjsh} = \sum_r [X_{irh} \cdot X_{jsh}] = X_{jsh} \sum_r [X_{irh}] = X_{jsh} \quad \forall i, j > i, s, h; \quad r \in R_{ih}, \quad s \in R_{jh} \quad (20)$$

$$0 \leq Z_{irjsh} \leq 1 \quad \forall i, r, j > i, s, h; \quad r \in R_{ih}, \quad s \in R_{jh} \quad (21)$$

Formulation HLDP is composed of constraints (11)-(21) excluding constraint (15). We use the CPLEX MILP solver accessed through the GAMS modeling environment to solve both SSPD and HLPD. This optimization step is integrated with CHARMM using a FORTRAN 90 interface.

(iv) *Backbone Relaxation*: The optimization step described above may lead to a number of new residues and/or rotamers for the hybrid structures. These new side-chains and/or conformations may no longer be optimally interacting with the previous backbone. To remedy this, a backbone relaxation step is included here allowing for dihedral angles to vary while the bond lengths and angles are constrained to their original values using

equations (1) and (2). Note that each hybrid structure undergoes a separate backbone relaxation procedure to optimize the backbone conformation with respect to its associated rotamers. Here the side-chain conformations are fixed while the backbone torsion angles are optimized over the same 10-15 residue window using the Adopted Basis-set Newton-Raphson (ABNR) algorithm within CHARMM and the same energy function used for sequence design [61]. A maximum of 4,000 steps are allotted for backbone relaxation though energy minimization.

(v) *Ligand Re-docking*: Because of the alterations in the backbone and the change of rotamers/residue type, the location of the ligands may need to be adjusted with respect to the new structure. Therefore, the ligands are re-docked separately for each of the hybrid and parental sequences using the ZDOCK docking software [241, 242]. This re-docking step is performed only after a number of prespecified design cycles to cut down on computational requirements. Tight bounds are introduced into ZDOCK to constrain ligand placement in only the relevant pocket or active site. The ligand re-docking step using the ZDOCK software is integrated with the backbone relaxation and side-chain optimization steps using a FORTRAN interface.

(vi) *Accepting/Rejecting Moves*: Following the re-docking step, the average score of the hybrid library is calculated and the perturbation imparted in Step (i) is accepted or rejected on the basis of the difference between the final and starting average scores according to the Metropolis criterion. The procedure is repeated for 200-10,000 iterations depending on the complexity and size of the design study.

Upon completion, IPRO provides a set of low energy solutions and associated mutations to be performed within the parental sequences whose propagation to the hybrid

library improves the average score of the library. Due to the decomposable structure of the parental sequence redesign problem, most of the computation can be done in parallel with little information cross-flow. Specifically, hybrid structure refinement, backbone relaxation, backbone perturbation, calculation of rotamer-backbone and rotamer-rotamer energies, and ligand docking for each hybrid are performed on separate processors. Following the rotamer-backbone and rotamer-rotamer energy calculations for each hybrid, the information is fed as parameters to the “master” processor which subsequently solves the MILP model (i.e., SSPD or HLDP) to determine the optimal residues at each of the design positions in the parental sequence(s). The choice of the residues/rotamers determined using the MILP for each of the hybrids is then passed to the “slave” processors for further backbone relaxation and ligand docking. All computational studies listed in this work were performed on a Linux PC cluster using a 3.06GHz Xeon CPU/4GB RAM.

Section 6.3: Hybrid Construction and Functional Screening

Section 6.3.1: Construction of DHFR hybrid libraries

Previously constructed plasmids pAZE-BE and pAZE-EB [232] were used in this work to construct plasmids for the generation of the *L. casei* – *B. subtilis* DHFR libraries in both orientations (pAZE-LB and pAZE-BL). First, the *E. coli* DHFR fragments containing residues 1-120 and 31-159 were removed from pAZE-EB and pAZE-BE plasmids by *NdeI/BamHI* and *PstI/SpeI* restriction digests, respectively. The *L. casei* DHFR fragments 1-124 and 30-162 were obtained by *NdeI/BamHI* and *PstI/SpeI* restriction digests of pAZE-EL and pAZE-LE plasmids (gift from Alex R. Horswill). The *L. casei* DHFR fragment 1-124 was then inserted into the cut pAZE-EB by ligation, taking advantage of the complementary *NdeI* and *BamHI* sites. Analogously, the *L. casei*

DHFR fragment containing residues 30-162 was inserted into the cut pAZE-BE by ligation. Plasmids pAZE-LB (*L. casei* residues 1-124 – *B. subtilis* residues 31-159) and pAZE-BL (*B. subtilis* residues 1-121 – *L. casei* residues 30-162) were confirmed by sequencing at the Nucleic Acids Facility of Pennsylvania State University.

To construct the hybrid libraries, plasmids pAZE-LB and pAZE-BL were linearized at a unique *Sall* site between the *L. casei* and *B. subtilis* DHFR fragments. Incremental truncation for the creation of hybrid enzymes (ITCHY) method was used to construct libraries of hybrid *L. casei* - *B. subtilis* DHFRs in both orientations [245]. Libraries were transformed and stored in *E. coli* strain DH5 α .

Section 6.3.2: Selection and determination of specific activities of active DHFR hybrids

The plasmids containing the hybrid DHFR genes were purified and electroporated into modified *E. coli* strain MH829, which has a deletion of DHFR (*folA*) gene. Transformed cells were washed twice in minimal media A (MMA) and plated on MMA agar plates supplemented with 0.5 % glycerol, 0.6 mM arginine, 50 μ g/mL thymidine, 25 μ g/mL kanamycin, 100 μ g/mL ampicillin, 1 mM MgSO₄, and 100 μ M isopropyl β -D-thiogalactose. The plates were allowed to grow for 5 days at room temperature and colonies were picked and restreaked onto the same media and grown at 30 °C for 24 hours. The selectants were sequenced at the Nucleic Acids Facility of Pennsylvania State University to identify crossover positions and confirm the absence of insertions, deletions, or mutations.

The specific activities of hybrid DHFRs were determined in cell-free lysates as previously described [232]. Briefly, the plasmid pAZE was used to express all DHFR hybrids. To increase expression levels, *lacI* gene was destroyed on all plasmids by *EcoRV*

and *SfoI* restriction digests. Plasmids were transformed into the strain MH829 and 50 mL cultures were grown at 30 °C in LB broth with 100 µg/mL ampicillin, 50 µg/mL thymidine, and 0.5 mM isopropyl β-D-thiogalactoside. Cultures were grown to OD₆₀₀ of 1.0, centrifuged, washed with 25 ml of buffer (20 mM Tris, pH 7.7, 2 mM DTT) and resuspended in 1 mL of buffer. The cells were broken by sonication and insoluble material was removed by centrifugation. The lysates were assayed at 25 °C in MTAN buffer at pH 7.0 using the Cary 100 Bio UV-Vis spectrophotometer by Varian. Cell-free lysate was preincubated with 100 µM cofactor NADPH and the reaction was initiated by adding substrate dihydrofolate to 100 µM. Reaction progress was monitored by following absorbance at 340 nm (NADPH absorbance maximum) ($\Delta\epsilon = 13,200 \text{ } \mu\text{M}^{-1}\text{cm}^{-1}$).

Section 6.4: Application Example:

Section 6.4.1: DHFR Library Characterization and Analysis

The construction, identification, and characterization of the above discussed sixteen *E. coli*/*B. subtilis* DHFR hybrids were described previously [232]. *E. coli* and *B. subtilis* DHFRs share a 28 % sequence identity at the protein level. Below is discussed the isolation and characterization of ten *B. subtilis*/*L. casei* DHFR hybrids used here to validate the computationally derived overall binding scores. The *B. subtilis*/*L. casei* DHFR hybrid library was constructed from the *B. subtilis*/*L. casei* DHFR pair sharing a 36 % sequence identity at the protein level. A previously developed [232] genetic selection utilizing an *E. coli* strain containing a complete deletion of chromosomal DHFR (*folA*) was used to select hybrid enzymes with DHFR activity from the library. For this reason, it was necessary to use inactive DHFR fragments to make the ITCHY libraries, which limited the crossover window to residues 31-121. The combined library put

through the selection included $\sim 2.1 \times 10^6$ members. There are $(90 \times 3)^2$ or 72,900 possible hybrid proteins. To determine the number of library members that must be examined for complete library coverage, the number of hypothetical members is typically multiplied by ten. Since we examined more than 729,000 members, complete library coverage can be assumed. From the DHFR enzymes that passed the selection, 40 hybrids were randomly chosen and sequenced. Only two contained insertions, the remaining 38 were free of insertions, deletions, and mutations. Ten out of 38 hybrids were chosen for this study based on their even distribution of crossover positions over the 90 amino acid crossover position window (see Table 6.1). The crossover position in the *B. subtilis/L. casei* hybrids is defined as the last residue (by alignment position) of *B. subtilis* DHFR. It is clear from the number of active DHFR hybrids identified that 36 % sequence identity on the amino acid level between two DHFR proteins can be sufficient for the generation of active hybrids.

Specific activities ($\mu\text{mol}/\text{min}/\text{mg}$) of the *B. subtilis/L. casei* hybrid enzymes were measured in order to compare these values to the overall binding scores obtained using the SSDP formulation. The *B. subtilis/L. casei* hybrids with the highest activities were found to have crossover positions close to the N- or C- terminus. These hybrid proteins consist mostly of one DHFR (i.e. *B. subtilis* or *L. casei*) and have only a short amino acid sequence replaced by the sequence of the other DHFR at either the N- or the C- terminus. Consequently, these hybrids have a relatively small number of new interactions since a large percentage of the sequence is retained from one species. The hybrids with the lowest activities have their crossover positions in the central region of the crossover position window, between amino acids 53 and 103. This region belongs to the adenosine

binding subdomain of DHFR, which is involved in binding of the cofactor NADPH [197]. These hybrids contain long sequence fragments from both *B. subtilis* – and *L. casei* DHFRs and are thus expected to have many new interactions not present in the wild type proteins. Similar results were seen for the *E. coli/B. subtilis* DHFR hybrids; the lowest specific activities were found for the hybrids with crossover positions in the central region consisting of amino acids 55-96.

Section 6.4.2: IPRO Analysis of DHFR Libraries

In this section we provide a step-by-step application of the IPRO procedure, starting with the SSDP formulation, to test whether it is feasible to improve the computationally derived overall binding scores of two separate DHFR hybrid systems: (i) sixteen *E. coli/B. subtilis* and (ii) ten *B. subtilis/L. casei* hybrid DHFR sequences. These results are contrasted against the experimentally determined specific activity values to check whether the trends observed for the specific activity can be explained using the computed binding scores. First we apply the SSDP formulation to individually design each one of the sixteen *E. coli/B. subtilis* DHFR hybrids considering two different sets of design positions followed by the HLDP formulation which is used to optimize the average binding energy of the sixteen *E. coli/B. subtilis* DHFR hybrids.

Starting with Step (a), IPRO first generates the sequences for the sixteen *E. coli/B. subtilis* and (ii) ten *B. subtilis/L. casei* DHFR hybrids corresponding to the crossover positions shown in Table I. This simply involves splicing of the parental sequence fragments consistent with the given crossover positions. Putative structures for two different sets of DHFR hybrids are generated as described in Step (b). The alignment of the parental structures required for this step is performed using the Combinatorial Extension

(CE) method [246]. An approximate structure of each of the hybrid sequences is constructed by concatenating the corresponding parental structure fragments obtained from the aligned structures. The structures of the *E. coli* (PDB code: 1RX2) and *L. casei* (PDB code: 1AO8) parental sequences were obtained from the Protein Data Bank [100] while the structure of the *B. subtilis* DHFR was provided to us by Dr. Petsko at Brandeis University (*Biochemistry*, 2005, *in preparation*). Each one of these putative structures was refined by allowing the backbone around the junction point (14 residue window) to relax through energy minimization and subsequently the hydrogen atoms were added as described in Step (b). While no residue changes are made, SSDP is used to drive side-chains movements (rotamer changes and/or backbone relaxation) for best binding. The optimized binding scores (kcal/mol) for these hybrid sequences were then contrasted against the experimentally measured specific activities ($\mu\text{mol}/\text{min}/\text{mg}$). The specific activity values of the *B. subtilis/L. casei* and *E. coli/B. subtilis* hybrids [232] are shown in (Table I). The calculated binding scores in each case is found to be linearly correlated to the natural log of the specific activities suggesting that binding energy is a good predictor of specific activity (see Figures 6.6a and 6.6b corresponding to *E. coli/B. subtilis* and *B. subtilis/L. casei* DHFR hybrid sequences respectively). Specifically, 72.7% of the variance in the specific activity trend for the *E. coli/B. subtilis* DHFR hybrids and 75.4% for the *B. subtilis/L. casei* DHFR hybrids is explained by the log-linear relation with the binding scores.

The next step involves the redesign of each one of the sixteen *E. coli/B. subtilis* DHFR hybrid sequences individually using SSDP formulation to enhance their computationally derived binding energies. Two separate sets of design positions were

considered, as required in Step (c), for mutation: (i) positions that were identified to be involved in clashes [231, 232] and (ii) all residues within the binding pocket (i.e., within 4 Å distance from the substrate) that are likely to contribute directly to the binding score. Clashing positions for each one of the hybrid structures was determined using the Clashmap [231] and FamClash [232] procedures. Positions that were frequently involved in clashes were identified and considered for redesign. The same design positions were considered for all the hybrid sequences in order to identify any significant patterns in the residue substitutions. On average, 20 design positions were considered in either case and each run was submitted to an individual processor for a total of 1,000 iterations for binding score minimization using SSDP. Interestingly, out of twenty positions considered for redesign, we found that only seven positions (results shown in Table 6.2) are mutated away from the wild-type. The maximum number of mutations introduced in any one hybrid sequence did not exceed four mutations (see Table 6.2). Notably, a number of mutations are prevalent in all designs. Also many residues that are within or close to the binding pocket persist at the wild-type type even though they are treated as design candidates.

Redesigning the clashing positions (a total of 17 positions) provides approximately the same improvement (-6.9 Kcal/Mol) in the average binding score as compared to designing only the binding pocket residues (-6.2 Kcal/Mol) including 22 residues. This means that at least in this study relieving clashes can indirectly improve binding at the same extent as active site residue redesign. The binding scores of the hybrid sequences before and after design for the two set of design positions are compared in Figures 6.7a and 6.7b, respectively. Notably, when only clashing residue positions are considered for redesign, most of the improvement in the binding scores of the hybrid sequences (average score = -149.0

Kcal/Mol) is found to be the result of a single mutation in the *B. subtilis* DHFR sequence fragment (S64R) and two mutations in the *E. coli* sequence fragment (S64R and T68F). On the other hand when only binding pocket residues are considered for redesign, a single mutation in the *E. coli* (W30F) and a single mutation in the *B. subtilis* (Y30F) DHFR sequence fragments appear to contribute most to the improvement in the binding score (average score= -148.3 Kcal/Mol). Not surprisingly, these mutations are found to be consistently occurring in the design of most of the hybrid sequences. Many alternate mutations leading to the same binding score improvement are found particularly for design positions 65, 67, and 68 (see Table 6.2b).

The results highlighted above describe the application of the SSDP optimization formulation which enables the one-by-one optimization of each one of the 14 hybrids. Note that mutations predicted for the same position can vary for different hybrids. Next, we describe the application of HLDP which unlike the SSDP formulation enforces the same set of mutations for all hybrids. The objective here is to contrast the overall results obtained from the two optimization formulations. Both the clashing positions and residues within the binding pocket are considered simultaneously. The HLDP formulation was run on a 16 node Linux PC cluster with 3.06GHz Xeon CPU/4GB RAM, with one node assigned to each sequence (14 hybrid sequences and two parental sequences). One of these nodes served as the “master” node that solved the HLDP framework every iteration. This procedure was run for a total of 48 hours that permitted on average 315 design iterations. The energy profile of the library before and after the redesign of the parental sequences is shown in Figure 6.8. Note that even though we obtained an improvement in the binding scores (see Table 3) for all hybrid sequences, this may not always be the case as the

improvement in the average binding score of the library may be in some cases due to a handful of hybrid sequences. We find that the most prevalent mutations based on the SSDP results are again present. HLDP identified mutations at only three positions in the parental sequences (positions 30, 64, and 68) that yielded an average binding score of -149.0 Kcal/Mol. Notably, this is very close to the average binding score of the library where each sequence is individually redesigned. While the upstream parental redesign using HLDP requires in total only five mutations in the parental sequences, the downstream hybrid sequence design involve up to four different mutations for each hybrid sequence. This example, therefore, demonstrates that upstream parental sequence redesign can indeed optimize all resulting hybrids in one step in contrast to one-by-one redesign of the hybrid sequences.

Examination of the resulting structures of the redesigned sequences reveals that most of the improvement in the average binding score of the library results from a new salt bridge between the substituted arginine at position 64 and the cofactor NADPH (Figure 6.9a). Moreover, substitution of tyrosine and tryptophan at position 30 with a smaller aromatic residue phenylalanine perhaps reduces steric hindrance with the substrate DHF (Figure 6.9b). We also find that the designs identified using the IPRO procedure are consistent with the residue types observed in the DHFR protein family sequences (at position 30: F=15.73% and at position 64: R=57.98 %). It is important to note that no information of the protein family sequences was *a priori* provided to the IPRO model.

Section 6.5: Summary and Discussion

In this work we introduced the computational framework IPRO for the computational design of protein combinatorial libraries. IPRO identifies targeted

mutations in the parental sequences that when propagated in the combinatorial library systematically optimizes a computationally accessible quantitative metric of library quality (e.g., stability, binding affinity, specific activity, etc.). A new design paradigm is thus proposed that improves the entire library in one step instead of “rescuing” individual hybrids one at a time. IPRO allows for ligand re-docking and backbone movement while a globally convergent mixed-integer linear program (MILP) formulation drives side-chain selection. Two separate MILP formulations (SSDP and HLDP) are included in the IPRO procedure that allow for both the downstream redesign of promising hybrids and the upstream redesign of parental sequences, respectively. Sixteen different *E. coli/B. subtilis* DHFR hybrids were computationally redesigned individually, (i.e., one-by-one using the SSDP formulation) and as well as in a single step through parental sequence redesign (i.e., HLDP formulation). We found similar improvements in the binding energy for both cases demonstrating the feasibility of redesigning combinatorial libraries in a single step.

The current implementation of IPRO can only handle design objectives exemplified by a single energy-based surrogate function, (e.g., binding score as a measure of specific activity). However, in many cases, library quality depends on multiple, and sometimes competing, requirements. For example, altering ligand (or substrate) specificity requires redesigning the binding pocket to recognize the new ligand but also eliminate any affinity for the old one(s). Future work include extending IPRO using a two-stage optimization procedure where the outer problem drives residue mutations by minimizing the binding energy with respect to the new ligand while the inner problem ensures that the new design does not bind the old ligand(s) for any rotamer combination. While modifying an existing active site to accommodate new interacting partners can be

achieved by targeted point mutations as described before, introducing a completely new functionality in an existing protein scaffold requires a new computational design paradigm. IPRO procedure can be further extended to allow for the “grafting” of binding sites from one protein to another. Again this leads to a nested optimization structure where the outer problem performs active site geometry optimization while the inner problem tests/prevent distortion of the grafted binding site upon energy minimization.

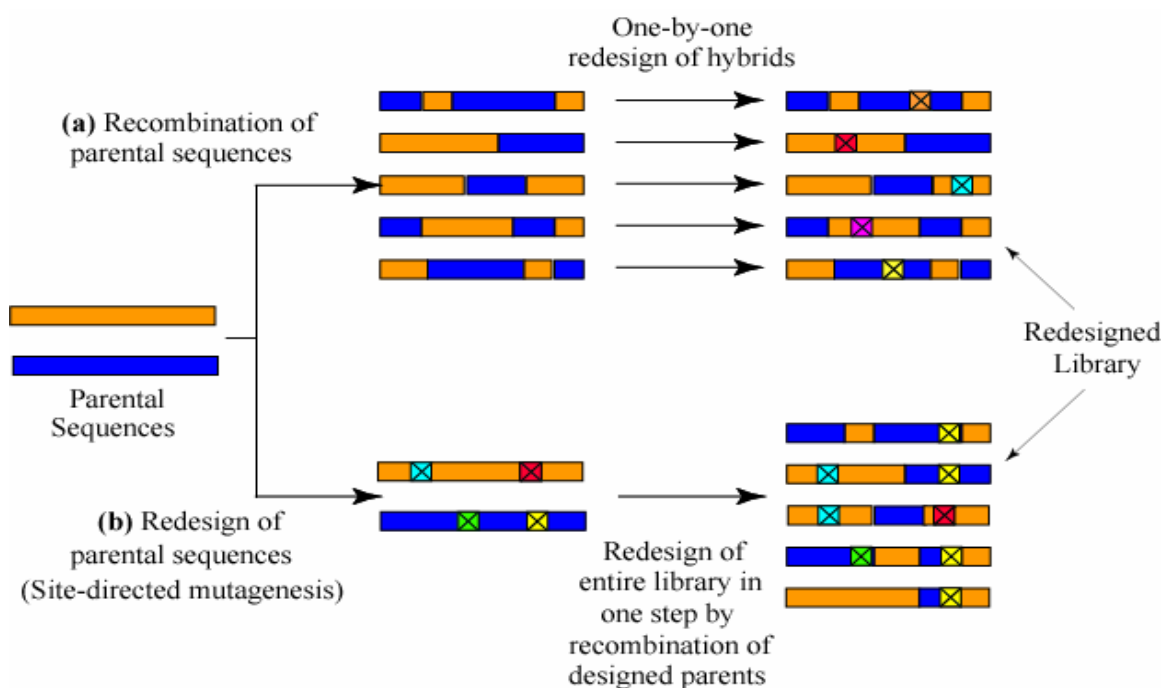


Figure 6.1: (a) Promising hybrid sequences from the library are selected for downstream redesign that involves either random or site-directed mutagenesis. (b) Illustration of the upstream parental sequence redesign. Note that the mutations in the parental sequences propagate downstream into the combinatorial library effectively designing the combinatorial library at once, thereby improving the overall quality of the library.

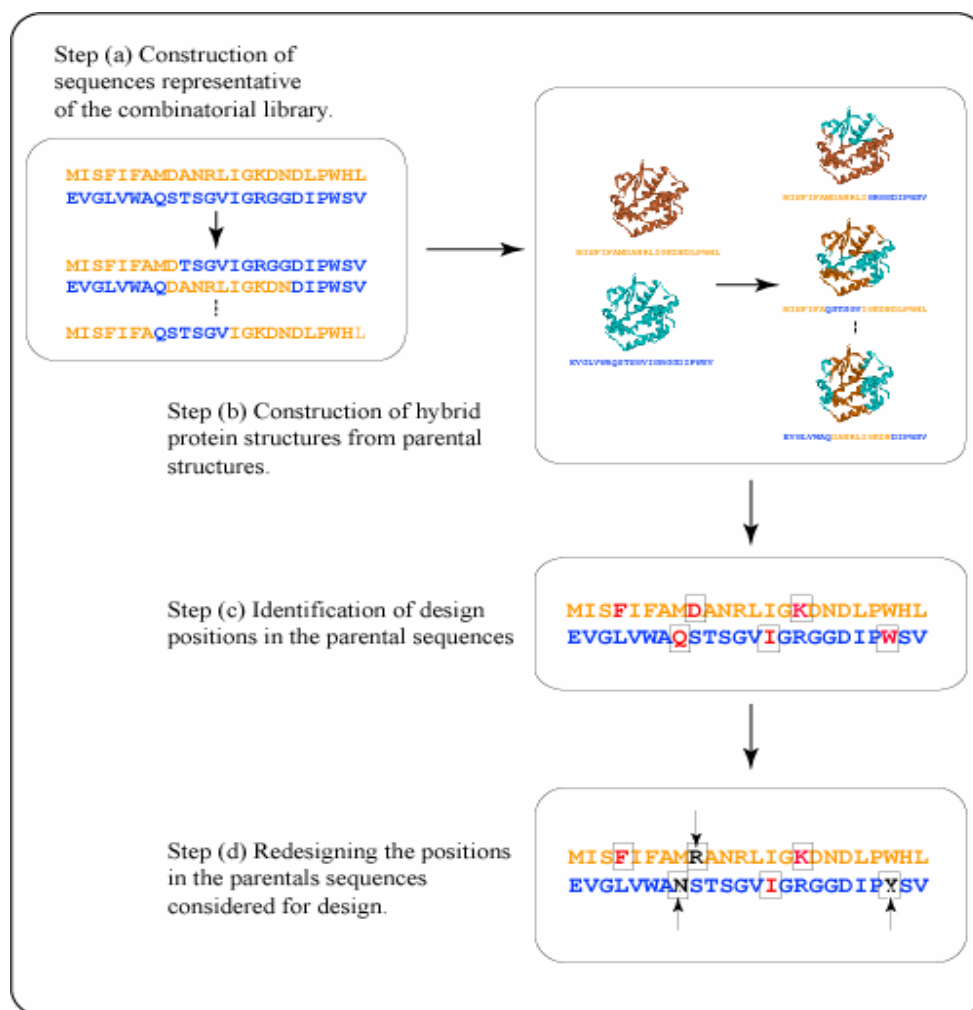


Figure 6.2: The four key steps involved in the IPRO procedure. Details of each of these steps are described separately in the text.

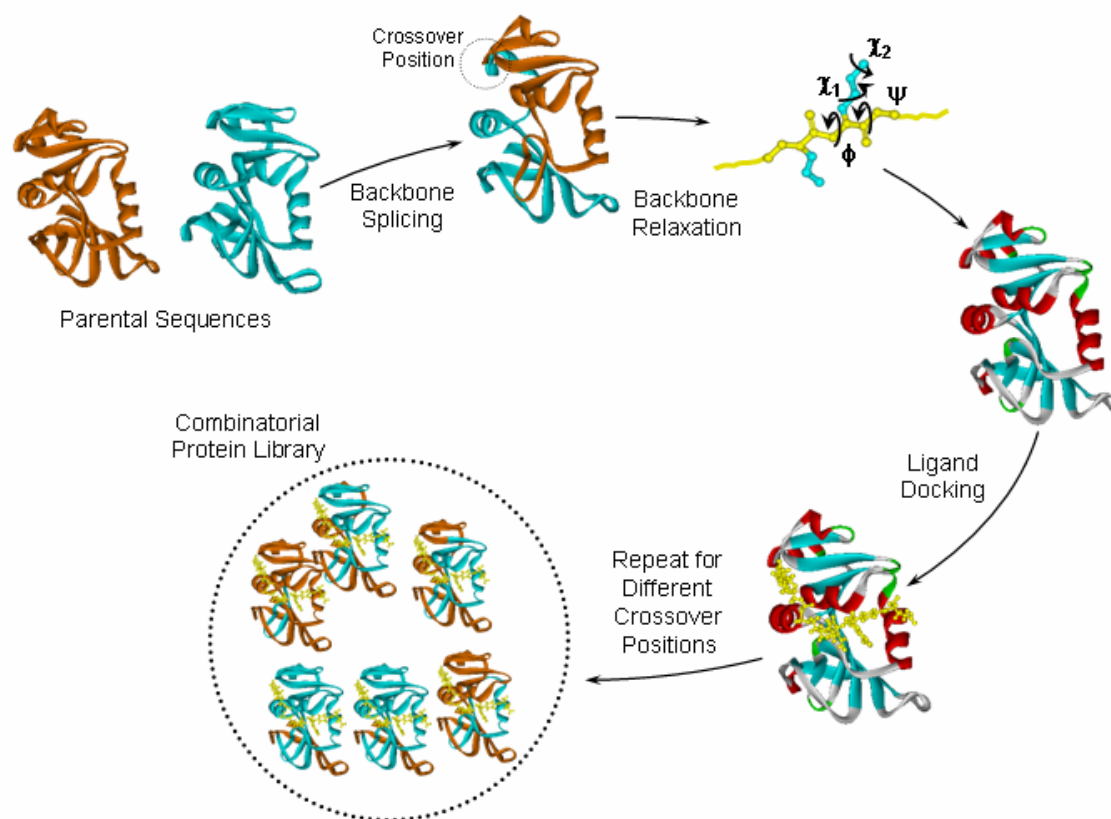


Figure 6.3: This figure highlights the key steps for constructing the initial structure of a hybrid protein from a set of parental structures with known crossover position(s). These involve (i) backbone splicing, (ii) backbone relaxation at the crossover positions, and (iii) ligand re-docking. These steps are repeated for different crossover positions to generate the combinatorial library.

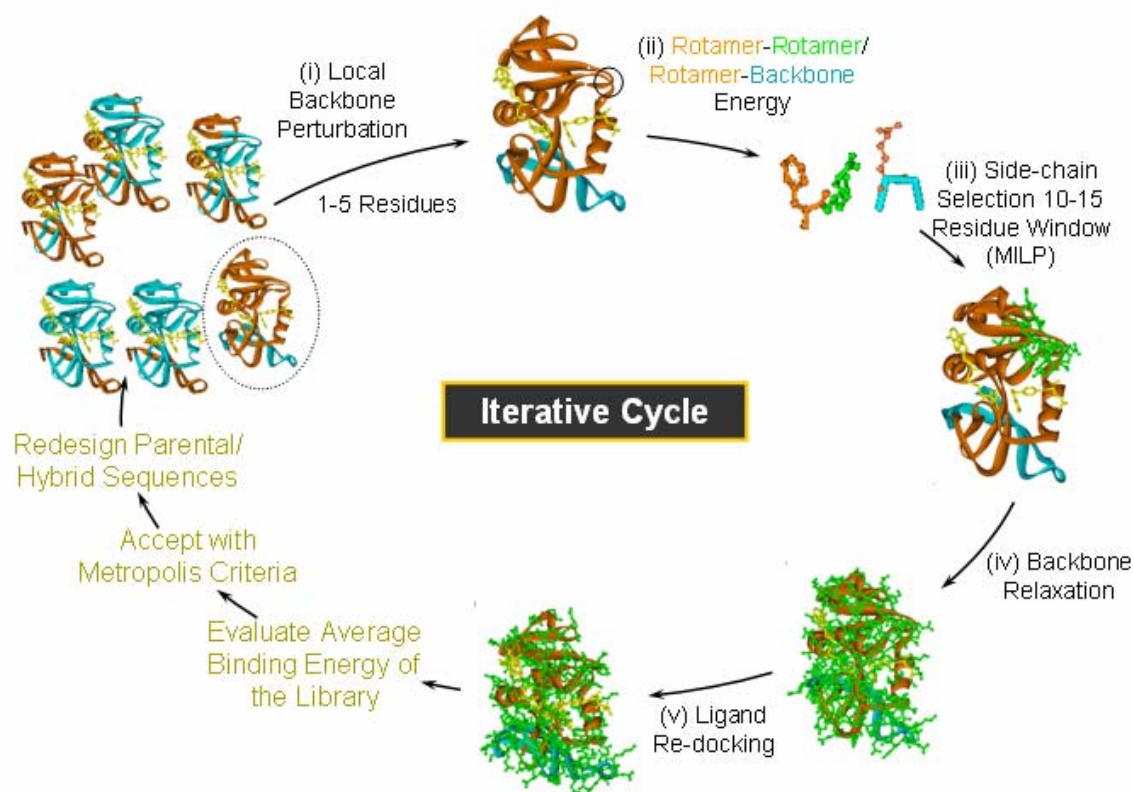


Figure 6.4: IPRO is an iterative protein redesign software that includes the following steps: **(i)** A local region of the protein (1-5 consecutive residues as shown in black circle) is randomly selected for perturbation. The backbone torsion angles of these residues are perturbed by up to ± 5 degrees. **(ii)** All amino acid rotamers consistent with these torsion angles are selected at each position from the Dunbrack and Cohen rotamer library[247]. Rotamer-backbone and rotamer-rotamer energies are calculated for all the selected rotamers using a suitable energy function [33]. **(iii)** A mixed-integer linear programming formulation is used to select the optimal rotamer at each of these positions such that the binding energy is minimized. **(iv)** The backbone of the protein is relaxed through energy minimization in order to allow it to adjust to these new side-chains. **(v)** The ligand

position is re-adjusted with respect to the modified backbone and side-chains using the ZDOCK[242] docking software. **(vi)** The binding energy of the protein-ligand complex is evaluated and the move is accepted or rejected using the Metropolis criterion.

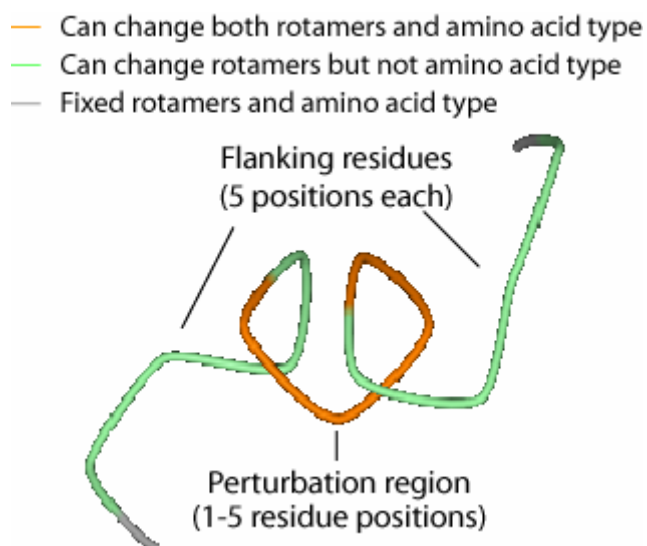


Figure 6.5: The design positions within the perturbation region (shown in orange) are permitted to change amino acid type while the flanking residue positions (5 residues on either side shown in green) can only change rotamers but not the residue type. Position outside this 10-15 residue window (gray) are fixed and cannot change either rotamer or residue type.

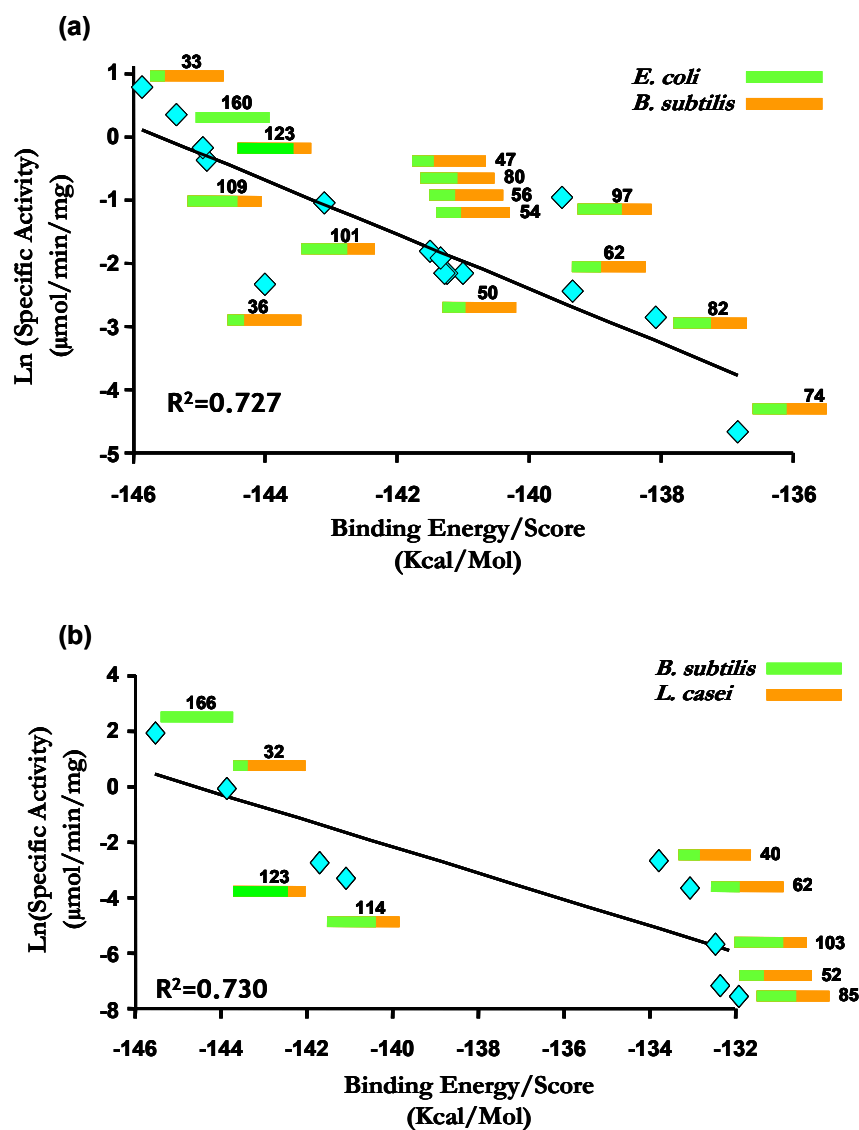


Figure 6.6: Plot of the natural log of the specific activities against the binding scores for two different types of DHFR hybrids (a) *E. coli*/*B. subtilis* and (b) *B. subtilis*/*L. casei*. Along each point is shown the corresponding hybrid sequence with its crossover position.

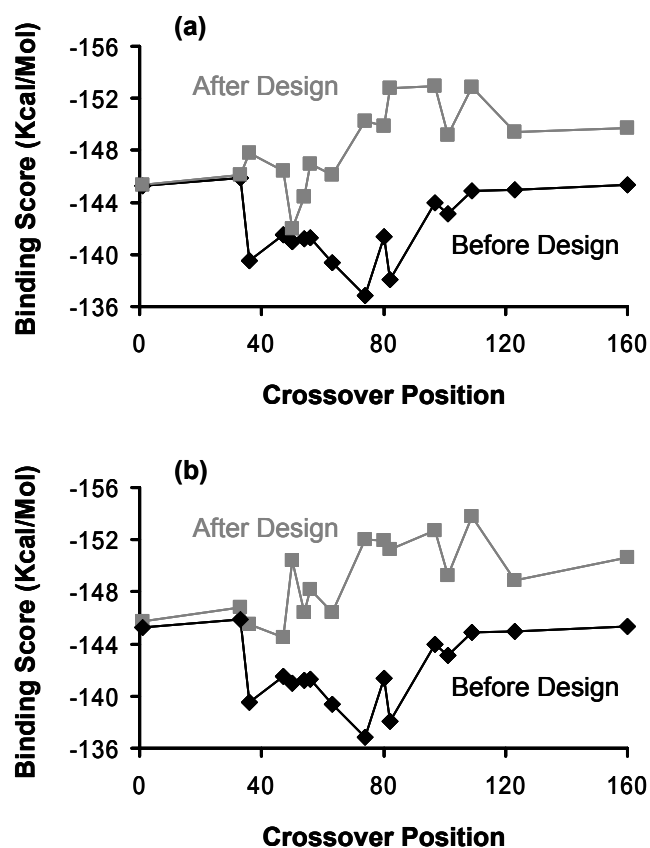


Figure 6.7: Binding score profile before and after redesign of the *E. coli/B. subtilis* DHFR hybrids using the SSDP framework when **(a)** only clashing residue positions are considered and **(b)** only binding pocket residues are considered for redesign.

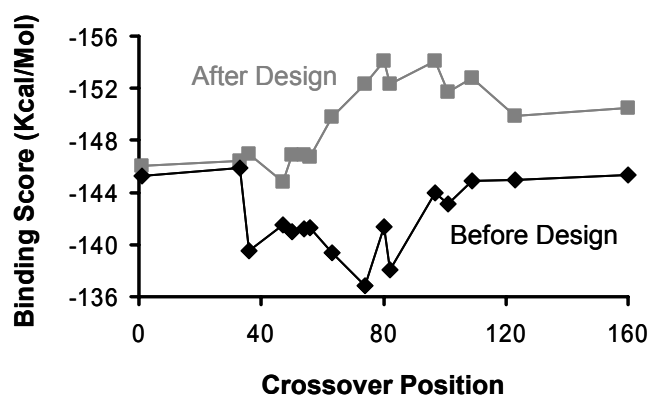


Figure 6.8: Binding score profile before and after redesign of parental *E. coli* and *B. subtilis* DHFR sequences using the HLDP framework. Both clashing residue positions and the binding pocket residues are considered for design.

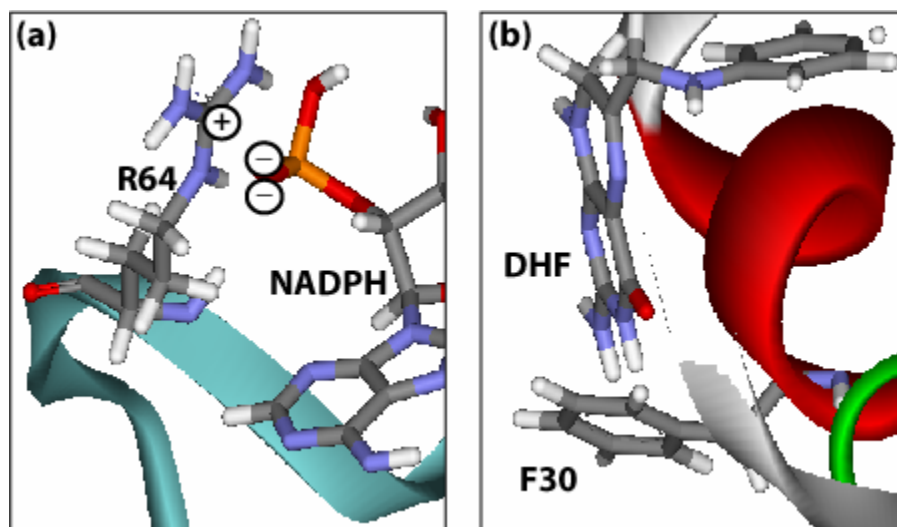


Figure 6.9: (a) Substitution of serine with an arginine at position 64 stabilizes the binding with the cofactor NADPH due to formation of a new salt bridge. (b) Substitution of tyrosine and tryptophan at position 30 with a smaller aromatic residue phenylalanine perhaps reduces steric hindrance with the substrate DHF.

Table 6.1: Crossover positions for the *E.coli/B.subtilis* and *B.subtilis/L. casei* DHFR hybrids and their specific activities (mmol/min/mg).

<i>E.coli/B.subtilis</i>		<i>B.subtilis/L. casei</i>	
Crossover Position	Specific Activity	Crossover Position	Specific Activity
0	20.22	0	0.197 ± 0.114
32	2.17	32	0.915 ± 0.086
35	0.39	40	0.067 ± 0.008
46	0.17	53	0.001 ± 0.000
49	0.12	62	0.025 ± 0.004
53	0.12	85	0.001 ± 0.000
55	0.12	103	0.003 ± 0.001
62	0.09	114	0.035 ± 0.016
73	0.01	123	0.063 ± 0.005
79	0.15	160	6.622 ± 0.157
81	0.06		
96	0.10		
100	0.36		
108	0.70		
122	0.84		
159	1.43		

*The errors in the specific activity for the *B. Subtilis/L. Casei* hybrids are given at 95% confidence interval.

*The crossover positions for the *E. coli/B. subtilis* and *B. subtilis/L. casei* hybrids are defined as the last residue position (in alignment) of the *E. coli* and *B. subtilis* DHFR sequences, respectively

Table 6.2: Individual redesigns of the **(a)** clashing positions and **(b)** binding site residues for the *E.coli/B.subtilis* hybrid DHFR sequences.

(a)	30	62	63	96	97	98	103
B.subt	Y	V	T	G	A	Q	L
E.coli	W	L	S	G	G	R	F
0	F						
33	F				K		
36	F				Q		
47	F						
50	F				K		
54	F						
56	F						
62	F/A						
73	F		T			K	M
79	F						
81	F						
96	F						
101	H					K	
109	H/F					K	
123	F			Q			L
160	F		A			K	L

(b)	57	61	63	64	65	67	68
B.subt	R	V	S	S	A	D	S
E.coli	R	I	T	S	Q	G	T
0		T		R	R/Q	R/D	R/F
33		T		R	Q	R	E
36				R	R/Q	R/D	R/Y
47		T		R	Q	E	Q
50		I		K	Q	K	R
54				R	Q		
56	N			R	K	T	Q
62				R	H	K	D
73	K		A	R	R		Q
79			A	R	H		F
81			A	R	R		F
96	T			R	R/Q		F
101				R	R		F
109	N			R	R		F
123				R	T		Y
160			A	R	R		F

The original *B. subtilis* and *E. coli* residues are shown in red and blue respectively. Positions with consistent mutations are shown in color. Note that position 0 corresponds to the *B. subtilis* parental sequence while 160 corresponds to *E. coli* sequence.

Table 6.3: Redesign of parental *E. coli* and *B. subtilis* DHFR sequences.

	30	64	68
B.subt	Y	S	S
E.coli	W	S	T
0	F	R	
33	F	R	
36	F	R	
47	F	R	
50	F	R	
54	F	R	
56	F	R	
62	F	R	
73	F	R	F
79	F	R	F
81	F	R	F
96	F	R	F
101	F	R	F
109	F	R	F
123	F	R	F
160	F	R	F

Chapter 7: Conclusions

Section 7.1: Summary

This thesis describes the development of computational tools for the design and optimization of combinatorial protein libraries. The application of these tools allows for appropriate allocation of diversity in the library while correcting problematic residue combinations. Therefore, experimental resources can now be focused towards the most promising regions of sequence space, thus increasing the chances of identifying novel engineered proteins. Specifically, molecular modeling techniques and bioinformatics-based approaches utilizing sequence and structure information encoded in the parental/family sequences were developed to target two critical problems in protein engineering: (i) *a priori* prescreening protein hybrids for their potential of being stably folded and functional, and (ii) identifying what sequence/residue permutations are the most promising in terms of improving/preserving protein structure and function.

In response to these problems we first developed a bioinformatics-inspired approach Residue Correlation Analysis (RCA), described in Chapter 1, for predicting functionally important domains from protein family sequence data that are less likely to tolerate uncoordinated mutations. Specifically, RCA is comprised of two major steps: (i) identifying pairs of residue positions that mutate in a coordinated manner, and (ii) using these results to identify protein regions that interact with an uncommonly high number of other residues. We hypothesized that strongly correlated pairs result not only from contacting pairs, but also from residues that participate in conformational changes involved during catalysis or important interactions necessary for retaining functionality. The results show that highly mobile loops that assist in ligand association/dissociation tend to exhibit high correlation. RCA results exhibit good agreement with the findings of

experimental and molecular dynamics studies for the three protein families that are analyzed: (i) DHFR (dihydrofolate reductase), (ii) cyclophilin, and (iii) formyl-transferase. Specifically, the specificity (percentage of correct predictions) in all three cases is substantially higher than those obtained by entropic measures or contacting residue pairs.

While RCA in Chapter 1 identifies regions less tolerable to mutations and crossovers, the next logical step in the research was to identify what types of residue pairs in these hybrids result into clashes. To this end we introduced a protein sequence data-based approach to characterize all possible residue pairs present in protein hybrids for inconsistency with protein family structural features. This approach is based on examining contacting residue pairs with different parental origins for different types of potentially unfavorable interactions (i.e., electrostatic repulsion, steric hindrance, cavity formation and hydrogen bond disruption). The identified clashing residue pairs between members of a protein family were then contrasted against functionally characterized hybrid libraries. Comparisons for five different protein recombination studies available in the literature revealed that the pattern of identified clashing residue pairs were surprisingly consistent with experimentally found patterns of functional crossover profiles. Specifically, we showed that the proposed residue clash maps are on average 5.0 times more effective than randomly generated clashes and 1.6 times more effective than residue contact maps or SCHEMA at explaining the observed crossover distributions among functional members of hybrid libraries. Therefore, this research indicates that residue clash maps can provide quantitative guidelines for the placement of crossovers in the design of protein recombination experiments. The shortcoming, however, of this

approach is that it can only classify hybrids as functional or nonfunctional with no information on the activity levels of the hybrids.

Clearly, the next step was to move beyond the simple classification as functional or nonfunctional and to correlate the rank ordering of these hybrids based on their activity levels. For this purpose, we developed a computational procedure FamClash for analyzing incompatibilities in engineered protein hybrids by using protein family sequence data. In this approach all pairs of residue positions in the sequence alignment that conserve the property triplet of charge, volume, and hydrophobicity are first identified and significant deviations are denoted as residue–residue clashes. The hybrids were then rank-ordered based on their activity levels depending on the number of clashes they contained. Experimental testing of this approach was performed in parallel to assess the predictive ability of FamClash. As a model system, single-crossover ITCHY (incremental truncation for the creation of hybrid enzymes) libraries were prepared from the *Escherichia coli* and *Bacillus subtilis* dihydrofolate reductases, and the activities of functional hybrids were determined. Comparisons of the predicted clash map as a function of crossover position revealed good agreement with activity data, reproducing the observed V shape and matching the location of a local peak in activity.

Having developed approaches to identify incompatible residue pairs, the latter part of the research focused on developing methods for identifying residue redesign candidates for restoring lost functionality. Specifically, two separate tracks of computational procedure (OPTCOMB and IPRO) were developed for designing proteins/protein libraries corresponding to the two experimental paradigms for library

generation: (1) recombination of parental segments and (2) parental/hybrid redesign through point mutations.

OPTCOMB (Optimal Pattern of Tiling for COMBinatorial library design) procedure is directly applicable to oligonucleotide ligation-based protocols such as GeneReassembly, DHR, SISDC, and many more. Given a set of parental sequences and the size ranges of the parental sequence fragments, OPTCOMB determines the optimal junction points (i.e., crossover positions) and the fragment contributing parental sequences at each one of the junction points. By rationally selecting the junction points and the contributing parental sequences, the number of clashes (i.e., unfavorable interactions) in the library is systematically minimized with the aim of improving the overall library quality. Using OPTCOMB, hybrid libraries containing fragments from three different dihydrofolate reductase sequences (*Escherichia coli*, *Bacillus subtilis*, and *Lactobacillus casei*) were computationally designed. Notably, we found that there exists an optimal library size when both the number of clashes between the fragments composing the library and the average number of clashes per hybrid in the library are minimized.

The computational procedure IPRO (Iterative Protein Redesign and Optimization procedure) developed in collaboration with Gregory Moore, unlike OPTCOMB, is geared towards the redesign of an entire combinatorial protein library in one step using energy based scoring functions. IPRO relies on identifying mutations in the parental sequences that when propagated downstream in the combinatorial library improves the average quality of the library (e.g., stability, binding affinity, specific activity, etc.). Residue and rotamer design choices are driven by a globally convergent Mixed-Integer Linear

Programming (MILP) formulation. Unlike many of the available computational approaches, the procedure allows for backbone movement as well as re-docking of the associated ligands after a pre-specified number of design iterations. IPRO can also be used, as a limiting case, for the redesign of a single or handful of individual sequences. The application of IPRO was highlighted through the redesign of a sixteen member library of *E. coli*/*B. subtilis* dihydrofolate reductase hybrids, both individually and through upstream parental sequence redesign, for improving the average binding energy. Computational results demonstrate that it is indeed feasible to improve the overall library quality as exemplified by binding energy scores through targeted mutations in the parental sequences.

Design results from both the models, OPTCOMB and IPRO, revealed that best library designs typically involve multiple mutations or complex tiling patterns of parental segments of unequal size that are hard to infer without relying on computational means. Therefore, by introducing formal systems information technology and molecular modeling approaches to the field, a broad impact in the general area of protein engineering can be made. The computational models introduced in this research thus provide a novel and versatile tool box for aiding the identification of design targets as well as guiding protein library designs.

Section 7.2: Future Perspectives

Currently much of the research in protein engineering is devoted to the development and application of molecular evolution and screening techniques. Concurrent with the progress in the directed evolution, several significant advances in rational sequence/structure-based design, driven by the emergence of automated protein

design programs, have occurred recently. One of the major difficulties in the development of algorithms for structure-based design arises from the immense complexity of all the degrees of freedom associated with a protein structure and the size of the number of all possible sequences. Although reduction in the degrees of freedom are achieved by assuming a fixed backbone and representing amino acid side-chains by a small number of low-energy conformations, or rotamers, the search space is still too large for systematic study. Of course one does not need to design an entire protein from scratch given that proteins in a family have similar folds. However, even to date, there are many protein families for which no structures have been characterized. Even more troubling is the realization that it is important to move away from strict adherence to the inverse folding constraints and incorporate backbone movement. Incorporating this flexibility makes the protein design problem computationally intractable. These insights have highlighted the need for better and faster algorithms and establishment of a proper trade-off between modeling accuracy and computational speed.

In addition to the computational complexity of the problem, there is still a lack of an accurate representation of molecular interactions within proteins. No doubt, significant advances have been made in this direction and reasonable representations of these interactions by self-consistent semi-empirical potential/scoring functions are well established. In spite of these advances, to accurately capture subtle interactions that are important for protein stability, further improvement in the representation of these interactions, specifically solvation and electrostatic interactions, are required. A recent impressive contribution along these lines is the *in silico* design and verification of a novel

fold by Baker's group [61] where the scoring functions are heavily parameterized to predict existing folds found in the Protein Data Bank [248].

Other challenges lie in the rational design of protein function: most computational approaches aim to design proteins so as to maximize their stability. There are ample experimental evidences that show that proteins have not evolved to maximize their stability. Therefore, using stability or binding energy as a surrogate of functionality is not always accurate. Consequently, the systematic structure-based engineering of substrate specificity, catalysis, and control of activity are all unsolved problems. One way around this difficulty is to utilize sequence information gleaned from protein family databases (e.g., Pfam [100]). For example, protein family sequence data, spanning all of nature's known solutions, can be used to constrain the solutions for various protein engineering problems. In one such study, Lockless and Ranganathan [107] have found that statistical sequence database-derived coupling energies correlate with thermodynamic coupling free energies in a small protein domain. Yet, another approach that many of the researchers are taking is combining the computational procedures with experimental testing to gain new insights into the protein functionality.

Although we have only begun to scratch the surface of computational protein engineering, it is already clear that automated design procedures have emerged as a useful approach to drive manipulations of protein structure and function, both for study of fundamental principles of structure and function, and for development of new technologies. The computational design now offers the possibility to address tremendous combinatorial complexity impossible to handle experimentally.

Bibliography

1. Stemmer, W.P.C., *Rapid evolution of a protein in vitro by DNA shuffling*. Nature, 1994. **370**: p. 389-391.
2. Stemmer, W.P.C., *DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution*. Proc. Natl. Acad. Sci. USA, 1994. **91**: p. 10747-10751.
3. Zhao, H., et al., *Molecular evolution by staggered extension process (StEP) in vitro recombination*. Nat. Biotechnol., 1998. **16**(3): p. 258-261.
4. Ostermeier, M., et al., *Combinatorial protein engineering by incremental truncation*. Proc Natl Acad Sci U S A, 1999. **96**(7): p. 3562-7.
5. Lutz, S., et al., *Creating multiple-crossover DNA libraries independent of sequence identity*. Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11248-53.
6. Sieber, V., C.A. Martinez, and F.H. Arnold, *Libraries of hybrid proteins from distantly related sequences*. Nat Biotechnol, 2001. **19**(5): p. 456-60.
7. Rui, L., et al., *Protein Engineering of Epoxide Hydrolase from Agrobacterium radiobacter AD1 for Enhanced Activity and Enantioselective Production of (R)-1-Phenylethane-1,2-Diol*. Applied and Environmental Microbiology, 2005. **71**(7): p. 3995-4003.
8. Griswold, K.E., et al., *Evolution of highly active enzymes by homology-independent recombination*. Proc Natl Acad Sci U S A, 2005. **102**(29): p. 10082-7.
9. Varadarajan, N., et al., *Engineering of protease variants exhibiting high catalytic activity and exquisite substrate selectivity*. Proc Natl Acad Sci U S A, 2005. **102**(19): p. 6855-60.
10. Franco, R., et al., *Porphyrin-substrate binding to murine ferrochelatase: effect on the thermal stability of the enzyme*. Biochem J, 2005. **386**(Pt 3): p. 599-605.
11. Minagawa, H., J. Shimada, and H. Kaneko, *Effect of mutations at Glu160 and Val198 on the thermostability of lactate oxidase*. Eur J Biochem, 2003. **270**(17): p. 3628-33.
12. Harvey, D.M. and C.T. Caskey, *Inducible control of gene expression: prospects for gene therapy*. Curr Opin Chem Biol, 1998. **2**(4): p. 512-8.
13. Fussenegger, M., *The impact of mammalian gene regulation concepts on functional genomic research, metabolic engineering, and advanced gene therapies*. Biotechnol Prog, 2001. **17**(1): p. 1-51.
14. Notley-McRobb, L. and T. Ferenci, *Substrate specificity and signal transduction pathways in the glucose-specific enzyme II (EII(Glc)) component of the Escherichia coli phosphotransferase system*. J Bacteriol, 2000. **182**(16): p. 4437-42.

15. Kobayashi, H., et al., *Programmable cells: interfacing natural and engineered gene networks*. Proc Natl Acad Sci U S A, 2004. **101**(22): p. 8414-9.
16. Yokobayashi, Y., R. Weiss, and F.H. Arnold, *Directed evolution of a genetic circuit*. Proc Natl Acad Sci U S A, 2002. **99**(26): p. 16587-91.
17. Bishop, A., et al., *Unnatural ligands for engineered proteins: new tools for chemical genetics*. Annu Rev Biophys Biomol Struct, 2000. **29**: p. 577-606.
18. Wong, K.F., et al., *Impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate reductase*. Proc Natl Acad Sci U S A, 2005. **102**(19): p. 6807-12.
19. Benkovic, S.J. and S. Hammes-Schiffer, *A perspective on enzyme catalysis*. Science, 2003. **301**(5637): p. 1196-202.
20. Saraf, M.C., G.L. Moore, and C.D. Maranas, *Using multiple sequence correlation analysis to characterize functionally important protein regions*. Protein Eng., 2003. **16**(6): p. 397-406.
21. Zhao, H. and F.H. Arnold, *Optimization of DNA shuffling for high fidelity recombination*. Nucleic Acids Res., 1997. **25**: p. 1307-1308.
22. Ostermeier, M., et al., *Combinatorial protein engineering by incremental truncation*. Proc. Natl. Acad. Sci. USA, 1999. **96**(7): p. 3562-3567.
23. Martin, A., V. Sieber, and F.X. Schmid, *In-vitro selection of highly stabilized protein variants with optimized surface*. J. Mol. Biol., 2001. **309**: p. 717-726.
24. Sakamoto, T., et al., *Laboratory evolution of toluene dioxygenase to accept 4-picoline as a substrate*. Appl. Env. Microbiol., 2001. **67**(9): p. 3882-3887.
25. Dalby, P.A., *Optimising enzyme function by directed evolution*. Curr. Opin. Struct. Biol., 2003. **13**(4): p. 500-505.
26. Bacher, J.M., B.D. Reiss, and A.D. Ellington, *Anticipatory evolution and DNA shuffling*. Genome Biol., 2002. **3**(8): p. REVIEWS1021.
27. Brakmann, S., *Discovery of superior enzymes by directed molecular evolution*. ChemBioChem, 2001. **2**(12): p. 865-871.
28. Petrounia, I.P. and F.H. Arnold, *Designed evolution of enzymatic properties*. Curr. Opin. Biotechnol., 2000. **11**(4): p. 325-330.
29. Schmidt-Dannert, C., *Directed evolution of single proteins, metabolic pathways, and viruses*. Biochemistry, 2001. **40**(44): p. 13125-13136.
30. Dwyer, M.A., L.L. Looger, and H.W. Hellinga, *Computational design of a biologically active enzyme*. Science, 2004. **304**(5679): p. 1967-71.
31. Allert, M., et al., *Computational design of receptors for an organophosphate surrogate of the nerve agent soman*. Proc Natl Acad Sci U S A, 2004. **101**(21): p. 7907-12.
32. Korkegian, A., et al., *Computational thermostabilization of an enzyme*. Science, 2005. **308**(5723): p. 857-60.
33. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy*. Science, 2003. **302**(5649): p. 1364-8.
34. Ostermeier, M., A.E. Nixon, and S.J. Benkovic, *Incremental truncation as a strategy in the engineering of novel biocatalysts*. Bioorg Med Chem, 1999. **7**(10): p. 2139-44.

35. Richardson, T.H., et al., *A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable alpha-amylase*. J Biol Chem, 2002. **277**(29): p. 26501-7.
36. Hiraga, K. and F.H. Arnold, *General method for sequence-independent site-directed chimeragenesis*. J Mol Biol, 2003. **330**(2): p. 287-96.
37. Sieber, V., C.A. Martinez, and F.H. Arnold, *Libraries of hybrid proteins from distantly related sequences*. Nat. Biotechnol., 2001. **19**(5): p. 456-460.
38. Voigt, C.A., et al., *Computational method to reduce the search space for directed protein evolution*. Proc. Natl. Acad. Sci. USA, 2001. **98**(7): p. 3778-3783.
39. Koehl, P. and M. Delarue, *Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy*. J. Mol. Biol., 1994. **239**(2): p. 249-275.
40. Lee, C., *Predicting protein mutant energetics by self-consistent ensemble optimization*. J. Mol. Biol., 1994. **236**(3): p. 918-939.
41. Bogarad, L.D. and M.W. Deem, *A hierarchical approach to protein molecular evolution*. Proc. Natl. Acad. Sci. USA, 1999. **96**(6): p. 2591-2595.
42. Perelson, A.S. and C.A. Macken, *Protein evolution on partially correlated landscapes*. Proc. Natl. Acad. Sci. USA, 1995. **92**(21): p. 9657-9661.
43. Kauffman, S. and S. Levin, *Towards a general theory of adaptive walks on rugged landscapes*. J. Theor. Biol., 1987. **128**(1): p. 11-45.
44. Harbury, P.B., B. Tidor, and P.S. Kim, *Repacking protein cores with backbone freedom: structure prediction for coiled coils*. Proc. Natl. Acad. Sci. USA, 1995. **92**(18): p. 8408-8412.
45. Harbury, P.B., et al., *High-resolution protein design with backbone freedom*. Science, 1998. **282**(5393): p. 1462-1467.
46. Klepeis, J.L., et al., *Integrated computational and experimental approach for lead optimization and design of compstatin variants with improved activity*. J Am Chem Soc, 2003. **125**(28): p. 8422-3.
47. Keating, A.E., et al., *Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils*. Proc. Natl. Acad. Sci. USA, 2001. **98**(26): p. 14825-14830.
48. Larson, S.M., et al., *Thoroughly sampling sequence space: large-scale protein design of structural ensembles*. Protein Sci., 2002. **11**(12): p. 2804-2813.
49. Kraemer-Pecore, C.M., J.T. Lecomte, and J.R. Desjarlais, *A de novo redesign of the WW domain*. Protein Sci., 2003. **12**(10): p. 2194-2205.
50. Dunbrack Jr., R.L., *Rotamer libraries in the 21st century*. Curr. Opin. Struct. Biol., 2002. **12**(4): p. 431-440.
51. MacKerell, A.D., et al., *CHARMM: The energy function and its parameterization with an overview of the program*, in *The Encyclopedia of Computational Chemistry*, R. Schleyer, Editor. 1998, John Wiley & Sons: Chichester. p. 271-277.
52. Mayo, S.L., B.D. Olafson, and W.A. Goddard, *DREIDING - A Generic Force-Field for Molecular Simulations*. J. Phys. Chem., 1990. **94**(26): p. 8897-8909.
53. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids and organic molecules*. J. Am. Chem. Soc., 1995. **117**: p. 5179-5197.

54. Scott, W.R., et al., *The GROMOS biomolecular simulation program package*. J. Phys. Chem. A, 1999. **103**(19): p. 3596-3607.
55. Chiu, T.L. and R.A. Goldstein, *Optimizing potentials for the inverse protein folding problem*. Protein Eng., 1998. **11**(9): p. 749-752.
56. Kuhlman, B. and D. Baker, *Native protein sequences are close to optimal for their structures*. Proc. Natl. Acad. Sci. USA, 2000. **97**(19): p. 10383-10388.
57. Looger, L.L. and H.W. Hellinga, *Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics*. J. Mol. Biol., 2001. **307**(1): p. 429-445.
58. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy functions for protein design*. Curr. Opin. Struct. Biol., 1999. **9**(4): p. 509-513.
59. Dwyer, M.A., L.L. Looger, and H.W. Hellinga, *Computational design of a Zn²⁺ receptor that controls bacterial gene expression*. Proc. Natl. Acad. Sci. USA, 2003. **100**(20): p. 11255-11260.
60. Kortemme, T., et al., *Computational redesign of protein-protein interaction specificity*. Nat Struct Mol Biol, 2004. **11**(4): p. 371-9.
61. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy*. Science, 2003. **302**(5649): p. 1364-1368.
62. Kraemer-Pecore, C.M., A.M. Wollacott, and J.R. Desjarlais, *Computational protein design*. Curr. Opin. Chem. Biol., 2001. **5**(6): p. 690-695.
63. Pokala, N. and T.M. Handel, *Review: protein design--where we were, where we are, where we're going*. J. Struct. Biol., 2001. **134**(2-3): p. 269-281.
64. Sarisky, C.A. and S.L. Mayo, *The beta-beta-alpha fold: explorations in sequence space*. J. Mol. Biol., 2001. **307**(5): p. 1411-1418.
65. Desjarlais, J.R. and T.M. Handel, *De novo design of the hydrophobic cores of proteins*. Protein Sci., 1995. **4**(10): p. 2006-2018.
66. Lazar, G.A., J.R. Desjarlais, and T.M. Handel, *De novo design of the hydrophobic core of ubiquitin*. Protein Sci., 1997. **6**(6): p. 1167-1178.
67. Yang, W., et al., *Structural analysis, identification, and design of calcium-binding sites in proteins*. Proteins, 2002. **47**(3): p. 344-356.
68. Hellinga, H.W. and F.M. Richards, *Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry*. J. Mol. Biol., 1991. **222**(3): p. 763-785.
69. Shimaoka, M., et al., *Computational design of an integrin I domain stabilized in the open high affinity conformation*. Nat. Struct. Biol., 2000. **7**(8): p. 674-678.
70. Looger, L.L., et al., *Computational design of receptor and sensor proteins with novel functions*. Nature, 2003. **423**(6936): p. 185-190.
71. Benson, D.E., A.E. Haddy, and H.W. Hellinga, *Converting a maltose receptor into a nascent binuclear copper oxygenase by computational design*. Biochemistry, 2002. **41**(9): p. 3262-3269.
72. Voigt, C.A., et al., *Protein building blocks preserved by recombination*. Nat. Struct. Biol., 2002. **9**(7): p. 553-558.
73. Aloy, P., et al., *Automated structure based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from*

- homology in genome annotation and to protein docking.* J. Mol. Biol., 2001. **311**: p. 395-408.
74. Armon, A., D. Graur, and N. Ben-Tal, *ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.* J. Mol. Biol., 2001. **307**: p. 447-463.
 75. Zvelebil, M.J.J., et al., *Prediction of protein secondary structure and active sites using the alignment of homologous sequences.* J. Mol. Biol., 1987. **195**: p. 957-961.
 76. Moore, G.L., et al., *Modelling and optimization of DNA recombination.* Comp. Chem. Eng., 2000. **24**: p. 693-699.
 77. Moore, G.L., et al., *Predicting crossover generation in DNA shuffling.* Proc. Natl. Acad. Sci. USA, 2001. **98**(6): p. 3226-3231.
 78. Voigt, C.A., et al., *Computationally focusing the directed evolution of proteins.* J. Cell. Biochem. Suppl., 2001. **37**: p. 58-63.
 79. Cannon, W.R., B.J. Garrison, and S.J. Benkovic, *Electrostatic characterization of enzyme complexes: Evaluation of the mechanism of catalysis of dihydrofolate reductase.* J. Am. Chem. Soc., 1997. **119**: p. 2386-2395.
 80. Gong, X.S., et al., *The role individual lysine residues in the basic patch on turnip cytochrome F for electrostatic interactions with plastocyanin in vitro.* Eur. J. Biol. Chem./FEBS, 2000. **267**: p. 3461-3468.
 81. Fisher, B.M., L.W. Schultz, and R.T. Raines, *Coulombic effects of remote subsites on the active site of ribonuclease A.* Biochemistry, 1998. **37**: p. 17386-17401.
 82. Radkiewicz, J.L. and C.L. Brooks, *Protein dynamics in enzymatic catalysis: exploration of dihydrofolate reductase.* J. Am. Chem. Soc., 2000. **122**: p. 225-231.
 83. Lichtarge, O. and M.E. Sowa, *Evolutionary predictions of binding surfaces and interactions.* Curr Opin Struct Biol, 2002. **12**(1): p. 21-7.
 84. Landgraf, R., I. Xenarios, and D. Eisenberg, *Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins.* J Mol Biol, 2001. **307**(5): p. 1487-502.
 85. Shindyalov, I.N., N.A. Kolchanov, and C. Sander, *Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?* Protein Eng., 1994. **7**(3): p. 349-358.
 86. Baldwin, E.P., W.A. Hajiseyedjavadi, and B.W. Matthews, *The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme.* Science, 1993. **262**: p. 1715-1718.
 87. Lim, V.I. and O.B. Ptitsyn, *On the constancy of hydrophobic nucleus volume in molecules of myoglobin and hemoglobins.* Mol. Biol. (USSR), 1970. **4**: p. 372-382.
 88. Gobel, U., et al., *Correlated mutations and residue contacts in proteins.* Proteins, 1994. **18**(4): p. 309-317.
 89. Altschuh, D., et al., *Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus.* J. Mol. Biol., 1987. **193**: p. 693-707.
 90. Bordo, D. and P. Argos, *Evolution of protein cores: constraints in point mutations as observed in globin tertiary structures.* J. Mol. Biol., 1990. **211**: p. 975-988.

91. Lesk, A.M. and C. Chothia, *How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins*. J. Mol. Biol., 1980. **136**: p. 225-270.
92. Oosawa, K. and M. Simon, *Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in Escherichia coli*. Proc. Natl. Acad. Sci. USA, 1986. **83**: p. 6930-6934.
93. Neher, E., *How frequent are correlated changes in families of protein sequences?* Proc. Natl. Acad. Sci. USA, 1994. **91**(1): p. 98-102.
94. Taylor, W.R. and K. Hatrick, *Compensating changes in protein multiple sequence alignments*. Protein Eng., 1994. **7**(3): p. 341-348.
95. Fukami-Kobayashi, K., D.R. Schreiber, and S.A. Benner, *Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences*. J. Mol. Biol., 2002. **319**: p. 729-743.
96. Olmea, O., B. Rost, and A. Valencia, *Effective use of sequence correlation and conservation in fold recognition*. J. Mol. Biol., 1999. **295**: p. 1221-1239.
97. Miller, G.P. and S.J. Benkovic, *Deletion of a highly motional residue affects formation of the michaelis complex for Escherichia coli dihydrofolate reductase*. Biochemistry, 1998. **37**: p. 6327-6335.
98. Osborne, M.J., et al., *Backbone dynamics in dihydrofolate reductase complexes: Role of loop flexibility in the catalytic mechanism*. Biochemistry, 2001. **40**: p. 9846-9859.
99. Miller, G.P., D.C. Wahnou, and S.J. Benkovic, *Interloop contacts modulate ligand cycling during catalysis by Escherichia coli Dihydrofolate Reductase*. Biochemistry, 2001. **40**: p. 867-875.
100. Bateman, A., et al., *The Pfam protein families database*. Nucleic Acids Res., 2002. **30**(1): p. 276-280.
101. Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt, *A model of evolutionary change in proteins*. Atlas of Protein Sequence and Structure, 1978. **5**: p. 345-352.
102. Lim, W.A. and R.T. Sauer, *Alternative packing arrangements in the hydrophobic core of lambda repressor*. Nature, 1989. **399**: p. 31-36.
103. Grantham, R., *Amino acid difference formula to help explain protein evolution*. Science, 1974. **185**(154): p. 862-864.
104. Levitt, M., *A simplified representation of protein conformations for rapid simulation of protein folding*. J. Mol. Biol., 1976. **104**: p. 59-107.
105. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks*. Proc. Natl. Acad. Sci. USA, 1992. **89**: p. 10915-10919.
106. McLachlan, A.D., *Tests for comparing related amino-acid sequences. cytochrome C and cytochrome C 551*. J. Mol. Biol., 1971. **61**(2): p. 409-424.
107. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families*. Science, 1999. **286**(5438): p. 295-299.
108. Thomas, D.J., G. Casari, and C. Sander, *The prediction of protein contacts from multiple sequence contacts*. Protein Eng., 1996. **9**: p. 941-948.
109. Olmea, O. and A. Valencia, *Improving contact predictions by the combination of correlated mutations and other sources of sequence information*. Fold Des., 1997. **2**: p. S25-S32.

110. Lund, O., et al., *Protein distance constraints predicted by neural networks and probability density functions*. Protein Eng., 1997. **10**: p. 1241-1248.
111. Fariselli, P., et al., *Progress in prediction inter-residue contacts of proteins with neural networks and correlated mutations*. Proteins, 2001. **5**: p. 157-162.
112. Vendruscolo, M., E. Kussell, and E. Domany, *Recovery of protein structure from contact maps*. Fold Des., 1997. **2**: p. 295-306.
113. Fariselli, P. and R. Casadio, *Neural network based predictor of residue contacts in proteins*. Protein Eng., 1999. **12**: p. 15-21.
114. Higgins, D.G. and P.M. Sharp, *CLUSTAL: A package for performing multiple sequence alignment on a micro computer*. Gene, 1988. **73**: p. 237-244.
115. Higgins, D.G., J.D. Thompson, and T.J. Gibson, *Using CLUSTAL for multiple sequence alignments*. Methods Enzymol., 1996. **266**: p. 383-402.
116. Mendes, J., et al., *Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants*. J. Comput. Aided Mol. Des., 2001. **15**(8): p. 721-740.
117. Agarwal, P.K., et al., *Network of coupled promoting motions in enzyme catalysis*. Proc. Natl. Acad. Sci. USA, 2002. **99**(5): p. 2794-2799.
118. Falzone, C.J., S.J. Benkovic, and P.E. Wright, *Partial ¹H NMR assignments of the Escherichia coli dihydrofolate reductase complex with folate: evidence for a unique conformation of bound folate*. Biochemistry, 1990. **29**: p. 9667-9677.
119. Lau, E.Y. and J.T. Gerig, *Effects of fluorine substitution on the structure and dynamics of dihydrofolate reductase*. Biophys. J., 1997. **73**: p. 1579-1592.
120. Li, L., et al., *Functional role of a mobile loop of Escherichia coli dihydrofolate reductase in transition-state stabilization*. Biochemistry, 1992. **31**(34): p. 7826-7833.
121. Eisenmesser, E.Z., et al., *Enzyme dynamics during catalysis*. Science, 2002. **295**: p. 1520-1523.
122. Kallen, J. and M.D. Walkinshaw, *The X-ray structure of a tetrapeptide bound to the active site of human cyclophilin A*. FEBS Letters, 1992. **300**: p. 286-290.
123. Zhao, Y. and H. Ke, *Crystal structure implies that cyclophilin predominantly catalyzes the trans to cis isomerization*. Biochemistry, 1996. **35**: p. 7356-7361.
124. Chen, P., et al., *Crystal structure of glycineamide ribonucleotide transformylase from Escherichia coli at 3.0 Å resolution*. J. Mol. Biol., 1992. **227**: p. 283-292.
125. Almasy, R.J., et al., *Structures of apo and complexed Escherichia coli glycineamide ribonucleotide transformylase*. Proc. Natl. Acad. Sci. USA, 1992. **89**: p. 6114-6118.
126. Klein, C., et al., *Towards structure-based drug design: crystal structure of a multisubstrate adduct complex of glycineamide ribonucleotide transformylase at 1.96 Å resolution*. J. Mol. Biol., 1995. **249**: p. 153-175.
127. Inglese, J., J.M. Smith, and S.J. Benkovic, *Active-site mapping and site-specific mutagenesis of glycineamide ribonucleotide transformylase from Escherichia coli*. Biochemistry, 1990. **29**: p. 6678-6687.
128. Morikis, D., et al., *Protein transfer dynamics of GART: The pH-dependent catalytic mechanism examined by electrostatic calculations*. Protein Sci., 2001. **10** (11): p. 2379-2392.

129. Ouali, M. and R.D. King, *Cascaded multiple classifiers for secondary structure prediction*. Protein Sci., 2000. **9**: p. 1162-1176.
130. Moore, J.C., et al., *Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences*. J Mol Biol, 1997. **272**(3): p. 336-47.
131. Wang, P.L., *Creating hybrid genes by homologous recombination*. Dis Markers, 2000. **16**(1-2): p. 3-13.
132. Saraf, M.C., G.L. Moore, and C.D. Maranas, *Using multiple sequence correlation analysis to characterize functionally important protein regions*. Protein Eng, 2003. **16**(6): p. 397-406.
133. Voigt, C.A., et al., *Computationally focusing the directed evolution of proteins*. J Cell Biochem Suppl, 2001. **Suppl 37**: p. 58-63.
134. Bogarad, L.D. and M.W. Deem, *A hierarchical approach to protein molecular evolution*. Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2591-5.
135. Voigt, C.A., et al., *Protein building blocks preserved by recombination*. Nat Struct Biol, 2002. **9**(7): p. 553-8.
136. Meyer, M.M., et al., *Library analysis of SCHEMA-guided protein recombination*. Protein Sci, 2003. **12**(8): p. 1686-93.
137. Moore, G.L. and C.D. Maranas, *Identifying residue-residue clashes in protein hybrids by using a second-order mean-field approach*. Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5091-6.
138. Oldfield, T.J., *Data mining the protein data bank: residue interactions*. Proteins, 2002. **49**(4): p. 510-28.
139. Glusker, J.P., *Structural aspects of metal liganding to functional groups in proteins*. Adv Protein Chem, 1991. **42**: p. 1-76.
140. Fischer, D., et al., *Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding*. Protein Sci, 1994. **3**(5): p. 769-78.
141. Wallace, A.C., N. Borkakoti, and J.M. Thornton, *TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites*. Protein Sci, 1997. **6**(11): p. 2308-23.
142. Copley, R.R. and G.J. Barton, *A structural analysis of phosphate and sulphate binding sites in proteins. Estimation of propensities for binding and conservation of phosphate binding sites*. J Mol Biol, 1994. **242**(4): p. 321-9.
143. Chakrabarti, P., *Anion binding sites in protein structures*. J Mol Biol, 1993. **234**(2): p. 463-82.
144. Westbrook, J., et al., *The Protein Data Bank: unifying the archive*. Nucleic Acids Res, 2002. **30**(1): p. 245-8.
145. Gobel, U., et al., *Correlated mutations and residue contacts in proteins*. Proteins, 1994. **18**(4): p. 309-17.
146. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path*. Protein Eng, 1998. **11**(9): p. 739-47.

147. Schwede, T., et al., *SWISS-MODEL: An automated protein homology-modeling server*. Nucleic Acids Res, 2003. **31**(13): p. 3381-5.
148. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.
149. van Gunsteren, et al., *Biomolecular Simulations: The GROMOS96 Manual and User Guide*, 1996.
150. Munson, M., et al., *What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties*. Protein Sci, 1996. **5**(8): p. 1584-93.
151. Dupraz, P., et al., *Point mutations in the proximal Cys-His box of Rous sarcoma virus nucleocapsid protein*. J Virol, 1990. **64**(10): p. 4978-87.
152. Ratnaparkhi, G.S. and R. Varadarajan, *Thermodynamic and structural studies of cavity formation in proteins suggest that loss of packing interactions rather than the hydrophobic effect dominates the observed energetics*. Biochemistry, 2000. **39**(40): p. 12365-74.
153. Song, K.S., et al., *A molecular model of a point mutation (Val297Met) in the serine protease domain of protein C*. Exp Mol Med, 1999. **31**(1): p. 47-51.
154. Vriend, G., *WHAT IF: a molecular modeling and drug design program*. J Mol Graph, 1990. **8**(1): p. 52-6, 29.
155. Agarwal, P.K., et al., *Network of coupled promoting motions in enzyme catalysis*. Proc Natl Acad Sci U S A, 2002. **99**(5): p. 2794-9.
156. Loll, B., et al., *Functional role of C(alpha)-H...O hydrogen bonds between transmembrane alpha-helices in photosystem I*. J Mol Biol, 2003. **328**(3): p. 737-47.
157. Ostermeier, M., *Theoretical distribution of truncation lengths in incremental truncation libraries*. Biotechnol Bioeng, 2003. **82**(5): p. 564-77.
158. Hansson, L.O., et al., *Evolution of differential substrate specificities in Mu class glutathione transferases probed by DNA shuffling*. J Mol Biol, 1999. **287**(2): p. 265-76.
159. Hansson, L.O., et al., *Structural determinants in domain II of human glutathione transferase M2-2 govern the characteristic activities with aminochrome, 2-cyano-1,3-dimethyl-1-nitrosoguanidine, and 1,2-dichloro-4-nitrobenzene*. Protein Sci, 1999. **8**(12): p. 2742-50.
160. Kikuchi, M., K. Ohnishi, and S. Harayama, *An effective family shuffling method using single-stranded DNA*. Gene, 2000. **243**(1-2): p. 133-7.
161. Peitsch, M.C., *ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling*. Biochem Soc Trans, 1996. **24**(1): p. 274-9.
162. Joern, J.M., P. Meinhold, and F.H. Arnold, *Analysis of shuffled gene libraries*. J Mol Biol, 2002. **316**(3): p. 643-56.
163. Schwede, T., et al., *Protein structure computing in the genomic era*. Res Microbiol, 2000. **151**(2): p. 107-12.
164. Ostermeier, M., *Synthetic gene libraries: in search of the optimal diversity*. Trends Biotechnol, 2003. **21**(6): p. 244-7.
165. Ness, J.E., et al., *Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently*. Nat Biotechnol, 2002. **20**(12): p. 1251-5.

166. Moore, G.L. and C.D. Maranas, *Computational Challenges in Combinatorial Library Design for Protein Engineering*. AIChE Journal, 2003: p. accepted.
167. Saraf, M.C. and C.D. Maranas, *Using a Residue Clash Map to Functionally Characterize Protein Hybrids*. Protein Eng, 2003: p. accepted.
168. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families*. Science, 1999. **286**(5438): p. 295-9.
169. Gaucher, E.A., et al., *Predicting functional divergence in protein evolution by site-specific rate shifts*. Trends Biochem Sci, 2002. **27**(6): p. 315-21.
170. del Sol Mesa, A., F. Pazos, and A. Valencia, *Automatic methods for predicting functionally important residues*. J Mol Biol, 2003. **326**(4): p. 1289-302.
171. Liang, M.P., D.L. Brutlag, and R.B. Altman, *Automated construction of structural motifs for predicting functional sites on protein structures*. Pac Symp Biocomput, 2003: p. 204-15.
172. Larson, S.M., A.A. Di Nardo, and A.R. Davidson, *Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions*. J Mol Biol, 2000. **303**(3): p. 433-46.
173. Gaucher, E.A., et al., *Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins*. Nature, 2003. **425**(6955): p. 285-8.
174. Fukami-Kobayashi, K., D.R. Schreiber, and S.A. Benner, *Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences*. J Mol Biol, 2002. **319**(3): p. 729-43.
175. Govindarajan, S., et al., *Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation*. J Mol Biol, 2003. **328**(5): p. 1061-9.
176. Di Nardo, A.A., S.M. Larson, and A.R. Davidson, *The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core*. J Mol Biol, 2003. **333**(3): p. 641-55.
177. Taylor, W.R. and K. Hatrick, *Compensating changes in protein multiple sequence alignments*. Protein Eng, 1994. **7**(3): p. 341-8.
178. Altschuh, D., et al., *Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus*. J Mol Biol, 1987. **193**(4): p. 693-707.
179. Ptitsyn, O.B. and A.V. Finkel'shtein, *[Predicting the spiral portions of globular proteins from their primary structure]*. Dokl Akad Nauk SSSR, 1970. **195**(1): p. 221-4.
180. Shindyalov, I.N., N.A. Kolchanov, and C. Sander, *Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?* Protein Eng, 1994. **7**(3): p. 349-58.
181. Lesk, A.M. and C. Chothia, *How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins*. J Mol Biol, 1980. **136**(3): p. 225-70.
182. Rod, T.H., J.L. Radkiewicz, and C.L. Brooks, 3rd, *Correlated motion and the effect of distal mutations in dihydrofolate reductase*. Proc Natl Acad Sci U S A, 2003. **100**(12): p. 6980-5.

183. Gong, X.S., et al., *The role of individual lysine residues in the basic patch on turnip cytochrome f for electrostatic interactions with plastocyanin in vitro*. Eur J Biochem, 2000. **267**(12): p. 3461-8.
184. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins*. Nat Struct Biol, 2003. **10**(1): p. 59-69.
185. Lutz, R. and H. Bujard, *Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements*. Nucleic Acids Res, 1997. **25**(6): p. 1203-10.
186. Ostermeier, M., J.H. Shim, and S.J. Benkovic, *A combinatorial approach to hybrid enzymes independent of DNA homology*. Nat Biotechnol, 1999. **17**(12): p. 1205-9.
187. Horswill, A.R. and S.J. Benkovic, (Unpublished results), 2003.
188. Ausubel, F.M., et al., *Short Protocols in Molecular Biology*. John Wiley & Sons, Inc., New York, 2002.
189. Lutz, S., M. Ostermeier, and S.J. Benkovic, *Rapid generation of incremental truncation libraries for protein engineering using alpha-phosphothioate nucleotides*. Nucleic Acids Res, 2001. **29**(4): p. E16.
190. Herrington, M.B. and N.T. Chirwa, *Growth properties of a folA null mutant of Escherichia coli K12*. Can J Microbiol, 1999. **45**(3): p. 191-200.
191. Rajagopalan, P.T., S. Lutz, and S.J. Benkovic, *Coupling interactions of distal residues enhance dihydrofolate reductase catalysis: mutational effects on hydride transfer rates*. Biochemistry, 2002. **41**(42): p. 12618-28.
192. Bateman, A., et al., *The Pfam protein families database*. Nucleic Acids Res, 2002. **30**(1): p. 276-80.
193. Klein, P., M. Kanehisa, and C. DeLisi, *Prediction of protein function from sequence properties. Discriminant analysis of a data base*. Biochim Biophys Acta, 1984. **787**(3): p. 221-6.
194. Krigbaum, W.R. and A. Komoriya, *Local interactions as a structure determinant for protein molecules: II*. Biochim Biophys Acta, 1979. **576**(1): p. 204-48.
195. Cid, H., et al., *Hydrophobicity and structural classes in proteins*. Protein Eng, 1992. **5**(5): p. 373-5.
196. Engelman, D.M., T.A. Steitz, and A. Goldman, *Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins*. Annu Rev Biophys Biophys Chem, 1986. **15**: p. 321-53.
197. Sawaya, M.R. and J. Kraut, *Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence*. Biochemistry, 1997. **36**(3): p. 586-603.
198. Radkiewicz, J.L. and C.L. Brooks, 3rd, *Protein Dynamics in Enzyme Catalysis: Exploration of Dihydrofolate Reductase*. J. Am. Chem. Soc., 2000. **122**: p. 225-231.
199. Fierke, C.A., K.A. Johnson, and S.J. Benkovic, *Construction and evaluation of the kinetic scheme associated with dihydrofolate reductase from Escherichia coli*. Biochemistry, 1987. **26**(13): p. 4085-92.
200. Pan, H., J.C. Lee, and V.J. Hilser, *Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble*. Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12020-5.

201. Huang, Z., C.R. Wagner, and S.J. Benkovic, *Nonadditivity of mutational effects at the folate binding site of Escherichia coli dihydrofolate reductase*. Biochemistry, 1994. **33**(38): p. 11576-85.
202. Stemmer, W.P., *Rapid evolution of a protein in vitro by DNA shuffling*. Nature, 1994. **370**(6488): p. 389-91.
203. Moore, G.L. and C.D. Maranas, *Computational Challenges in Combinatorial Library Design for Protein Engineering*. AIChE Journal, 2004. **50**(2): p. 262-272.
204. Coco, W.M., et al., *Growth factor engineering by degenerate homoduplex gene family recombination*. Nat Biotechnol, 2002. **20**(12): p. 1246-50.
205. Patrick, W.M., A.E. Firth, and J.M. Blackburn, *User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries*. Protein Eng, 2003. **16**(6): p. 451-7.
206. Osborne, M.J., et al., *Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism*. Biochemistry, 2001. **40**(33): p. 9846-59.
207. Saraf, M.C., et al., *FamClash: a method for ranking the activity of engineered enzymes*. Proc Natl Acad Sci U S A, 2004. **101**(12): p. 4142-7.
208. Saraf, M.C. and C.D. Maranas, *Using a residue clash map to functionally characterize protein recombination hybrids*. Protein Eng, 2003. **16**(12): p. 1025-34.
209. Endelman, J.B., et al., *Site-directed protein recombination as a shortest-path problem*. Protein Eng Des Sel, 2004.
210. Lichtarge, O., H.R. Bourne, and F.E. Cohen, *An evolutionary trace method defines binding surfaces common to protein families*. J Mol Biol, 1996. **257**(2): p. 342-58.
211. Nagi, A.D. and L. Regan, *An inverse correlation between loop length and stability in a four-helix-bundle protein*. Fold Des, 1997. **2**(1): p. 67-75.
212. Glover, F., *Improved linear integer programming formulations of nonlinear integer problems*. Management Science, 1975. **22**(4): p. 455-460.
213. Bateman, A., et al., *The Pfam protein families database*. Nucleic Acids Res, 2004. **32 Database issue**: p. D138-41.
214. ILOG, I., *ILOG AMPL CPLEX System Version 7.0 User's Guide*. 2000.
215. Brooke, A., et al., *GAMS: A user's guide*. 1998.
216. Looger, L.L., et al., *Computational design of receptor and sensor proteins with novel functions*. Nature, 2003. **423**(6936): p. 185-90.
217. Moore, G.L. and C.D. Maranas, *Computational Challenges in Combinatorial Library Design for Protein Engineering*. AIChE J., 2004: p. in press.
218. Lovell, S.C., et al., *The penultimate rotamer library*. Proteins, 2000. **40**(3): p. 389-408.
219. Miyazaki, K., et al., *Directed evolution study of temperature adaptation in a psychrophilic enzyme*. J. Mol. Biol., 2000. **297**(4): p. 1015-1026.
220. Baik, S.H., et al., *Significantly enhanced stability of glucose dehydrogenase by directed evolution*. Appl. Microbiol. Biotechnol., 2003. **61**(4): p. 329-335.

221. Reetz, M.T., et al., *Directed Evolution of an Enantioselective Enzyme through Combinatorial Multiple-Cassette Mutagenesis*. Angew. Chem. Int. Ed. Engl., 2001. **40**(19): p. 3589-3591.
222. Horsman, G.P., et al., *Mutations in distant residues moderately increase the enantioselectivity of Pseudomonas fluorescens esterase towards methyl 3bromo-2-methylpropanoate and ethyl 3phenylbutyrate*. Chemistry, 2003. **9**(9): p. 1933-9.
223. Carr, R., et al., *Directed Evolution of an Amine Oxidase Possessing both Broad Substrate Specificity and High Enantioselectivity*. Angew. Chem. Int. Ed. Engl., 2003. **42**(39): p. 4807-4810.
224. Furukawa, K., *Engineering dioxygenases for efficient degradation of environmental pollutants*. Curr. Opin. Biotechnol., 2000. **11**: p. 244-249.
225. Wackett, L.P., *Directed evolution of new enzymes and pathways for environmental catalysis*. Ann. NY Acad. Sci., 1998. **864**: p. 142-152.
226. Bruhlmann, F. and W. Chen, *Tuning biphenyl dioxygenase for extended substrate specificity*. Biotechnol. Bioeng., 1999. **63**(5): p. 544-551.
227. Whalen, R.G., et al., *DNA shuffling and vaccines*. Curr. Opin. Mol. Ther., 2001. **3**(1): p. 31-36.
228. Patten, P.A., R.J. Howard, and W.P. Stemmer, *Applications of DNA shuffling to pharmaceuticals and vaccines*. Curr. Opin. Biotechnol., 1997. **8**(6): p. 724-733.
229. Marzio, G., et al., *In vitro evolution of a highly replicating, doxycycline-dependent HIV for applications in vaccine studies*. Proc. Natl. Acad. Sci. USA, 2001. **98**(11): p. 6342-6347.
230. Moore, J.C., et al., *Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences*. J. Mol. Biol., 1997. **272**: p. 336-347.
231. Saraf, M.C. and C.D. Maranas, *Using a Residue Clash Map to Functionally Characterize Protein Recombination Hybrids*. Protein Eng., 2003: p. in press.
232. Saraf, M.C., et al., *FamClash: A Method for Ranking the Activity of Engineered Enzymes*. Proc. Natl. Acad. Sci. USA, 2004: p. accepted.
233. Moore, G.L. and C.D. Maranas, *Identifying residue-residue clashes in protein hybrids by using a second-order mean-field approach*. Proc. Natl. Acad. Sci. USA, 2003. **100**(9): p. 5091-5096.
234. Maheshri, N. and D.V. Schaffer, *Computational and experimental analysis of DNA shuffling*. Proc. Natl. Acad. Sci. USA, 2003. **100**(6): p. 3071-3076.
235. Richardson, T.H., et al., *A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable alpha-amylase*. J. Biol. Chem., 2002. **277**(29): p. 26501-26507.
236. Hiraga, K. and F.H. Arnold, *General method for sequence-independent site-directed chimeragenesis*. J. Mol. Biol., 2003. **330**(2): p. 287-296.
237. Coco, W.M., et al., *DNA shuffling method for generating highly recombined genes and evolved enzymes*. Nat. Biotechnol., 2001. **19**(4): p. 354-359.
238. Ridder, L., et al., *Quantum mechanical/molecular mechanical free energy simulations of the glutathione S-transferase (M1-1) reaction with phenanthrene 9,10-oxide*. J Am Chem Soc, 2002. **124**(33): p. 9926-36.

- 239. Bates, P.A., et al., *Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM*. Proteins, 2001. **Suppl 5**: p. 39-46.
- 240. Bonneau, R., et al., *Rosetta in CASP4: progress in ab initio protein structure prediction*. Proteins, 2001. **Suppl 5**: p. 119-26.
- 241. Chen, R., et al., *ZDOCK predictions for the CAPRI challenge*. Proteins, 2003. **52**(1): p. 68-73.
- 242. Chen, R., L. Li, and Z. Weng, *ZDOCK: an initial-stage protein-docking algorithm*. Proteins, 2003. **52**(1): p. 80-7.
- 243. Pardalos, P.M. and H. Wolkowicz, *Preface*. J. Comb. Optim., 2002. **6**(3): p. 235-236.
- 244. Looger, L.L. and H.W. Hellinga, *Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics*. J Mol Biol, 2001. **307**(1): p. 429-45.
- 245. Lutz, S., M. Ostermeier, and S.J. Benkovic, *Rapid generation of incremental truncation libraries for protein engineering using alpha-phosphothioate nucleotides*. Nucleic Acids Res., 2001. **29**(4): p. E16.
- 246. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path*. Protein Eng., 1998. **11**(9): p. 739-747.
- 247. Dunbrack, R.L., Jr. and F.E. Cohen, *Bayesian statistical analysis of protein side-chain rotamer preferences*. Protein Sci, 1997. **6**(8): p. 1661-81.
- 248. Westbrook, J., et al., *The Protein Data Bank: unifying the archive*. Nucleic Acids Res., 2002. **30**(1): p. 245-248.

VITA

Manish C. Saraf

Education

Pennsylvania State University, University Park, PA 16802, USA
Ph.D. in Chemical Engineering
Spring, 2006

Pennsylvania State University, University Park, PA 16802, USA
Minor in Operations Research
Spring, 2006

Indian Institute of Technology (IIT), Mumbai, India
Bachelor of Technology (B. Tech) in Chemical Engineering
May, 2001

PhD Thesis Topic

Development of Computational Tools for the Design and Optimization of Combinatorial Protein Libraries

Honors

- 2005, **Pan American Study Institute on Process Systems Engineering** Travel Grant, Carnegie Mellon University.
- 2002, **Chemical Engineering Graduate Fellowship**, Penn State University.
- 2002, **Graduate School Fellowship**, Penn State University.
- 1997, Secured 3rd place out of 30,000 applicants in Roorkee Architecture and Engineering Entrance Examination (INDIA).

Publications

- Saraf, M.C., Moore, G.L., Goodey, N.M., Cao, V.Y., Benkovic, S.J., and Maranas, C.D. (2005), "IPRO: An Iterative Computational Protein Library Redesign and Optimization Procedure," *Biophysical Journal*, Vol. 90(11).
- Saraf, M.C., Gupta, A., and Maranas, C.D. (2005), "Design of Combinatorial Protein Libraries of Optimal Size," *PROTEINS: Structure, Function, and Bioinformatics*, Vol. 60(4). 767-777.
- Saraf, M.C., Horswill, A.R., Benkovic, S.J., and Maranas, C.D. (2004), "FamClash: A Novel Method for Predicting Activity of Engineered Enzymes," *Proc Natl Acad Sci USA*, Vol. 101 (12). 4142-4147.
- Saraf, M.C., and Maranas, C.D. (2003), "Using a Residue Clash Map to Functionally Characterize Protein Recombination Hybrids," *Protein Engineering*, Vol. 16(12). 1025-1034.
- Saraf, M.C., Moore, G.L., and Maranas, C.D. (2003), "Using Multiple Sequence Correlation Analysis to Characterize Functionally Important Protein Regions," *Protein Engineering*, Vol. 16 (6). 397-406