Computationally Optimizing the Directed Evolution of Proteins

Thesis by

Christopher Ashby Voigt

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

California Institute of Technology

Pasadena, California, USA

2002

(Submitted July 25, 2002)

Abstract

Directed evolution has proven a successful strategy for protein engineering. To accelerate the discovery process, we have developed several computational methods to optimize the mutant libraries by targeting specific residues for mutagenesis, and subunits for recombination. In achieving this goal, a statistical model was first used to study the dynamics of directed evolution as a search algorithm. These simulations improved our understanding of the relationship between parameters describing the search space (e.g., interactions between amino acids) and experimental search parameters (e.g., mutation rate and library size). Based on these simulations, a more detailed model was used to calculate the structural tolerance of each residue to amino acid substitutions. Further, a computational model was developed to optimize recombination experiments, based on the three-dimensional structure. Together, these computational techniques represent a major step towards information-driven combinatorial protein design.

Acknowledgements

Just before I started writing my thesis, I imagined that I would probably write the title page, copyright page, an over-ambitious table of contents, and acknowledgements first. I would then procrastinate for the next six months in putting together the bulk of the writing. In reality, the thesis has come together in the opposite way. I have found writing the acknowledgements to be much more difficult than expected, and this has been compounded as I put each chapter together and remember the enormous group of people that have made this possible.

Foremost, I need to acknowledge my three advisors at Caltech (in order of appearance): Zhen-Gang Wang, Frances Arnold, and Stephen Mayo. They provided a wonderful collaborative environment, which gave me the freedom to explore different ideas. I am grateful for the unique training in theoretical, computation, and experimental methods that they provided. Working with them has made for a very enjoyable graduate experience.

There are members of all three groups with whom I have collaborated directly. When I first arrived in the Mayo group, Ben Gordon and Arthur Street introduced me to the computational methods. I enjoyed trading wacky ideas on combinatorial optimization with Ben and am still in awe of Arthur's programming abilities. When my project took an experimental turn, Carlos Martinez of the Arnold group and Rhonda Digiusto of the Mayo group introduced me to the necessary molecular biology. I am indebted to Jonathan Silberg, Michelle Meyer, Kaori Hiraga, Radu Georgesco, and Chris Otey in the Arnold group for testing the ideas in this thesis on many types of enzyme systems. In the Mayo group, I have worked closely

with Deepshikha Datta in trying to discover structural switches in proteins. From the Wang group, I have had many stimulating conversations on the dynamics of polymers with Andy Spakowitz.

Over time, I have interacted with many new and old members of the three groups. Their ideas have subtly altered the direction of my research interests. In the Mayo group, I would like to thank Oscar Alvizo, Julie Archer, Mary Ary, Dan Bolon, Cynthia Carlson, Eun Jung Choi, Deepshikha Datta, Geoffrey Hom, Possu Huang, Shira Jacobson, Caglar Tanrikulu, Kirsten Lasilla, John Love, Jessica Mao, Andrei Marinescu, Shannon Marshall, Chantel Morgan, J. J. Plecs, Scott Ross, Cathy Sarisky, Julia Schifman, Premal Shah, Pavel Strop, and Eric Zollars. A special thanks goes to Dr. Love for teaching me the sins of Vegas and taking my money. Luckily, Possu, Geoff, and Premal were also playing so I could ultimately break even. Currently, Peter Samuelson, a freshman, is programming Web applications based on this thesis. He will probably be done before I finish writing this. The lab would not run nearly as smoothly if it were not for Cynthia Carlson. She made everything come together. Finally, Darryl Willick, Ryan Martin, and Hezekiah McMurray need to be thanked for their behind-the-scenes work to keep the computers running.

The people in the Arnold group have been very helpful in directing my ideas by teaching me how computation can influence the experiments. When I joined the lab, Ann Gershenson and Patrick Wintrode introduced me to the concept of directed evolution and the potential for computational techniques. Other past and present members with whom I have interacted include Geethani Bandara, Thomas Buelter, Patrick Cirino, Cynthia Collins, Edgardo Farinas, Ann Fu, Radu Georgescu, Kaori

Hiraga, John Joern, Oliver May, Kimberly Mayer, Peter Meinhold, Kentaro Miyazaki, Michelle Meyer, Peter Nguyen, Chris Otey, Ioanna Petrounia, Claudia Schmidt-Dannert, Rebecca Schulman, Uli Schwaneberg, Volker Sieber, Jonathan Silberg, Lianhong Sun, Todd Thorson, Alex Tobias, Andrew Udit, Daisuke Umeno, Alex Volcov and Yohei Yokobayashi. I should thank John Joern for letting me live with him, which is not a minor thing if you have seen my office. During recruitment, Pat Cirino convinced me to come to Caltech and has subsequently stomached my rants regarding graduate life.

Niles Pierce has played an active role in my graduate career, both as a postdoc in the Mayo lab and as a professor. I particularly appreciate advice that he gave me in using Matlab and applying for jobs. His wife, Gillian, was very helpful in editing a large book chapter that I wrote. Near the end of my time here, I have gotten to know one of his students, Robert Dirks, mostly through his pervasive use of computer time.

I also need to thank the members of my thesis committee that are not advisors: Doug Rees, Richard Roberts, and Walter Fontana. Walter is a scientist at the Santa Fe Institute, a research center in New Mexico that focuses on interdisciplinary theoretical problems. Frances was mostly responsible for introducing me to the SFI and for that I thank her. Stu Kauffman deserves some recognition for overseeing the writing of a book chapter. As an undergraduate, I was inspired by his book, “Origins of Order,” so it was an honor to work with him. Throughout the last year, I was able to work with Lauren Ancel to develop a workshop on robustness and evolvability. This was a lot of fun and was made possible through an initiative headed by Erica Jen.

Beyond the SFI, I have collaborated with researchers at other institutions. Dane Wittrup at MIT is one of the first people to test the ideas in this thesis on creating libraries on antibodies (Chapter 4). His student, Brenda Kellogg, was responsible for lab work for this project. I have also collaborated with Jim Bull, George Georgiou, and Brent Iverson at the University of Texas, Austin. Their ability to produce massive amounts of data inspired some of the work presented here.

Prior to graduate school, there were numerous teachers and professors that had very positive and influential roles. When I was a student at Brandywine High School, Ronald Eschelman managed to motivate me in chemistry. In addition, I was inspired by Vincent Pro (history), Jeff Stugard (drafting), and Tom Twilley (biology). At the University of Michigan, Robert Ziff introduced me to statistical mechanics, and more importantly, the concept of doing research. While I was still at Michigan, I met Richard Goldstein in the biophysics research division. Through his language and colorful use of analogies, Richard has greatly influenced my speaking style.

There are many friends and family that provided tremendous support throughout graduate school. My parents, Henry and Julie Voigt, were very encouraging, especially at the beginning. They really helped me keep everything in perspective. My sister Emily was an English major at Columbia and her skills with language were put to use when she edited several grant applications. Finally, I would like to thank Hyon-Jee for keeping me sane the last four years. She has been an intelligent and beautiful classmate, friend, and girlfriend.

Table of Contents

Abstract iii
Acknowledgements iv
Table of Contents ix
List of Tables xi
List of Figures xii

Chapters

Chapter 1 1-1
Introduction to Directed Evolution Theory

Chapter 2 2-1
Modeling the Dynamics of Directed Evolution

Chapter 3 3-1
Targeting Mutagenesis with Structural Information

Chapter 4 4-1
Tolerance of the CDRs of Antibody D1.3

Chapter 5 5-1
Recombination Preserves Protein Building Blocks

Chapter 6 6-1
Comparing Search Algorithms in Protein Design

Chapter 7 7-1
Evolvable Systems in Biology

Appendixes

Appendix A A-1
Higher-Order Moments of the Mutant Distribution

Appendix B B-1
Dead-End Elimination and Monte Carlo Entropy Calculations

Appendix C C-1
Adding Ambient Temperature to the Sequence Entropy

Appendix D D-1
Calculating a Joint Entropy for Two Structures

Appendix E E-1
Combinatorial Libraries Based on Schema Disruption

Appendix F F-1
Non-homologous Recombination

References R-1

List of Tables

Table 1-1: Discovery times for evolutionary algorithms

Table 3-1: Site entropies and solvent accessibility of subtilisin E
Table 3-2: Site entropies and solvent accessibility of T4 lysozyme
Table 3-3: Saturation of high-entropy residues of β-lactamase

Table 4-1: Full ORBIT sequence design of CDR residues
Table 4-2: ORBIT design of targeted residues
Table 4-3: Comparison of experimental and computational results

Table 5-1: Designed TEM-1/PSE-4 hybrid β-lactamases

Table 6-1: Overview of the 20 protein test set
Table 6-2: Results of side chain placement calculations
Table 6-3: Times for side chain placement calculations
Table 6-4: DEE explosion behavior for protein sequence designs
Table 6-5: Core results for sequence design calculations
Table 6-6: Boundary results for sequence design calculations
Table 6-7: Surface results for sequence design calculations

List of Figures

Figure 1-1: Optimizing the directed evolution algorithm
Figure 1-2: Transitions in word space
Figure 1-3: A cartoon of the fitness landscape
Figure 1-4: Searching smooth and rugged landscapes
Figure 1-5: Mutational constraints in protein structures
Figure 1-6: Constraints lead to a decrease in entropy
Figure 1-7: Mutational capacity of different words
Figure 1-8: A schematic of protein design tools
Figure 1-9: Recombination accelerates word construction

Figure 2-1: Optimal mutation rates based on evolutionary parameters
Figure 2-2: Beneficial mutations are biased towards uncoupled residues
Figure 2-3: A transition for discovering compensating mutations
Figure 2-4: Modeling the mutant distribution
Figure 2-5: The mean and standard deviation of the mutant distribution

Figure 3-1: Sequence entropy versus stabilization energy
Figure 3-2: Sequence entropy freezes at high fitness
Figure 3-3: Sequence entropy and solvent accessibility of subtilisin E
Figure 3-4: Entropy mapped onto the structures of subtilisin E and T4 lysozyme
Figure 3-5: Beneficial mutations occur at high sequence entropy residues
Figure 3-6: Activity improvement versus sequence entropy for T4 lysozyme
Figure 3-7: Average sequence entropy versus generation for antibody 4-4-20
Figure 3-8: Comparison of entropy, solvent accessibility, and natural diversity
Figure 3-9: Sequence entropy versus functional tolerance
Figure 3-10: Mutated high entropy residues of β-lactamase

Figure 4-1: Structure of antibody D1.3 bound to HEL
Figure 4-2: Full sequence design of CDR residues
Figure 4-3: Partial sequence design of targeted residues
Figure 4-4: The sequence entropies of the CDR residues
Figure 4-5: The CDR entropies mapped onto the D1.3 structure
Figure 4-6: The entropy versus distance from binding interface
Figure 4-7: The residue entropies with and without solvation

Figure 5-1: An illustration of schema disruption
Figure 5-2: Calculating the probabilities for broken interactions
Figure 5-3: Comparison of schema disruption profiles with recombination data
Figure 5-4: Single-crossover disruption profile for GART
Figure 5-5: Schema disruption profile for TEM-1/PSE-4 β-lactamase
Figure 5-6: Schema mapped onto the β -lactamase structure
Figure 5-7: Interactions between schema
Figure 5-8: Designed TEM-1/PSE-4 hybrids
Figure 5-9: A transition in activity based on disruption

Figure 5-10: The parameter sensitivity of the schema disruption profile
Figure 5-11: The effect of parental identity on the schema disruption profile
Figure 5-12: The schema disruption profile and other domain algorithms

Figure 6-1: The fraction of incorrect rotamers found by various search algorithms
Figure 6-2: The energy difference between the GMEC and other solutions
Figure 6-3: Convergence time for DEE versus number of design positions
Figure 6-4: MCQ and SCMF results for core, boundary, and surface designs

Figure 7-1: Robustness and evolvability in parameter space
Figure 7-2: Robust and fragile architectures of nodes and edges
Figure 7-3: A comparison of knock out graphs for protein libraries
Figure 7-4: The segment polarization network of *Drosphila*
Figure 7-5: The range of carotenoids produced by directed evolution

Figure B-1: A schematic of the DEE-entropy algorithm
Figure B-2: Comparing mean-field and DEE calculated sequence entropies
Figure B-3: An example of Monte Carlo output
Figure B-4: Comparing mean-field and Monte Carlo calculated sequence entropies

Figure D-1: A cartoon of the mapping between sequence and structure space
Figure D-2: The joint sequence entropies for protein G and engrailed
Figure D-3: Common allowed amino acids for protein G and engrailed

Figure E-1: Designed TEM-1/PSE-4 libraries with targeted crossovers
Figure E-2: Fraction of low-disruption hybrids in targeted libraries
Figure E-3: Low-disruption hybrids from the MIN and MAX libraries
Figure E-4: Low-disruption hybrids from the MIN-MAX library
Figure E-5: Experimental results for the MIN-MAX library
Figure E-6: Properties of random seven-crossover libraries
Figure E-7: Comparison of random libraries and the schema profile

Figure F-1: β-lactamase schema without the probability matrix
Figure F-2: Schema abundance in protein structure database
Figure F-3: Non-homologous structure with β -lactamase schema
Figure F-4: Structural alignments of schema 1+2
Figure F-5: Sequence comparison of TEM-1 and MADS box

Chapter 1

Introduction to Directed Evolution Theory

Enzymes can catalyze a wide range of difficult reactions with high specificity in mild conditions. Despite these advantages, it is difficult to coax enzymes to work on an industrial scale (Thayer, 2001). Enzymes may have low activity towards desired, non-natural reactions and they are often destabilized by reactor conditions, such as high temperatures and organic solvents. It is desirable to optimize enzymes to reduce their disadvantages while retaining their beneficial properties. However, the development of reliable enzyme-modification methods is limited by multiple competing constraints in proteins. For example, a change that improves activity may reduce some other desired property, such as the stability or expression yield. An approach to this problem is to reproduce evolution *in vitro*, through iterative rounds of mutation and selection (Figure 1-1). Within this approach, there is potential for optimization based on principles gleaned from statistical mechanics, computer science, and protein design. The focus of this thesis is the development of computational techniques to model and accelerate the directed evolution of enzymes.

An enzyme is a catalytic protein, which is a linear polymer of amino acids that folds into a well-defined three-dimensional structure. At each monomeric unit, or residue, one of twenty possible amino acids can exist, where the amino acid identities differ in size, polarity, charge, and mobility. The amino acid sequence encodes the ability to fold into a three-dimensional structure and perform some biological function. Changes in the amino acid sequence can alter the thermodynamic and catalytic properties of an enzyme.

Those properties that incur survival in evolution are collectively referred to as the protein's fitness.

During evolution, nucleotide mutations are made spontaneously in the DNA of genes, which translate into amino acid substitutions in proteins. Those mutations that are either neutral or lead to an increase in fitness survive, whereas those mutants that have decreased fitness die. This process can be visualized as a random walk through sequence space – the hyper-dimensional set of all possible amino acid combinations, connected via single amino acid substitutions. A useful analogy in understanding sequence space is the concept of a word space (Figure 1-2) (Smith, 1970). In this space, words are connected by single letter substitutions and movements can be made if the substitution results in an English word. Similarly, mutations can cause drift in sequence space along paths where the intermediate sequences are adequately fit to survive selection.

In sequence space, each amino acid combination has an associated fitness. This additional dimension produces a characteristic fitness landscape (Figure 1-3). The topology of the fitness landscape affects the success of evolution as an optimization algorithm (Kauffman and Levin, 1987; Kauffman, 1993). If the space is very smooth with a single optimum, then any starting sequence can find the global optimum by mutating the sequence and accepting those mutations that increase the fitness (Figure 1-4). However, if the space is very rugged, with many local maxima, then it is far more difficult to optimize the sequence (Derrida, 1981; Macken and Perelson, 1989). An algorithm that accepts all randomly generated uphill steps is unlikely to discover the global optimum.

In proteins, rugged landscapes arise from competing interactions between residues (Figure 1-5). For example, if the side-chains of two amino acids interact, then mutating either residue first may lead to a disruption of the interaction and a decrease in fitness (Baase *et al.*, 1999). However, if both residues are mutated simultaneously, then this has the potential of replacing the interaction entirely, thus improving the likelihood of increasing the fitness. Another way to describe this scenario is to note that the sum of the individual changes in fitness is not equal to the change in fitness of the two mutations made simultaneously. This effect is referred to as non-additivity and it reflects the fact that the residues are interacting. This interaction leads to ruggedness in the fitness landscape and makes the search problem more difficult for evolutionary techniques.

Models have been developed in statistical mechanics to describe the effect of competing interactions on the set of energetic states of a system. In particular, spin-glasses, where the interactions between the spin states of atoms contribute to the ground state of the system, have been extensively studied (Sherrington and Kirkpatrick, 1975; Fischer and Hertz, 1991). Because of their ability to capture competing constraints on biomolecules, these simplified models have been used to study the dynamics of evolution (Anderson, 1983; Prügel-Bennet and Shapiro, 1994). In Chapter 2, a spin-glass-like model is introduced to study the effect of inter-residue interactions on the optimal evolutionary parameters.

The introduction of energetic constraints reduces the entropy of a system. For example, consider a closed box of molecules at constant volume and temperature (Figure 1-6) (Hill, 1960). When the molecules can access the entire volume of the box, the total number of states is Ω. This represents the case when the molecules can exist on the left-

and right-hand sides with equal probability. When the molecules are restricted to the left-hand side through the addition of a restraint, this reduces the number of states to Ω_L. The change in entropy ΔS is

$$\Delta S = S_L - S = -k \ln\left(\frac{\Omega}{\Omega_L}\right), \tag{1-1}$$

where k is Boltzman's constant. By dramatically decreasing the number of available states, the introduction of a restraint can significantly reduce the entropy of a system. The addition of a constraint can be thought of as the reduction or removal of a potential barrier or free energy (Hill, 1960).

The reduction of entropy through the introduction of constraints can be demonstrated by returning to the concept of word space (Shannon, 1951; Abramson, 1963). Considering two different starting points -WORD and ALSO- the number of single-mutant neighbors that are also English words can be enumerated (Figure 1-7 A). Due to the spelling rules of English, some of the letters are more easily substituted. Based on the alignment of words, the entropy of each position can be calculated (Figure 1-7 B). In terms of proteins, constraints are imposed at each residue by the particular three-dimensional topology of the backbone and interactions with other amino acids. These constraints restrict the number of amino acids that can be substituted at each position.

While the statistical models have proven useful in understanding the generalized dynamics of evolution, their simplicity impedes their ability to model specific enzymatic systems. To achieve this, more realistic energies have to be calculated based on the interactions between amino acids, a task that is suitable for computational protein design

tools (Figure 1-8). The original goal of protein design is computationally optimize an amino acid sequence to fold into a defined three-dimensional structure (Hellinga and Richards, 1994; Desjarlais and Handel, 1995; Dahiyat and Mayo, 1997b). The first step of protein design is to reduce the conformational complexity of amino acids by assigning a set of discrete rotamers for each amino acid at each residue. Then, all of the interaction energies between all pairs of rotamers are calculated. Finally, a minimization algorithm is used to find the amino acid sequence that has the lowest predicted energy in the folded state of the protein. These techniques are valuable in modeling molecular evolution because they can be used to calculate the constraints between amino acids for a particular three-dimensional structure. This facilitates the prediction of the evolutionary dynamics for experimentally relevant systems, as is done in Chapter 3.

The recombination of several homologous sequences has proven a successful strategy for directed evolution (Stemmer, 1994; Crameri *et al*., 1998). This technique creates a library of hybrid genes where each mutant has inherited portions of their genetic material from different parents. The power of recombination can be demonstrated by considering the construction of a library of sentences (Figure 1-9). Recombination promotes word swapping whereas mutations alone are more likely to destroy the integrity of a word. The utility of recombination as a search technique has been studied extensively in computer science (Holland, 1975; Mitchell, 1992). An interesting result from these studies is that crossovers are not universally advantageous for all search problems (Schaffer and Eschelman, 1991; Spears and De Jong, 1991; Mühlenbein, 1992). In fact, they have often been shown to hinder the search (Table 1-1) (Mitchell *et al*., 1994). The success of recombination is related to the topology of the fitness landscape

(Manderick *et al*., 1991; Kauffman, 1993; Hordijk and Manderick, 1995; Voigt *et al*., 2001). When crossovers do not divide the constraints of a system, then recombination is more likely to be successful. In the case of a sentence, this means that crossovers should occur between words. In protein evolution, crossovers should not divide interactions between amino acids. Methods to identify regions of protein structures where crossovers are likely to disturb interacting residues are presented in Chapter 5.

There are several future goals in expanding the development of computational tools for directed evolution. Foremost, the targeting strategies proposed here have to be assessed for their ability to accelerate the discovery of novel enzymatic properties. In constructing enriched libraries, two conditions need to be optimized. First, the destabilizing effects of mutations and crossovers have to be minimized. Second, the diversity of the library should be maximized. Balancing these two factors is nontrivial and could be aided by techniques borrowed from multi-objective optimization (Loughlin and Ranjithan, 1997). A second goal is to extend directed evolution theory to model and optimize the evolution of genetic circuits and metabolic networks (Cremeri *et al*., 1997; Schmidt-Dannert *et al*., 2000). Models of biochemical networks can be used to identify those components that are most likely to generate diverse network functions when mutagenized. By expanding the theory to explore the evolution of different hierarchies in biology, it may be possible to develop an understanding of the organizational strategies that lead to robust and evolvable systems.

Table 1-1. Discovery Times for Evolutionary Algorithms

Algorithm[a]	Time[b]
Steepest ascent hill-climbing[c]	>265,000
Random mutation hill-climbing[d]	6179
Genetic algorithm[e]	61334

a. Algorithms and data taken from Mitchell (1992).
b. The mean number of function evaluations required to find the optimal string for a "royal road" fitness landscape, averaged over 200 landscapes.
c. An initial string is chosen at random. All single mutations are attempted and the most fit mutant becomes the next parent. This process is repeated until no more fit mutations are found.
d. An initial string is chosen at random. Mutations are randomly attempted and the first mutant that has an increased fitness becomes the next parent. This process is repeated until convergence is achieved.
e. Two initial strings are chosen at random and a population of offspring is produced with crossover rate p_c and mutation rate p_m. Selection pressure is then applied to the population. Rounds of crossover, mutation, and selection are repeated until the population converges.

Figure 1-1:

A schematic is shown of the directed evolution algorithm with potential areas for optimization marked in red. The first step of directed evolution is to isolate the DNA sequence that encodes the wild-type protein. Through PCR techniques, a library of mutant or recombinant DNA sequences is produced. Each member of this library has a sequence that is slightly perturbed from the parent. This library is then screened for those properties in which the researcher is interested. The mutant with the best combination of properties (highest fitness) becomes the parent to the next round of mutation and selection. Within the directed evolution algorithm, there is much potential for optimization. The evolutionary parameters are interdependent, for example, the optimal mutation rate depends on the screening capacity. Further, after each generation, the list of mutant fitnesses contains information about the local search space. Rather than discarding this information, it is desirable to use it to optimize the evolutionary parameters for the next generation. Finally, information, such as the three-dimensional structure of the protein or alignments of naturally occurring sequences, is being rapidly accumulated for many enzymatic systems. This information has the potential to optimize the evolutionary search.

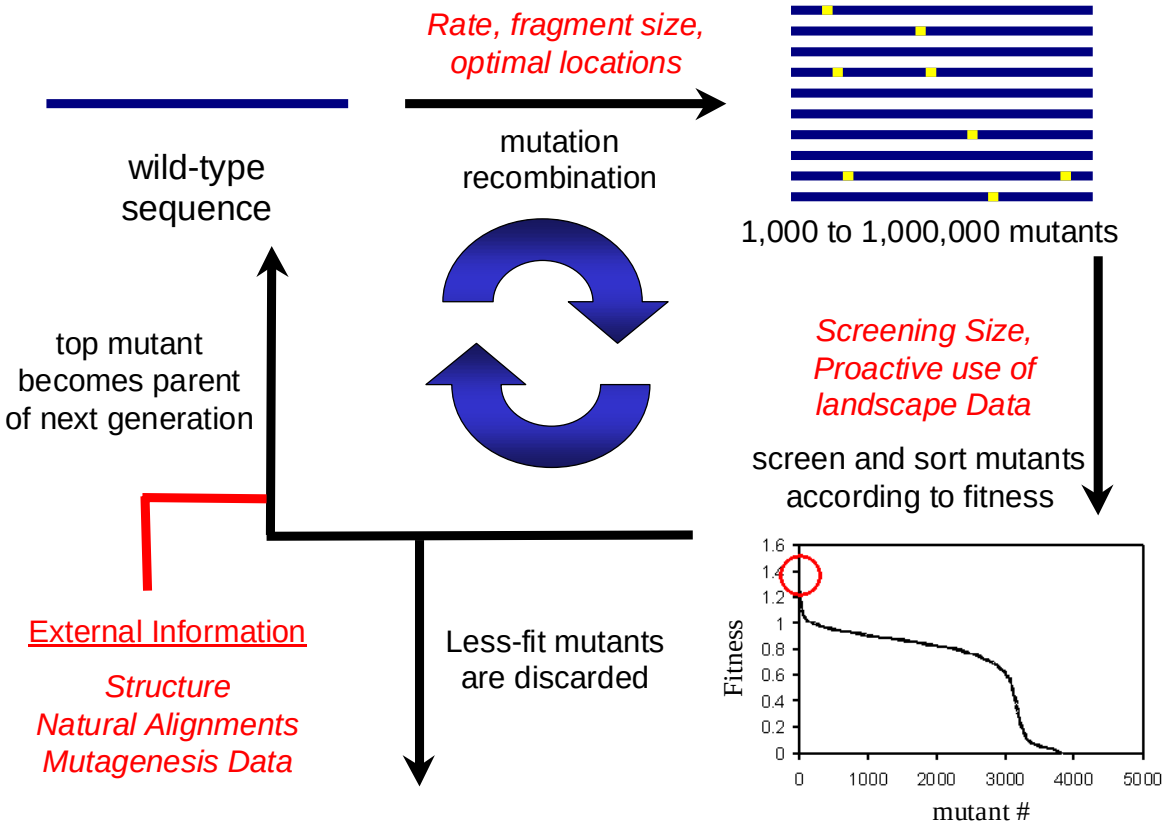

Rate, fragment size,
optimal locations
wild-type
sequence
mutation
recombination
1,000 to 1,000,000 mutants
top mutant
becomes parent
of next generation
Screening Size,
Proactive use of
landscape Data
screen and sort mutants
according to fitness
External Information
Structure
Natural Alignments
Mutagenesis Data
Less-fit mutants
are discarded
Fitness
mutant #
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
0
1000
2000
3000
4000
5000

Figure 1-2:

Transitions between words are similar to movements through sequence space (Smith, 1970). In this example, WORD is transformed into GENE via two paths consisting of single letter substitutions. A requirement is that each intermediate set of four letters composes an English word. Two independent paths are shown. The probability that a transition will occur between two sequences during evolution is related to the number of connected paths between the sequences.

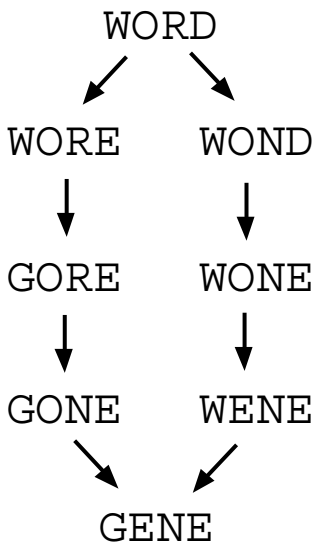

WORD
WORE
WOND
GORE
WONE
GONE
WENE
GENE

Figure 1-3:

A two-dimensional projection of the hyper-dimensional fitness landscape is shown. In this simplified representation, for a four-residue sequence is considered where the colors represent amino acid identities. The all-blue sequence is the global optimum whereas the lower fitness peaks are local optima. The problem of *in vitro* evolution is how to search this space effectively, without becoming trapped at a sub-optimal fitness.

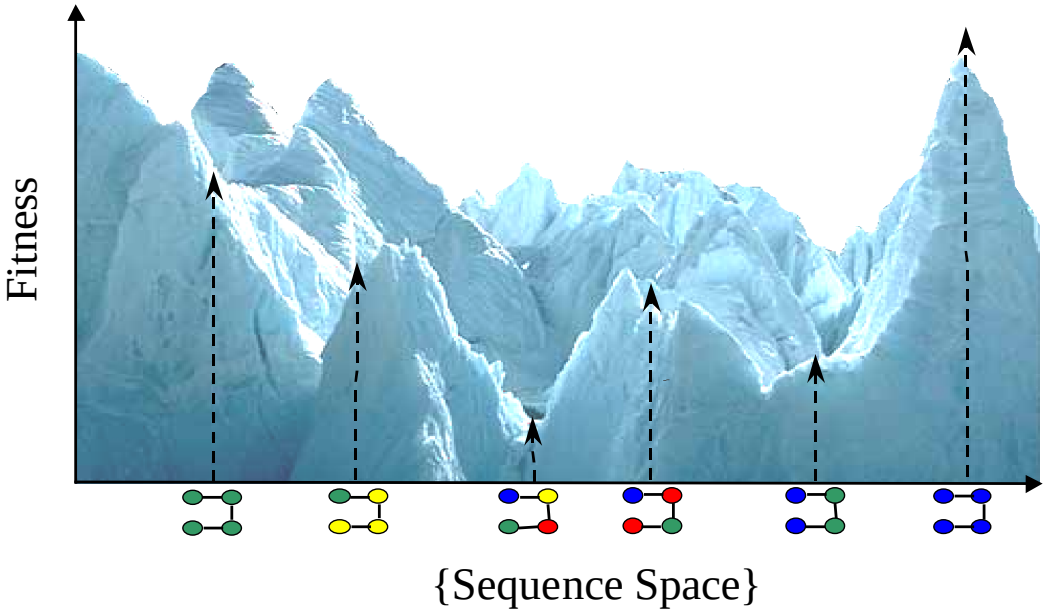
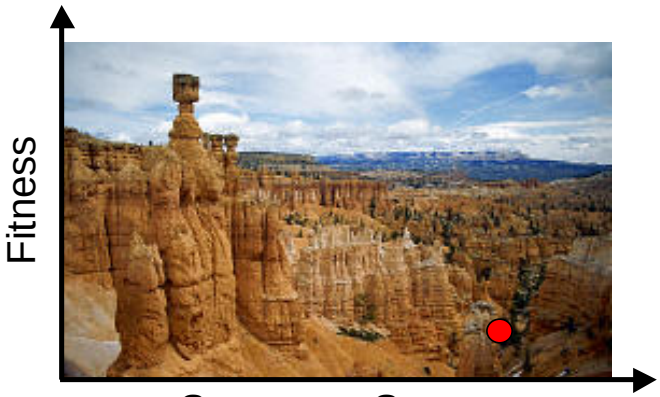

Fitness
{Sequence Space}

Figure 1-4:

An example is shown of a smooth **(A)** and rugged **(B)** fitness landscape. Rugged landscapes are characterized by multiple maxima, which act as traps for an evolving sequence. Conversely, a smooth landscape contains fewer such peaks. Hypothetical random starting points are marked by the red dots. From any starting point, the smooth landscape is easy to climb. Any algorithm of mutagenesis of selection is guaranteed to discover the global optimum. However, it is more difficult to optimize a sequence on a rugged landscape, as a steepest-ascent mutagenesis algorithm is likely to converge onto a sub-optimal peak.

(A)



(B)

Figure 1-5:

A simple example is shown that demonstrates constraints in the protein structure. In this example, a particular set of two mutations leads to an increase in fitness. However, if either mutation is made individually, this leads to a decrease in fitness (here, due to the over- or under-packing of atoms). Another way to describe this scenario is to note that the sum of the individual changes in fitness is not equal to the change in fitness of the two mutations made simultaneously.

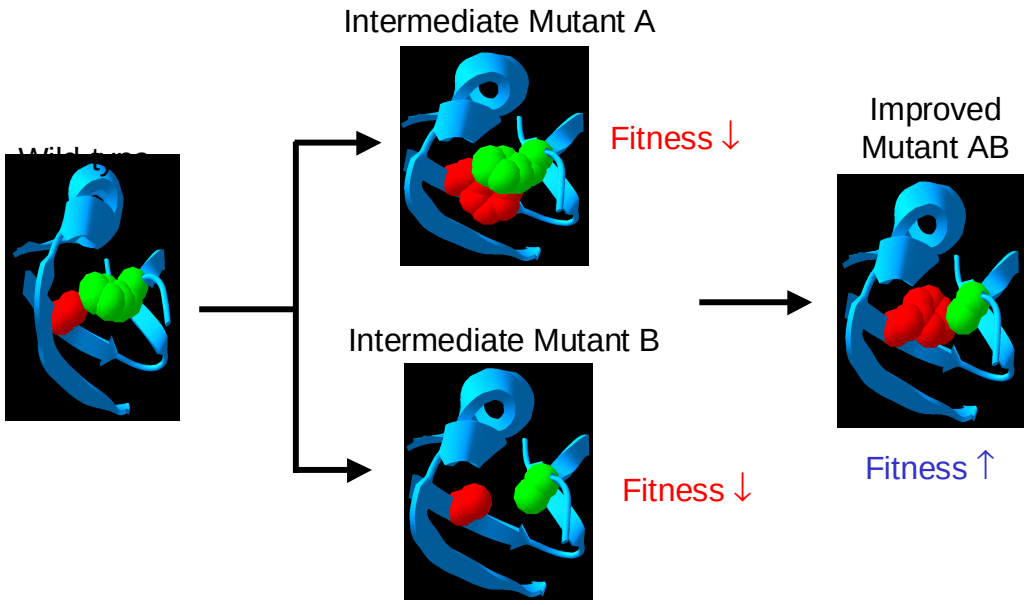

Intermediate Mutant A
Fitness ↓
Improved
Mutant AB
Intermediate Mutant B
Fitness ↓
Fitness ↑

Figure 1-6:

The box shown contains ten freely diffusing molecules. In this simplification, each molecule can exist in either one of two states: the left or right of the box (demarcated by the dashed gray line). When the molecules are allowed to diffuse freely, the probability that each molecule will exist on either side is 0.5. When a barrier is imposed on the system, the probabilities change to either 1.0 or 0.0 and the entropy of the system decreases. This example was inspired by a presentation by Jeffrey Saven (University of Pennsylvania).

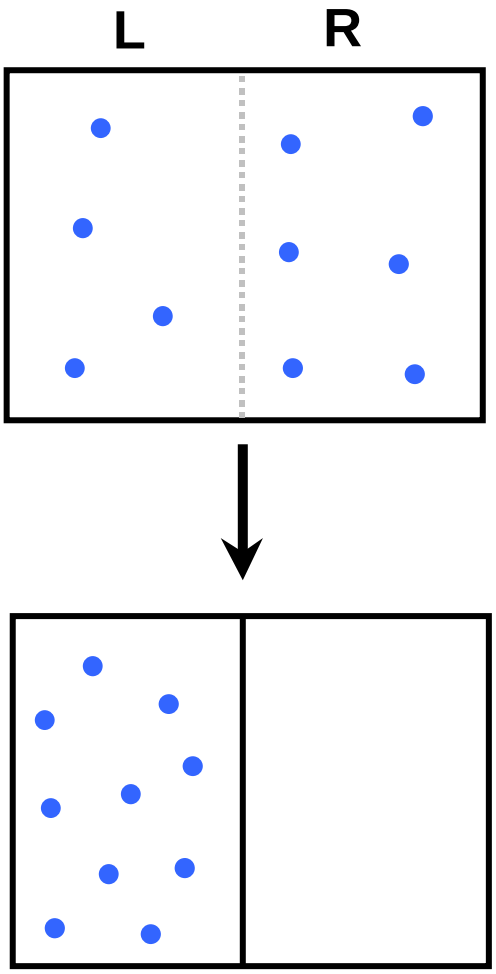

L
R

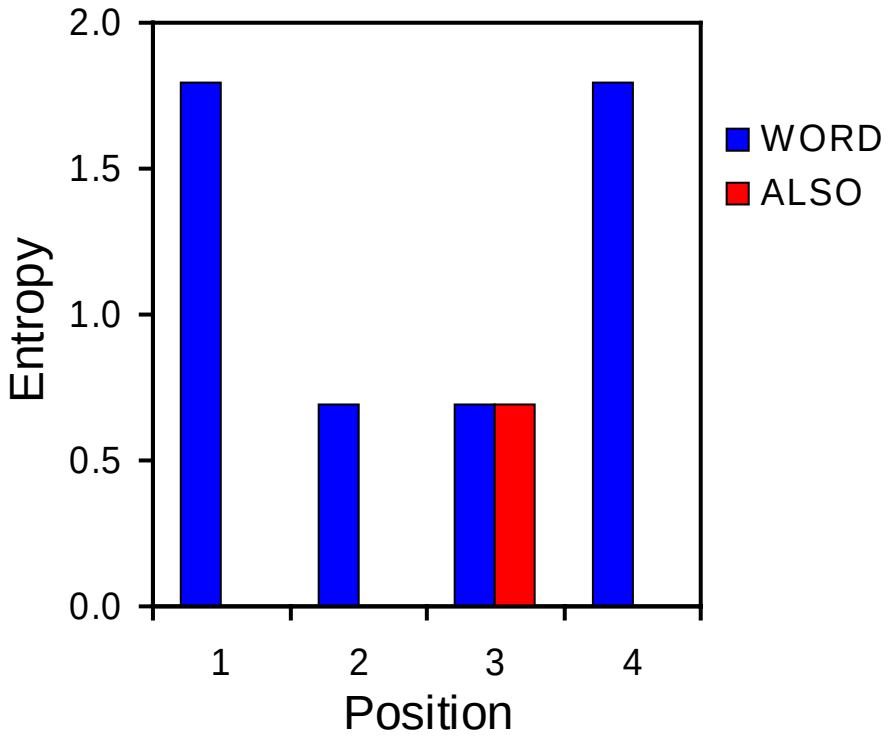
Figure 1-7:

(A) The sets of four letters attainable with single substitutions from WORD and ALSO. Those sets that correspond to English words are marked in red. The first and last letters in WORD are variable whereas the middle letters and the letters of ALSO are less mutable. **(B)** Based on the alignment in (A), the entropy of each position can be calculated. Low entropies indicate intolerance to substitutions due to constraints. Here, the constraints are the spelling rules of English. In protein structures, the constraints are interactions between residues.

WORD	WORD	WORD	WORD
AORD	WARD	WOAD	WORA
BORD	WBRD	WOBD	WORB
CORD	WCRD	WOCD	WORC
DORD	WDRD	WODD	WORE
EORD	WERD	WOED	WORF
FORD	WFRD	WOFD	WORG
GORD	WGRD	WOGD	WORH
HORD	WHRD	WOHD	WORI
IORD	WIRD	WOID	WORJ
JORD	WJRD	WOJD	WORK
KORD	WKRD	WOKD	WORL
LORD	WLRD	WOLD	WORM
MORD	WMRD	WOMD	WORN
NORD	WNRD	WOND	WORO
OORD	WPRD	WOOD	WORP
PORD	WQRD	WOPD	WORQ
QORD	WRRD	WOQD	WORR
RORD	WSRD	WOSD	WORS
SORD	WTRD	WOTD	WORT
TORD	WURD	WOUD	WORU
UORD	WVRD	WOVD	WORV
VORD	WWRD	WOWD	WORW
XORD	WXRD	WOXD	WORX
YORD	WYRD	WOYD	WORY
ZORD	WZRD	WOZD	WORZ
5	1	1	5

ALSO	ALSO	ALSO	ALSO
BLSO	AASO	ALAO	ALSA
CLSO	ABSO	ALBO	ALSB
DLSO	ACSO	ALCO	ALSC
ELSO	ADSO	ALDO	ALSD
FLSO	AESO	ALEO	ALSE
GLSO	AFSO	ALFO	ALSF
HLSO	AGSO	ALGO	ALSG
ILSO	AHSO	ALHO	ALSH
JLSO	AISO	ALIO	ALSI
KLSO	AJSO	ALJO	ALSJ
LLSO	AKSO	ALKO	ALSK
MLSO	AMSO	ALLO	ALSL
NLSO	ANSO	ALMO	ALSM
OLSO	AOSO	ALNO	ALSN
PLSO	APSO	ALOO	ALSP
QLSO	AQSO	ALPO	ALSQ
RLSO	ARSO	ALQO	ALSR
SLSO	ASSO	ALRO	ALSS
TLSO	ATSO	ALTO	ALST
ULSO	AUSO	ALUO	ALSU
VLSO	AVSO	ALVO	ALSV
WLSO	AWSO	ALWO	ALSW
XLSO	AXSO	ALXO	ALSX
YLSO	AYSO	ALYO	ALSY
ZLSO	AZSO	ALZO	ALSZ
0	0	1	0

(A)



(B)

Figure 1-8:

The typical set of computational tools for protein design is shown schematically. First, the three-dimensional structure is retrieved and the side chains are stripped, leaving a fixed backbone structure. Then, each residue is classified as existing in the core, boundary, or surface of the protein. Only hydrophobic amino acids are allowed in the core, hydrophilic amino acids at the surface, and all amino acids in the boundary. The flexibility of the amino acid side chains is captured using a set of discrete conformational rotamers. Next, all of the side-chain/side-chain and side-chain/backbone energies are calculated using a force field that includes terms for solvation, H-bonding, electrostatics, and van der Waals interactions. Finally, the optimal set of rotamers is obtained using a search algorithm, such as dead-end elimination or Monte Carlo simulated annealing (see Chapter 6).

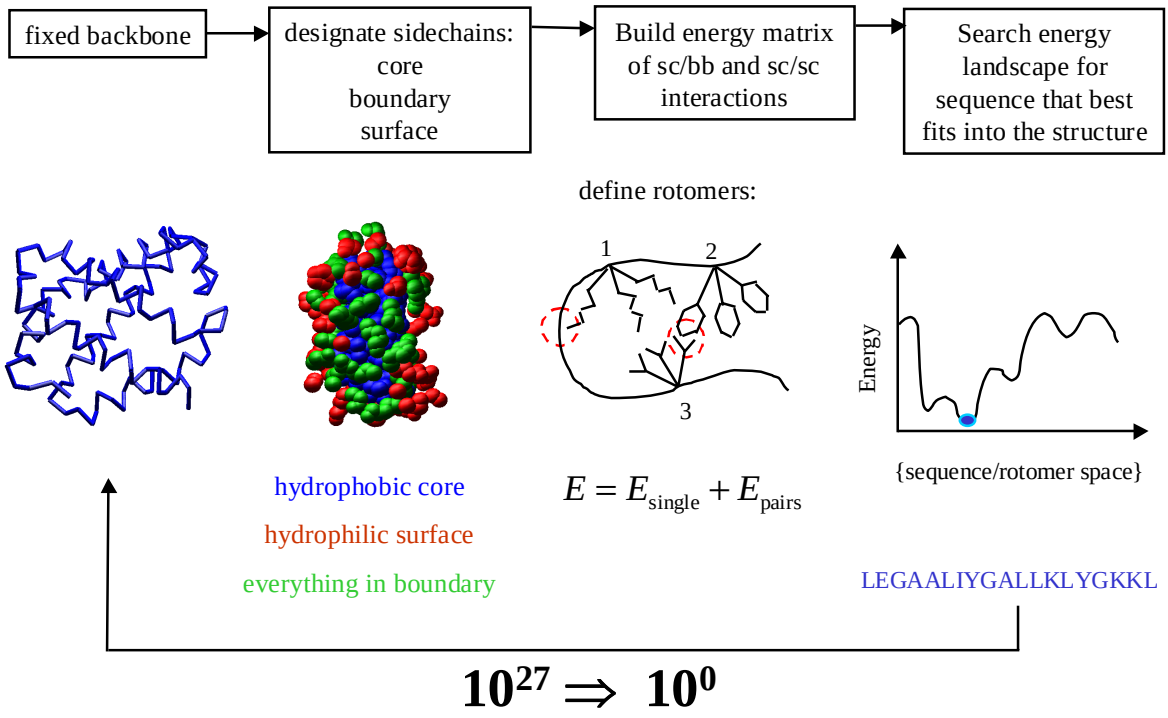

fixed backbone
designate sidechains:
core
boundary
surface
Build energy matrix
of sc/bb and sc/sc
interactions
Search energy
landscape for
sequence that best
fits into the structure
define rotomers:
1
2
3
Energy
{sequence/rotomer space}
hydrophobic core
hydrophilic surface
everything in boundary
$E = E_{\text{single}} + E_{\text{pairs}}$
LEGAALIYGALLKLYGKKL
$10^{27} \Rightarrow 10^{0}$

Figure 1-9:

(A) From a partial starting sentence, point mutations are unlikely to discover new sentences because the number of simultaneous mutations required is too large to be sampled in a reasonable amount of time. The vast majority of single- or multiple-substitution sentences will be nonsensical. **(B)** If recombination is allowed to swap the words from two sentences, then it is more likely to create a library of potentially new sentences. However, if recombination is allowed to divide the words, the library will be significantly less viable (Table 1-1).

THE HEAD AND IN A FRONTAL ATTACK

THE HRAD AND IN A FRONTAL ATTACK
THE HEAD AND IN A FOONTAL ATTACK
THE HEAD ANT IN A FRONTAL AUTACK
THE HEAD AND IN T FRONTAL ATTACK
THEXHEAD AND IN A FRONTAL ATTACK
THE HEMD AND IN A FRONTAL ATTACK
THE HEAD AND IQ A FRONTAL ATTACK
THE HEAD AND IN A FRONTPL ATTACK
THE HEAD AND IN I FRONTAL ATTACK
THE HRAD AND IN A FRONTAL ATTACK

(A)

THE HEAD AND IN A FRONTAL ATTACK
THIS POINT IS THEREFORE METHOD

THIS POINT IN A FRONTAL ATTACK
THE HEAD AND THEREFORE METHOD
THIS HEAD AND IN A METHOD ATTACK
THIS POINT IN A FRONTAL METHOD
THE POINT IS THEREFORE ATTACK
THIS POINT AND IN A FRONTAL ATTACK
THE HEAD POINT IS THEREFORE
THIS HEAD IS THEREFORE ATTACK
THE HEAD AND IS THEREFORE METHOD
THIS POINT IN A FRONTAL METHOD

(B)

Chapter 2

Modeling the Dynamics of Directed Evolution

Portions of this chapter are reproduced from:
Voigt, C. A., Mayo, S. L., Arnold, F. H., and Wang, Z-G. (2001). Computational method to reduce the search space for directed protein evolut
Voigt, C. A., Mayo, S. L., ion. Proc. Natl. Acad. Sci. USA 98, 3778-3783Arnold, F. H., and Wang, Z-G. (2001). Computationally focusing the directed evolution of proteins. J. Cell. Biol. 37, 58-63.
Voigt, C. A., Kauffman, S., and Wang, Z-G. (2001). Rational evolutionary design: The theory of in vitro protein evolution, Adv. Prot. Chem. 55, 79-159.

Abstract

Several models are introduced to study the directed evolution algorithm as a search technique. First, a spin-glass-like energy function is developed to capture the statistical features of fitness landscapes and is used to study the effect of a finite screening capacity on the optimal mutation rate. We demonstrate that the optimal mutation rate is low when the screening capacity is small, the parent sequence is highly fit, or there are many interacting residues. Further, when the mutation rate and the screening capacity are limited, the beneficial mutations discovered by directed evolution tend to be at uncoupled, or non-interacting, residues. Using a probabilistic model of the mutant library, a transition in the dynamics of directed evolution is shown where the benefit from simultaneously mutating coupled residues becomes significant in large libraries. Finally, we use mean-field theory to study the effect of the mutation rate on the moments of the mutant fitness distribution.

1. Introduction

A key constraint in directed evolution is the limited screening capacity. Typically, screening is limited to 10^3 to 10^6 mutants (Giver *et al*., 1998; Petrounia and Arnold, 2000; Daugherty *et al*., 2000). The state-of-the-art high throughput selection techniques, such as RNA-protein fusion, can only handle on the order of 10^{12} mutants (Roberts and Szostak, 1997). Despite impressive experimental advances, the sampling ability remains tiny when compared with the vastness of sequence space. To reduce the project time and cost of an experiment, it is desirable to optimize the search parameters, such that the maximum fitness improvements can be found with the minimum screening effort. Towards this goal, this chapter is devoted to a model describing the properties of small libraries of mutants, such as those generated by error-prone PCR.

Simulations using a statistical model of the fitness landscape demonstrate the relationship between the screening capacity, the parent fitness, the landscape ruggedness, and the optimal mutation rate. Further, we demonstrate that when the screening capacity and mutation rate are small, directed evolution tends to discover beneficial mutations at uncoupled residues. In addition, the mutant data collected from the screen contain information about the structure of the local search space. To extract this information, we first use a probabilistic model to analyze the transition at which beneficial coupled mutations dominate the mutant library. Next, a mean-field solution to the model is derived to study the statistics of the mutant distribution.

2. Modeling Directed Evolution

2.1. The Search Space

Our strategy for simulating the evolutionary dynamics is to start with a statistical description of the fitness landscape. The directed evolution algorithm is then tested on an ensemble of landscapes and the relationship between evolutionary parameters is observed. Ruggedness, caused by interacting residues, is the dominant feature of the fitness landscape that determines the success of an evolutionary search. There has been extensive research in statistical physics to quantify the relationship between coupling and the ruggedness of energy landscapes (i.e., frustration) (Sherrington and Kirkpatrick, 1975; Fischer and Hertz, 1991). Spin glasses are simple models of frustration and have been frequently used to model evolution (Anderson, 1983; Bryngelson and Wolynes, 1987; Prügel-Bennett and Shapiro, 1994).

Husimi and Aita used an uncoupled (fully additive) fitness landscape to compare the effectiveness of several evolutionary search strategies (Aita and Husimi, 1996; Aita and Husimi, 1998). This model can be expressed as the fitness function,

$$F = \sum_{i}^{N} \gamma(i_a), \tag{2-1}$$

where N is the number of residues and $\gamma(i_a)$ is the individual contribution of amino acid a at residue i to the total fitness of the sequence, F. The uncoupled fitness function corresponds to a fitness landscape with a single optimum, which is easily found by mutation and selection.

Mutational effects often appear remarkably additive in amino acid substitution experiments (Wells, 1990; Matsuura *et al.*, 1998; Brown and Sauer, 1999). The observed additivity of mutations is partially determined by the mutational distance from the wild-

type sequence. If only a few mutations are made on a large protein, then their effects could appear additive if the regions perturbed by each mutation do not overlap (Shoichet *et al*., 1995). As mutations are accumulated, it becomes more likely that non-additivity will be observed. This suggests that an additive fitness function may adequately describe the behavior of evolution up to some mutational distance from the wild-type sequence. Based on this argument, the fitness function can be written as an expansion

$$F = \sum_{i}^{N} \gamma(i_a) + \frac{1}{2} \sum_{i}^{N} \sum_{j \neq i}^{N} \gamma(i_a, j_b) + (3 - \text{body terms}) + \ldots \quad , \tag{2-2}$$

where the higher-order terms become increasingly important as the mutational distance from wild-type gets larger. Directed evolution generally makes on the order of 10 amino acid mutations on a 300–500 residue protein, indicating that the length of the walk is small with respect to the total number of residues (Arnold and Wintrode, 1999). However, some non-additive effects have been observed frequently in directed evolution experiments and are important in modeling the process (Moore and Arnold, 1996; Moore *et al*., 1997; Spiller *et al*., 1999).

Two-body coupling interactions have been added to model thermostability (Shakhnovich, 1994; Li *et al*., 1996; Dahiyat and Mayo, 1997; Saven and Wolynes, 1997) and catalytic activity (Matsuura *et al*., 1998). Equation (2-2) can be truncated to account for only one- and two-body terms

$$F = \sum_{i}^{N} \gamma(i_a) + \frac{b}{2} \sum_{i}^{N} \sum_{j \neq i}^{N} \gamma(i_a, j_b) \lambda_{ij} \ , \tag{2-3}$$

where b determines the relative strength of coupled versus uncoupled interactions and $\lambda_{ij} = 1$ if residues i and j are coupled and 0 if not. The form of Equation (2-3) is similar to an energy expression commonly used in protein design (Dahiyat and Mayo, 1997) and

has been used by Wolynes to study combinatorial libraries (Saven and Wolynes, 1997). The number of non-zero terms in λ is given by a model parameter τ, which determines the degree of coupling between residues and is therefore a measure of landscape ruggedness. The interactions are symmetric: if residue *i* interacts with residue *j*, then residue *j* interacts with residue *i*.

The property of smoothness (weak coupling) can be viewed as the tolerance of sequence positions for amino acid substitutions (Reidhaar-Olson and Sauer, 1988; Reidhaar-Olson and Sauer, 1990; Saven and Wolynes, 1997). In our model, tolerance arises out of two effects: (1) the fitness distribution of a site γ, and (2) coupled interactions. When the distribution is skewed towards low fitness, the position will be intolerant. A model where each residue has a different standard deviation in the distribution of fitnesses was used by Husimi and Aita to model the effects of tolerance on evolution (Aita and Husimi, 1996; Aita and Husimi, 1998). Tolerance is also related to the number of interactions in which a residue participates. Residues that are weakly coupled tend also to be tolerant, such as residues that lie on the surface (Reidhaar-Olson and Sauer, 1988; Saven and Wolynes, 1997, Brown and Sauer, 1999). The parameter *b* in Equation (2-3) can be viewed as determining the origin for tolerance. If *b* is small, effect (1) dominates whereas if *b* is large, effect (2) dominates.

At the beginning of the simulation, the fitness landscape is generated by randomly placing the τ interactions between *N* residues and randomly assigning the one-body $\gamma(i_a)$ and two-body $\gamma(i_a,j_b)$ fitness contributions from a Gaussian distribution. Both the placement of the interactions and their strengths remain quenched after the landscape has been initialized. The directed evolution algorithm of mutagenesis and screening is then

simulated from starting sequences with different fitnesses. These simulations are used to observe the effect of various evolutionary parameters (*i.e.*, mutation rate) on the properties of the mutant library.

In the simulations, mutations are made on the DNA level and transcribed to amino acid sequences through a representation of the genetic code. As a result of the special connectivity and degeneracy of the triplet code, some amino acid substitutions are impossible via a single nucleotide mutation. The gene on which random mutagenesis is performed is large, making it unlikely that two adjacent DNA mutations will occur in a single round of error-prone PCR. This reduces the number of possible paths in sequence space from $20N$ to about $5.7N$. Because the number of available single-point mutations decreases, it decreases the number of fitter sequences in the mutant library at each step.

A higher mutation rate increases the probability that a nucleotide substitution will lead to the creation of a stop codon. The presence of stop codons causes an acceleration in the generation of inactive mutants as the mutation rate increases, thus reducing the effective size of the library. The fraction of screened mutants that are dead due to a mutation to a stop codon is described by the binomial distribution

$$f_{stop} = \sum_{n=1}^{N} \frac{N!}{n!(N-n)!} q_{a\to stop}{}^{n} (1-q_{a\to stop})^{N-n} = 1-(1-q_{a\to stop})^{N}, \qquad (2\text{-}4)$$

where N is the number of codons and $q_{a\to stop}$ is the probability that a mutation will cause a transition to a stop codon. The average probability of appearance of a stop codon is $q_{a\to stop} = (3/63)Q$, where $Q = 1 - (1-p_m)^3$ is the probability that a codon is mutated, given the per-nucleotide mutation probability p_m. When an average of 5 DNA mutations are made on a 1000 nucleotide gene then $f_{stop} = 0.21$ and when the average is 20 then

f_{stop} = 0.61, indicating that even a moderate mutation rate can cause a significant fraction of mutant sequences to contain stop codons.

2.2. Optimal Search Parameters and Finite Screening Capacity

The model described by Equation (2-3) was used to investigate the effect of the finite screening capacity on libraries generated by mutagenesis (Voigt *et al*., 2001a; Voigt *et al*., 2001c). For a given screening capacity, there is an optimal mutation rate, defined as the rate that produces the largest fitness improvement for a given library size. This is a consequence of two opposing effects. On the one hand, a large enough mutation rate is required to generate adequate diversity in the mutants. On the other hand, because the probability of an individual mutation demonstrating improvement is small, multiple mutations on the same sequence (the result of large mutation rate) are generally deleterious. In a limited screening pool, the probability of observing improvement thus decreases rapidly as the number of mutations increases.

The optimal mutation rate is typically low (about one amino acid substitution per sequence) because the probability of an individual mutation demonstrating improvement is small (Moore and Arnold, 1996; Moore *et al*., 1997). When multiple mutations are accumulated, it is likely that most are deleterious and these mutations quickly erode the improvement from the few beneficial mutations that may exist. This effect worsens as the number of mutants that can be screened decreases (Figure 2-1A).

As the mutation rate increases, the number of possible combinations increases exponentially. Therefore, to adequately sample higher mutation rates, exponentially larger libraries are required. Similarly, as the fitness of the parent sequence increases, the

probability of improvement decreases, thus exaggerating the effect of deleterious mutations (Figure 2-1C). Thus, as the generations of mutation and selection progress, an exponentially increasing screening size is required (Macken and Perelson, 1989).

The probability of improvement is also affected by the ruggedness of the fitness landscape. As the number of interactions increases, the probability that a mutation is deleterious also increases. When multiple mutations are accumulated on a gene, a larger fraction of these mutations will decrease the fitness. This effect quickly erodes the beneficial effect of any positive mutations. Therefore, when a small library is used to search rugged landscapes, a smaller mutation rate is optimal (Figure 2-1B). If the topology of the protein structure is particularly tolerant to amino acid substitutions, thus creating a smooth fitness landscape, then fewer mutants must be screened in order to achieve the benefit of a higher mutation rate. The ability to absorb mutations without affecting fitness, or neutral evolution, allows sequences to drift through sequence space, improving the likelihood of discovering fitness improvements (Kimura, 1983; Fontana, 1987).

2.3. Beneficial Mutations Occur at Uncoupled Positions

The fitness model is also used to observe where beneficial mutations are found with respect to the protein structure (Voigt *et al*., 2001b). In this model, the structural topology is described by the pattern of τ interactions distributed among the N residues. We find that the probability of a beneficial mutation occurring at a highly coupled residue decreases significantly as the fitness of the parent increases (Figure 2-2). The bias towards mutating uncoupled residues late in evolution is a result of the finite screening

capacity. A highly coupled group of residues requires several simultaneous mutations to demonstrate improvement. When a mutation is made at a coupled residue, it is necessary to improve all the coupled terms in addition to the uncoupled term, the probability of which rapidly decreases as the sequence becomes more highly optimized. This result is independent of the specific form of Equation (2-3) and can be demonstrated using any model that incorporates a variable degree of coupling between residues, such as Kauffman's *NK*-model (Kauffman and Levin, 1987; Kauffman and Weinberger, 1989; Kauffman, 1993), lattice proteins (Shakhnovich, 1994; Li *et al*., 1996), or RNA secondary structure models (Fontana and Shuster, 1998).

2.4. The Probability of Coevolving Residues

In Section 2.3, we demonstrated that at low mutation rates and small library sizes, beneficial mutations will tend to occur at uncoupled residues. This implies a transition in the dynamics of directed evolution based on these evolutionary parameters. At some critical library size and mutation rate, pairs of beneficial coupled mutations will begin to be discovered. In this section, data from a large library of antibody 26-10 mutants is analyzed to demonstrate this transition (Daugherty *et al*., 2000).

A probabilistic model is developed to describe the fraction of the library that retains function P_f as the mutation rate m is increased (Figure 2-3). For low m, P_f decays rapidly, representing an accumulation of deleterious mutations. However, at some critical mutation rate, the behavior changes drastically. At this point, the slope of the decay decreases so that P_f remains relatively unchanged for large mutation rates. As the mutation rate continues to increase, P_f starts to degrade again, albeit much less rapidly

than the initial burst. The goal of our model is to capture the effects that underlie these transition points.

The initial decay represents the accumulation of deleterious single mutations and stop codons. First, using a variant of Equation (2-4), we separate the portion of the library that is free of stop codons P_s,

$$P_s = \left(\frac{60}{63} + \frac{3}{63}\left(1 - \frac{m}{N}\right)^3 \right)^{\frac{N}{3}}, \tag{2-5}$$

where N is the number of nucleotides. Further, we assume that a certain fraction of single-point mutations make the mutants non-functional f_d. The fraction of the mutant library that does not contain one of these lethal mutations P_d is

$$P_d = \left(1 - f_d \frac{m}{N}\right)^N. \tag{2-6}$$

The fit parameter f_d can be obtained easily from the low mutation rate data and is found to be $f_d = 0.7$. In this experiment, survival is determined by using a stringent screen for binding to antigen. The value of f_d may decrease as the definition of functionality is relaxed, or only the effect of the mutation rate on stability is measured.

The critical point occurs because a coupled interaction is improved by a simultaneous double mutation. The improvement initially overwhelms the damage caused by the deleterious single mutations, thus allowing a mutant to remain functional. This causes a decrease in the decay. The simplest way to model this effect is to assume that there are n_c coupled interactions in the protein. If the two coupled residues are mutated simultaneously to the correct amino acid state, then the mutant retains functionality. We

do not consider 3-way and higher-order coupling interactions because the probability that these will occur in a library of 10^5–10^7 mutants is negligible.

Here, we treat the coupled interactions on the nucleotide level. In this model, a single amino acid substitution that requires two simultaneous nucleotide changes is coupled. The simultaneous mutagenesis of coupled residues is modeled by considering the fraction of the library P_c that has at least one out of n_c possible coupling interactions mutated simultaneously to the proper nucleotide

$$P_c = 1 - \left(1 - \left(\frac{1}{3}\right)^2 \left(\frac{m}{N}\right)^2\right)^{n_c} . \tag{2-7}$$

The $(m/N)^2$ term is the probability of simultaneously mutating two positions and the $(1/3)^2$ term is the transition probability that the offspring has the proper two new nucleotides.

Finally, even a coupled interaction that initially provides a large fitness improvement is eventually degraded by the accumulation of too many single mutations. This can be accounted for by introducing a third parameter n_s which is the number of single mutations that will, on average, overcome the beneficial effects from an improved pair of coupled mutations. Assuming a Poisson distribution of mutations, the fraction of the library that has less than n_s mutations P_n is

$$P_n = \sum_{i=1}^{n_s - 1} \frac{e^{-m} m^i}{i!} . \tag{2-8}$$

Note that up to this step the Poisson assumption has not been invoked.

Combining these results, the fraction of functional mutants is given by

$$P_f = P_s(P_d + P_c P_n)\,. \tag{2-9}$$

In this treatment, there are three fit parameters: the fraction of single mutations that lead to loss-of-function f_d, the number of good potential coupling interactions n_c, and the number of mutations required to overcome a good coupled interaction n_s. The parameters f_d = 0.7, n_c = 110, and n_s = 12 fit the data well (Figure 2-3). There are two interesting behaviors in the class of curves generated by this approach. First, the transition point is robust with respect to n_c and n_s. By decreasing f_d, the transition point can be moved towards higher mutation rates. It may be possible to test this prediction by using a less-stringent definition of function (thus lowering f_d). A class of curves based on altering f_d would be very useful in developing a more refined model. Second, the value found for n_s (10–14 mutations) agrees well with the number of mutations found on the improved mutants in the antibody data set (Daugherty *et al.*, 2000).

2.5. Calculating the Moments of the Mutant Library

During the screening experiments, a large amount of fitness data is generated, but only the fitness information of the improved mutants is used to continue to the next round of evolution. The large ensemble of less-fit mutants provides a view of the local fitness landscape. By analyzing these data, certain statistical landscape parameters can be deduced, such as τ and b, which can then be used in conjunction with the simulation results to guide the setting of evolutionary parameters. In this analysis, sequencing is time consuming and expensive, so a sequence cannot be assigned to each measured fitness. The lack of sequencing data means that only the probability distribution of mutant fitnesses can be analyzed (Figure 2-4A). We can analytically model the behavior of the

moments of this distribution (mean and standard deviation) as the sequence ascends the fitness landscape under the influence of different per-nucleotide mutation rates, p_m. In our analysis, only the portion of the mutant distribution that is not dead (zero fitness) or parent (unmutated) is considered, thus removing the discontinuities in the mutant fitness distribution (Figure 2-4B).

To obtain the moments of the mutant fitness distribution, we average the change in fitness from the wild-type to mutant, $w = F_{mut} - F_{wt}$, over all sequences

$$< w >= \sum_{S_A} \sum_{S_B} w(S_A, S_B) P\{S_A, S_B\}, \tag{2-10}$$

where S_A indicates the sum over all wild-type sequences and S_B indicates the sum over all mutant sequences. The probability term can be split into the probability that sequence S_A exists and the transition probability that S_A mutates to S_B

$$P\{S_A, S_B\} = P\{S_A\} P\{S_A \rightarrow S_B\}. \tag{2-11}$$

Mean-field theory can then be invoked to further divide the probability $P\{S_A\}$ into the product of the probabilities $p(i_a)$ of amino acid a existing at residue i multiplied by the additional probability $q_{ia \rightarrow ib}$ that i_a mutates to i_b,

$$P\{S_A, S_B\} = \prod_{i=1}^{N} p(i_a) q_{i_a \rightarrow i_b} . \tag{2-12}$$

Inserting the one-body assumption for the fitness function (Equation 2-1) and Equation (2-12) into the mutant average gives

$$< w >= \sum_{i=1}^{N} \sum_{S_A} \sum_{S_B} [\gamma(i_b) - \gamma(i_a)] \prod_{j=1}^{N} p(j_a) q_{j_a \rightarrow j_b} . \tag{2-13}$$

We can rearrange the probabilities to group the $i = j$ terms,

$$< w >= \sum_{i=1}^{N} \sum_{i_a} \sum_{i_b} [\gamma(i_b) - \gamma(i_a)] p(i_a) q_{i_a \to i_b} \prod_{j \neq i}^{N} \sum_{j_a} \sum_{j_b} p(j_a) q_{j_a \to j_b} \,. \qquad (2\text{-}14)$$

If both probabilities are normalized such that the sum of $p(j_a)q_{ja\to jb}$ over j_a and j_b is equal to one, then Equation (2-14) can be reduced to

$$< w >= \sum_{i=1}^{N} \sum_{i_a} \sum_{i_b} \Delta\gamma_i \, p(i_a) q_{i_a \to i_b} \,, \qquad (2\text{-}15)$$

where

$$\Delta\gamma_i = \gamma(i_b) - \gamma(i_a) \,. \qquad (2\text{-}16)$$

However, we are interested in removing all the mutant sequences that contain at least one stop codon and all the mutant sequences that are identical on the DNA level to the wild-type. When the possibility of these transitions is removed, the probabilities are no longer normalized and removing the product of probabilities from Equation (2-13) is nontrivial.

The unmutated and stop-codon-containing mutants can be removed as follows. The transition probabilities sum such that

$$\sum_{i_b} q_{i_a \to i_b} + q^I + q_{i_a \to stop} = 1 \qquad \forall i_a \,, \qquad (2\text{-}17)$$

where $q^I = (1 - p_m)^3$ is the probability that the codon does not mutate and $q_{ia\to stop}$ is the probability of residue i mutating from amino acid a to a stop codon. The probability that the mutant sequence S_B contains no stop codons and at least one mutation is

$$P\{\text{no stop}; \geq 1 \,\text{mutation}\} \equiv D = \prod_{i=1}^{N} \left(1 - q_{i_a \to stop}\right) - \prod_{i=1}^{N} q^I \,. \qquad (2\text{-}18)$$

We can simplify the normalization procedure by assuming that all codons have an equal probability of mutating to a stop codon so

$$D = \left(1 - \frac{3}{63}Q\right)^N - (1-Q)^N, \tag{2-19}$$

where $Q = 1 - q^I$ is the probability that the codon was mutated. Equation (2-19) is used to normalize the transition probability $P\{S_A \rightarrow S_B\}$ so Equation (2-13) can be rewritten as

$$< w >= \frac{1}{D}\sum_{j=1}^{N}[\gamma(j_b) - \gamma(j_a)]\prod_{i=1}^{N}\sum_{i_a} p(i_a)\left(1 - \frac{3}{63}Q\right). \tag{2-20}$$

After, some rearrangement, the analogy with Equation (2-16) can be made,

$$< w >= C\sum_{i=1}^{N}\sum_{i_a}\sum_{i_b}\Delta\gamma_i P_i\,, \tag{2-21}$$

where

$$C = \frac{1}{1 - \left(\frac{1-Q}{1-\frac{3}{63}Q}\right)^N}, \tag{2-22}$$

and

$$P_i = \frac{p(i_a)\left[\frac{3}{63}Q + (1-Q)\delta_{a_i,b_i}\right]}{\left(1 - \frac{3}{63}Q\right)}, \tag{2-23}$$

where δ is a delta function. Thus, removing transitions to stop codons and unmutated sequences only requires renormalizing the probabilities $p(i_a)$ and adding a constant C. The average of the mutant distribution generated from the two-body fitness function can be found similarly,

$$< w >= C\left\{\sum_{i}^{N}\sum_{i_a,i_b}\Delta\gamma_i P_i + \frac{b}{2}\sum_{i}^{N}\sum_{j\neq i}^{N}\sum_{i_a,i_b}\sum_{j_a,j_b}\Delta\gamma_{ij}P_i P_j\right\}, \tag{2-24}$$

where

$$\Delta\gamma_{ij} = \left(\gamma(i_b, j_b) - \gamma(i_a, j_b)\right)\lambda_{ij} . \quad (2\text{-}25)$$

The second moment of the two-body mutant distribution can also be calculated, the details of which are shown in Appendix A.

Using the mean-field solution, the change in the fitness distribution is captured as the sequence ascends the fitness landscape (Figure 2-5). By increasing the number of two-body coupling interactions between residues, the effect of the landscape ruggedness on the moments is calculated. As the fitness of the wild-type increases, the first and second moments increase. In other words, as the sequence ascends the fitness landscape, the mutant distribution spreads out (diffuses) and becomes skewed towards less-fit mutants (drifts). In addition, the dependence of the moments on mutation rate can be predicted by recalculating the transition probabilities *q*. As the mutation rate increases, both the drift and the diffusion of mutants from the parent increases. Because rugged landscapes have less correlation between sequence points, the drift-diffusion effect becomes exaggerated as the coupling between residues increases.

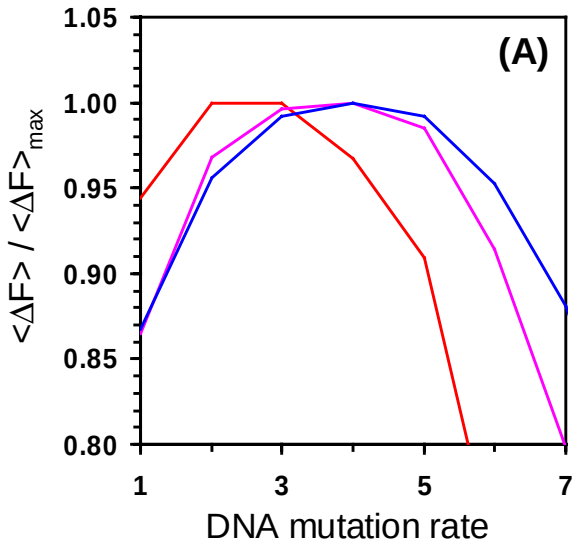
4. Conclusions

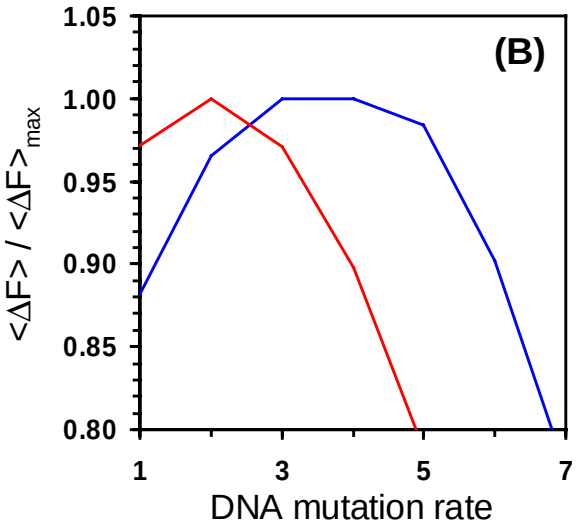
In this chapter, we have introduced several statistical models to study the dynamics of directed evolution. Using a spin-glass-like model, we explored the relationship between the optimal mutation rate, library size, fitness of the parents, and the interactions between amino acids. A bias was also discovered that mutations preferentially occur at uncoupled residues, when the mutation rate and number of mutants screened is small. A probabilistic model is then used to study the emergence of

compensating mutation in large libraries. Finally, the moments of the mutant fitness are calculated using mean-field theory and the fitness model. Ultimately, while these tools are useful in studying general trends in directed evolution, it became increasingly difficult to make the extension to real protein evolution systems. The large number of fit parameters that exist in the landscape-based models largely hindered this step. This difficulty is what inspired the use of inverse folding algorithms to calculate the energetics of interacting residues for specific protein structures, as described in Chapter 3.

Figure 2-1

The optimal DNA mutation rate as determined from a statistical model that captures the effect of interactions between amino acids. The genetic code is included in the model. The fitness improvement is the average maximum change in fitness for a given library size, as averaged over 10,000 landscapes. To compare the relative location of the optima, the curves have been scaled so that the optima are at 1.0. **(A)** The optimal mutation rate for the uncoupled landscape as the number of mutants screened increases M = 1000 (red line), 10,000 (purple line), and 50,000 (blue line). **(B)** The optimal mutation rate for a 1000-mutant library as the total number of interactions between residues (the "landscape ruggedness") increases. The number of coupling interactions is 75 (dotted line) and 0 (solid line). As the landscape ruggedness increases, the optimal mutation rate decreases. **(C)** The optimal mutation rate is shown as a function of the parental fitness for a smooth $\tau = 0$ (blue line) and rugged $\tau = 75$ (red line) landscape. As the parental fitness increases, the probability that a mutation is deleterious also increases, making a smaller mutation rate optimal.


(A)
1.05
1.00
0.95
0.90
0.85
0.80
<ΔF> / <ΔF>max
1
3
5
7
DNA mutation rate


(B)
1.05
1.00
0.95
0.90
0.85
0.80
<ΔF> / <ΔF>max
1
3
5
7
DNA mutation rate

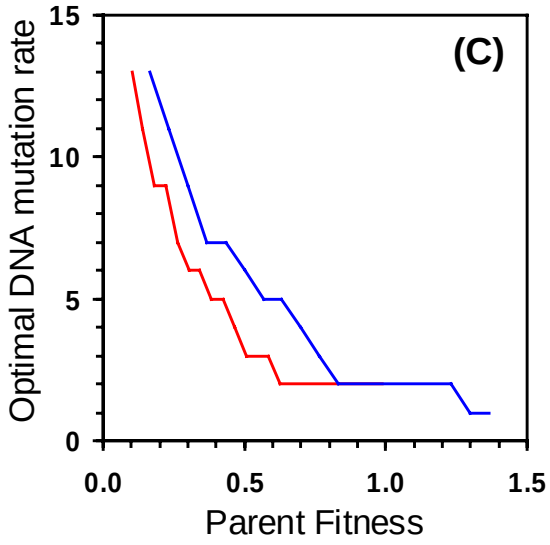

(C)
15
10
5
0
Optimal DNA mutation rate
0.0
0.5
1.0
1.5
Parent Fitness

Figure 2-2

The probability distribution $p(c)$ that a beneficial mutation is found by directed evolution at a residue with c coupled interactions. The distribution is shown at two fitness values as the sequence ascends the fitness landscape, F = 0.0 (O) and F = 17.0 (▲). Mutations were made on the DNA level and then translated into amino acid substitutions. A mutation rate of three nucleotide substitutions (corresponding to an average of one amino acid substitution) per gene was applied to a N = 50 amino acid residue sequence (b = 10.0). During each generation, 3000 mutants were screened and the coupling of the positions where mutations occurred on the most improved mutant was recorded.

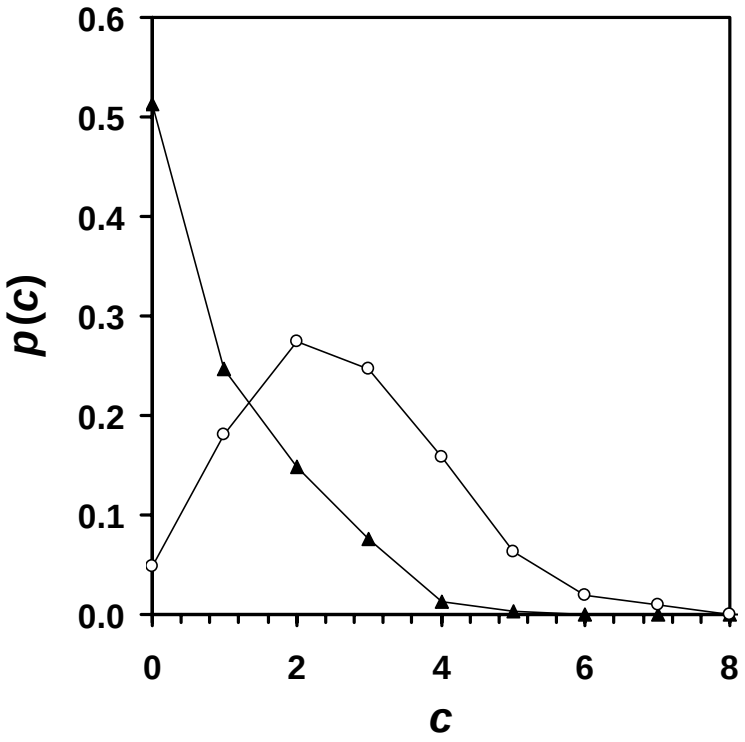

0.6
0.5
0.4
0.3
0.2
0.1
0.0
p(c)
0
2
4
6
8
c

Figure 2-3

Data are shown from screening libraries of antibody mutants (Daugherty *et al.*, 2000). The average mutation rate of a library m is plotted against the log fraction of functional mutants in the library P^f. After the initial exponential decline of P^f with m, a transition occurs and more mutants are functional at high m than is expected from the initial trend. This indicates that compensating mutations are being found in libraries at high mutation rates. The squares represent experimental data and the solid line is our model with f_d = 0.7, n_c = 110, and n_s = 12.

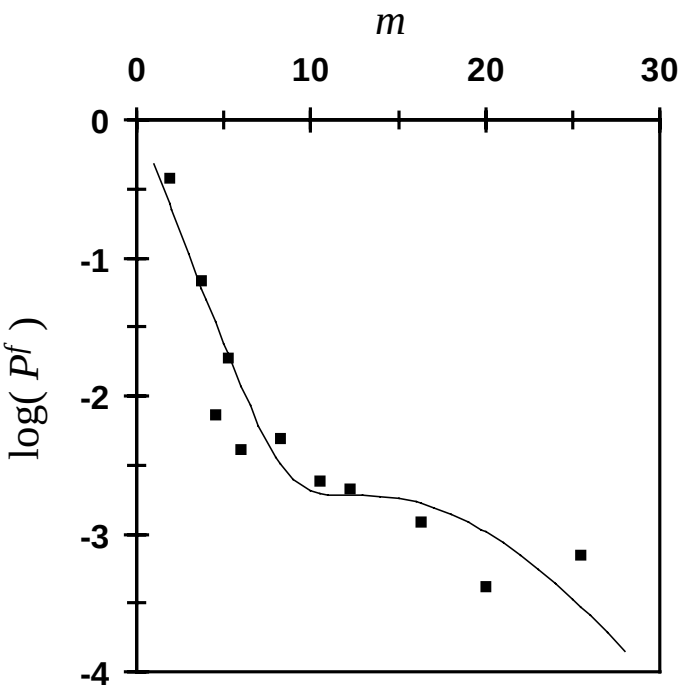
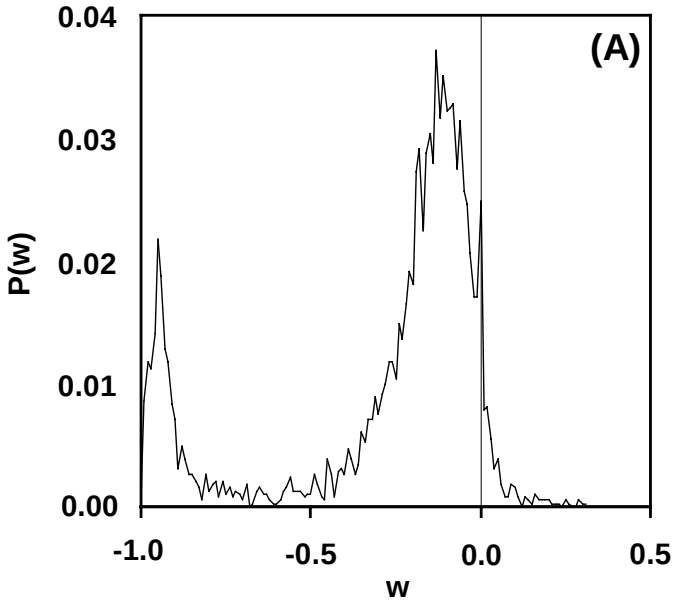

m
0
10
20
30
0
-1
-2
-3
-4
log(Pf)

Figure 2-4

The experimental and theoretical mutant fitness distributions. **(A)** An example of high-quality experimental mutant fitness data (May *et al*., 2000). The abscissa has been scaled so the range of fitnesses from parent to dead mutants is equal to one. The probability distribution has two discernable peaks: at $w = 0$, representing unmutated sequences, and at $w = -1.0$, representing non-functional mutants. **(B)** The theoretical mutant fitness distribution for the uncoupled fitness function as the sequence ascends the fitness landscape. The unmutated and stop-codon-containing mutants have been removed. The data is shown for $F = 0.0$ (black), 1.03 (red), and 1.83 (blue).


(A)
0.04
0.03
0.02
0.01
0.00
P(w)
-1.0
-0.5
0.0
0.5
w

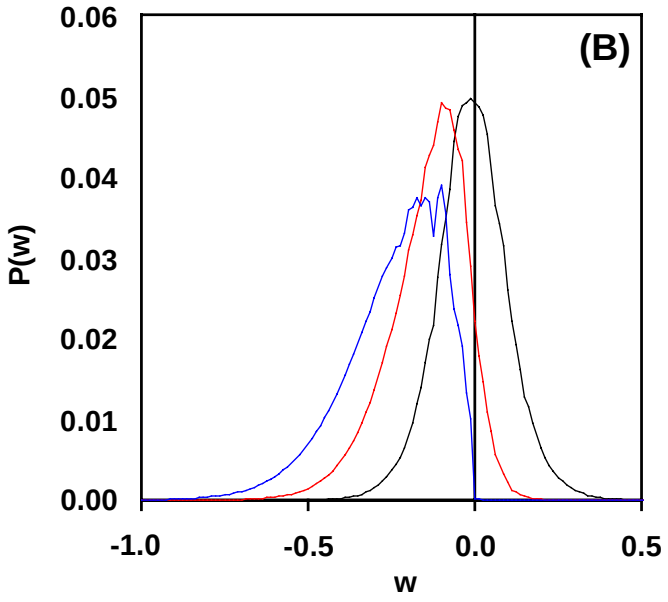
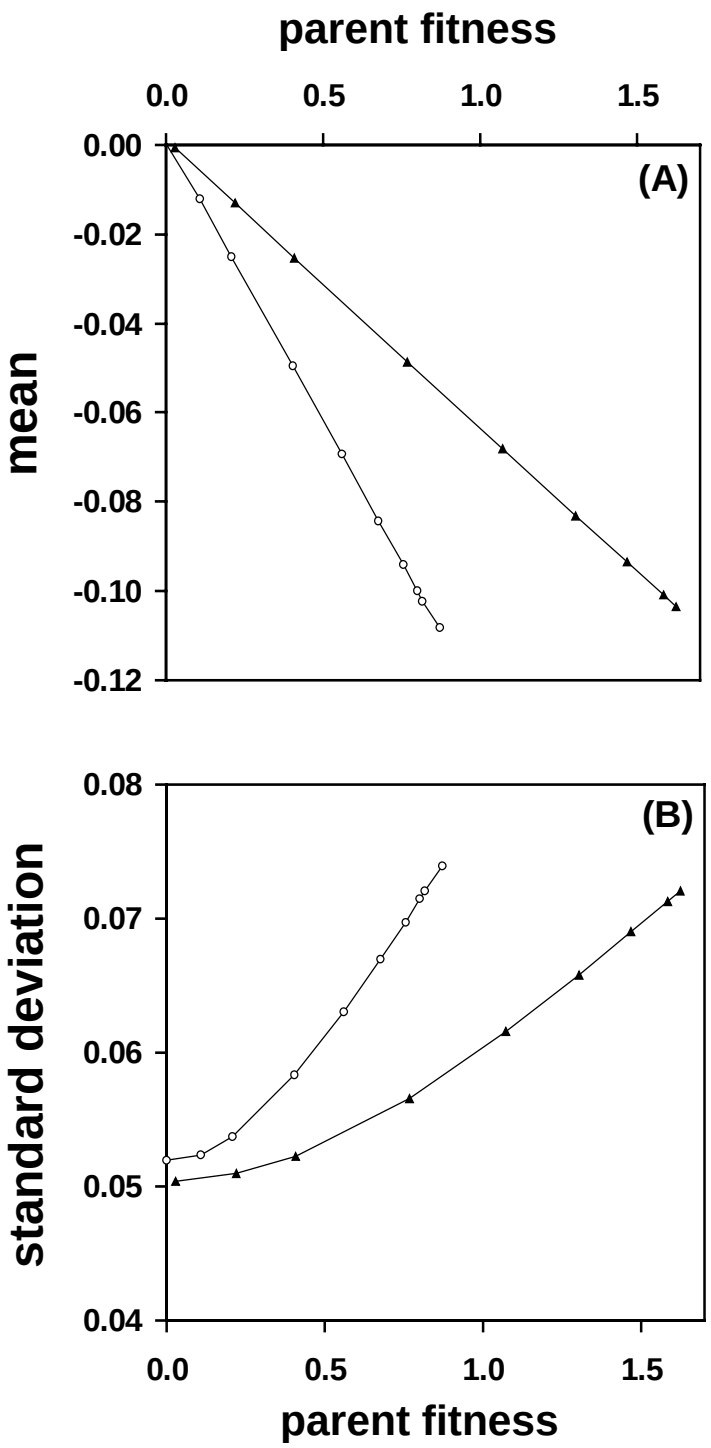

(B)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
P(w)
-1.0
-0.5
0.0
0.5
w

Figure 2-5:

The change in the mean **(A)** and standard deviation **(B)** of the mutant fitness distribution as determined using mean-field theory. The data is for an $N = 50$ residue sequence. As the fitness is increased, the distribution drifts to lower fitnesses (the mean decreases) and diffuses (the standard deviation increases). Two landscapes are shown: (▲) $\tau = 0$ and (O) $\tau = 30$. As the ruggedness of the landscape is increased, the distribution drifts and diffuses more rapidly.


parent fitness
0.0
0.5
1.0
1.5
0.00
-0.02
-0.04
-0.06
-0.08
-0.10
-0.12
mean
(A)
0.08
0.07
0.06
0.05
0.04
standard deviation
(B)
0.0
0.5
1.0
1.5
parent fitness

Chapter 3

Focusing Mutagenesis with Structural Information

Portions of this chapter are reproduced from:
Voigt, C. A., Mayo, S. L., Arnold, F. H., and Wang, Z-G. (2001). Computational method to reduce the search space for directed protein evolution. Proc. Natl. Acad. Sci. USA 98, 3778-3783

Abstract

We introduce a computational method to optimize the random mutagenesis of proteins. In Chapter 2, a statistical model of interacting residues was used to demonstrate that beneficial mutations tend to occur at amino acid positions that are tolerant to substitutions, in the limit of small libraries and low mutation rates. We transform this observation into a design strategy by applying mean-field theory to a structure-based computational model to calculate each residue's structural tolerance. Thermostabilizing and activity-increasing mutations accumulated during the experimental directed evolution of subtilisin E, T4 lysozyme, and antibody 4-4-20 are strongly biased to residues identified using this computational approach. This method can be used to predict positions where the probability of discovering beneficial mutations is maximized. Based on this strategy, we pick ten residues of β-lactamase that have high calculated structural tolerances and created libraries by mutagenizing these residues to all twenty amino acids. In seven out of ten of these libraries, amino acid substitutions are found that improve the antibiotic-resistance towards moxalactam. In contrast, beneficial mutations were not found at residues that are predicted to have low structural tolerances, but high solvent-exposed surface area or variability in an alignment of natural sequences.

1. Introduction

As techniques to alter the properties of proteins, directed evolution and computational design have matured separately. The aim of directed evolution is to accumulate stepwise improvements by iterations of random mutagenesis and screening (Moore and Arnold, 1996; Miyazaki *et al*., 2000). As a fundamentally different approach, the objective of computational protein design (Street and Mayo, 1999) is to solve the inverse folding problem by constructing a force field that describes the interactions between amino acids and then computing the globally optimal amino acid sequence (Dahiyat and Mayo, 1997; Malakaukas and Mayo, 1997). Directed evolution has the benefit of improving any enzyme property that can be captured by a screen, however, the search is restricted by the number of mutants that can be experimentally screened at each generation (~10^3–10^6). Conversely, computational design can effectively search a much larger number of sequences (>10^{26}) (Dahiyat and Mayo, 1997), but is limited as to the size of the protein and is currently restricted to optimizing the stabilization energy. This chapter introduces a new approach to protein engineering in which computational design is used as a guide to focus an evolutionary search, thus combining the benefits of both design strategies.

An effective and widely used directed evolution strategy is to produce a library of mutants from a parent sequence through random point mutagenesis using error-prone PCR (Moore and Arnold, 1996; Miyazaki *et al*., 2000). The usual practice of mutagenizing the whole gene has several problems. The probability that any single random mutation improves a property is small, and the probability of improvement decreases rapidly when multiple simultaneous mutations are made. Therefore, the limited

number of mutants that can be screened imposes a low upper limit on the mutation rate (Zhao and Arnold, 1999; Voigt *et al*., 2001a; Voigt *et al*., 2001b). Furthermore, the negligible probability that two or three mutations occur in a single codon and the significant biases of error-prone PCR severely restrict the possible amino acid substitutions. These effects can be overcome by intensely mutagenizing a limited number of positions (Skandalis *et al*., 1997; Nikolova *et al*., 1998; Miyazaki *et al*., 1999). The challenge, however, is to identify the residues where such experiments are likely to be successful, as beneficial mutations often appear far from sites that would be predicted heuristically (e.g., catalytic sites) (Moore and Arnold, 1996; Miyazaki *et al*., 2000). In Chapter 2, we used a simple fitness model to demonstrate that beneficial mutations preferentially occur at residues sharing the fewest interactions with the remainder of the protein. In Chapter 3, this observation is transformed into a design strategy through a detailed energetic model of structural interactions.

2. Computational Methods

2.1. Calculating the Tolerance of Protein Structures

As a strategy for directed evolution, concentrating mutagenesis on the regions of weak coupling to reduce the search space to the positions that are most likely to show improvement. We can extend this result from the simple model to make experimentally relevant predictions by using a detailed protein design model that calculates the stabilizing energy of a sequence folded onto a fixed backbone (Dahiyat and Mayo, 1997; Malakaukas and Mayo, 1998) to determine the coupling of each residue (Section 2.2). The protein backbones of subtilisin E (274 amino acids), T4 lysozyme (164 amino acids),

and antibody 4-4-20 (227 residues) were retrieved from high-resolution crystal structures (Matsumura *et al*., 1989; Whitlow *et al*., 1995; Jain *et al*., 1998), and the interactions between residues were calculated by coarse-graining the flexibility of each amino acid into rotamers and constructing a force field to calculate the rotamer/backbone and rotamer/rotamer stabilizing energies. An initial elimination of rotamers makes the problem computationally tractable; however, the combinatorial complexity remains enormous. The sequence space considered is hyper-astronomically large, for example, there are 10^{343} possible amino acid combinations for subtilisin E. Searching the entire space for the global optimum is intractable both computationally and experimentally.

To circumvent the combinatorial difficulties, we apply statistical mechanics to determine the coupling of each position, using structural tolerance towards amino acid substitutions as a measure of the coupling (Section 2.3). Structural tolerance is crucial for the success of directed evolution. Maintaining structure is required for the acquisition or fine-tuning of any other property, leading to the suggestion that properties such as stability and activity are correlated (Shoichet *et al*., 1995). A structurally tolerant protein has a larger number of allowed mutations that can potentially improve a property, making it more likely that there is a connected path in sequence space of single mutations that leads to regions of higher fitness. By reducing the evolutionary search to regions of sequence space that are consistent with the structure, functional space can be more thoroughly explored.

2.2. Force Field and Rotamer Library

Analogous in form to Equation (2-3), the energy term consists of two contributions: rotamer/backbone $e(i_r)$ and rotamer/rotamer $e(i_r,j_s)$:

$$E = \sum_{i=1}^{N} e(i_r) + \sum_{i=1}^{N-1} \sum_{j>i}^{N} e(i_r, j_s), \tag{3-1}$$

where N is the number of residues and i_r is rotamer r at position i. Because the backbone remains fixed, its internal energy contribution is not relevant to the optimization procedure. Note that the fitness functions described in Chapter 2 represent the negative of energy (*i.e.*, $F = -E$). Potential functions and parameters for van der Waals interactions, hydrogen bonding, and electrostatics are described in previous work (Dahiyat and Mayo, 1996; Dahiyat and Mayo, 1997). We use the DREIDING force field parameters for the atomic radii and internal coordinate parameters (Mayo *et al*., 1990). The van der Waals energies are modeled using a 6–12 Leonard-Jones potential with an additional 0.9 scale factor applied to the atomic radii to soften the lack of flexibility implied by the fixed backbone and the rotamer descriptions. A ceiling of 500 kcal/mol was set for the rotamer/rotamer energies to avoid unhindered van der Waals contributions and to expedite mean-field convergence. All rotamer/backbone and rotamer/rotamer energies are computed and stored prior to the mean-field calculation, requiring 165 (113, 138) minutes for subtilisin E (T4 lysozyme, antibody 4-4-20) on 10 Silicon Graphics R10000 processors running at 195 MHz.

The rotamer library is backbone-dependent as described by Dunbrack and Karplus (Dunbrack and Karplus, 1993; Dunbrack and Karplus, 1994). The following modifications were included as previously described (Dahiyat *et al*., 1997). The χ_3 angles that were undetermined from the database statistics were assigned the values: Arg,

−60°, 60°, and 180°; Gln, −120°, −60°, 0°, 60°, 120°, and 180°; Glu, 0°, 60°, and 120°; Lys, −60°, 60°, and 180°. The χ_4 angles that were undetermined from the database statistics were assigned the following values: Arg, −120°, −60°, 60°, 120°, and 180°; Lys, −60°, 60°, and 180°. Rotamers with combination of χ_3 and χ_4 resulted in sequential g^+/g^- or g^-/g^+ angles were eliminated.

Due to memory constraints, the entropy calculation is currently limited to handling up to 33000 rotamers on a SGI Origin 2000. Several filters had to be employed to reduce the number of rotamers below this limit. Rotamers that interact with the backbone with energies greater than 5 kcal/mol (subtilisin E), 20 kcal/mol (T4 lysozyme), and 1 kcal/mol (antibody 4-4-20) are eliminated from the calculation. The amino acids at residues 1−4 and 269−274 of subtilisin E are fixed in their wild-type identity and conformation. For subtilisin E, an average of 121 rotamers per residue are considered, corresponding to 3.2×10^4 one-body energies and 5.1×10^8 two-body energies. For T4 lysozyme, an average of 176 rotamers per residue are considered, corresponding to 2.9×10^4 one-body energies and 4.1×10^8 two-body energies. For antibody 4-4-20, an average of 144 rotomers per residue are considered, corresponding to 3.3×10^4 one-body energies and 5.3×10^8 two-body energies.

2.3. Mean-field Theory

The observation that some sequence positions are more tolerant to mutation initiated the application of information theory as a method to understand the importance of these residues to the structure and function of the protein (Reidhaar-Olson and Sauer, 1988; Saven and Wolynes, 1997). A residue that is intolerant to mutations has high

information content, whereas a residue that can be easily substituted has low information content. Directed evolution drives the system to minimum entropy by reducing the number of amino acid possibilities at each residue as the sequence becomes more optimized. We estimate the change in the number of sequences at increasing fitnesses by calculating the sequence entropy at a given energy, $S(E) = k_B \ln \Omega$, where the number of states Ω is the number of sequences at energy E.

The sequence entropy can be calculated from the probability distribution of allowed amino acid substitutions (Fontana and Shuster, 1987; Saven and Wolynes, 1997). The entropy s_i for a given residue i is calculated from

$$s_i(E) = -k_B \sum_a^A p(i_a) \ln p(i_a), \tag{3-2}$$

where A is the total number of amino acids, $p(i_a)$ is the probability that amino acid a exists at position i, and k_B is taken to be 1. A residue intolerant to mutations has a low entropy whereas a tolerant residue has high entropy. If all amino acids are equally likely, then $s_i = \ln A \approx 3.0$.

The total sequence entropy can be rewritten as the sum of the individual entropy of each residue,

$$S(E) = \sum_i^N s_i(E). \tag{3-3}$$

We apply mean-field theory to calculate the amino acid probabilities required by Equation (3-2), as a function of the energy (Lee, 1994; Koehl and Delarue, 1994; Koehl and Delarue, 1996; Saven and Wolynes, 1997). It is difficult to do the variation with a fixed energy. Instead, we use the thermodynamic equivalence of ensembles to work with

a fixed energy $\langle E \rangle_A$, where the average is over all sequences corresponding to a temperature T. Thus, the variational free energy

$$F = \langle E \rangle_A - TS_A \tag{3-4}$$

is minimized subject to the normalization condition,

$$\sum_r^{K_i} p(i_r) = 1, \tag{3-5}$$

for all i. The average energy is obtained from

$$\langle E \rangle_A = \sum_i^N \langle e(i_r) \rangle_A + \sum_i^N \sum_{j \neq i}^N \langle e(i_r, j_s) \rangle_A , \tag{3-6}$$

where the averages are taken over all amino acids at each position. Utilizing the mean-field approximation, this can be rewritten as

$$\langle E \rangle_A = \sum_i^N \sum_r^{K_i} p(i_r) e(i_r) + \sum_i^N \sum_{j \neq i}^N \sum_r^{K_i} \sum_s^{K_j} p(i_r) p(j_s) e(i_r, j_s) . \tag{3-7}$$

Introducing the Lagrange multiplier μ_i in the normalization of the probabilities for each site, the variational free energy is

$$\begin{aligned} F(T) = \sum_i^N \sum_r^{K_i} p(i_r) e(i_r) + \sum_i^N \sum_{j \neq i}^N \sum_r^{K_i} \sum_s^{K_j} p(i_r) p(j_s) e(i_r, j_s) + \\ k_B T \sum_i^N \sum_r^{K_i} p(i_r) \ln p(i_r) + \sum_i^N \mu_i \left(\sum_r^{K_i} p(i_r) - 1 \right) \end{aligned} . \tag{3-8}$$

Minimization of F is performed by setting the partial derivative $\partial F / \partial p(i_r)$ to zero for all i and r. After rearrangement, this gives

$$p(i_r) = \exp[-\beta \varepsilon(i_r) - 1 - \beta \mu_i], \tag{3-9}$$

where $\beta = 1/k_B T$ and

$$\varepsilon(i_r) \equiv e(i_r) + \sum_{j \neq i}^{N} \sum_{s}^{K_j} p(j_s) e(j_s) . \tag{3-10}$$

By solving for μ_i using the normalization condition (3-5), we find the partition function

$$e^{\beta\mu+1} = \sum_{r}^{K_i} e^{-\beta\varepsilon(i_r)} \equiv Z_i \tag{3-11}$$

and therefore,

$$p(i_r) = \frac{e^{-\beta\varepsilon(i_r)}}{Z_i} . \tag{3-12}$$

Equations (3-10) and (3-12) constitute a set of self-consistent equations for $p(i_r)$. The probability that an amino acid exists at a residue can then be calculated by summing over the rotamer probabilities for that amino acid, in other words,

$$p(i_a) = \sum_{r}^{K_{ia}} p(i_r) , \tag{3-13}$$

where K_{ia} is the total number of rotamers associated with amino acid *a* at residue *i*. The sequence entropy of each residue can then be calculated using Equation (3-2). Equation (3-13) assumes that the contribution of residue entropy to the free energy is small (e.g., the ambient temperature is 0 K). We solved the mean-field equations including a non-zero ambient temperature, but found that the resulting equations did not converge (Appendix D).

Operationally, the mean-field calculation is started by uniformly initializing the rotamer probabilities to $1/K_j$ and the mean-field energies are calculated via Equation (3-10) for each residue. The algorithm iterates between Equations (3-11) and (3-13) until self-consistency is achieved. Convergence is significantly improved if the probability vector *p* is updated with a memory of the previous step as described by Lee (Lee, 1994).

An initially high temperature (50,000 K) is set and the convergence algorithm is repeated as the temperature is lowered in increments of 100 K, until the final temperature (600 K for subtilisin E, 300 K for T4 lysozyme, and 500 K for antibody 4-4-20) is reached. The final temperature corresponds with an estimated energy above which the structural stability is compromised (Figure 3-1). The sequence entropy at this temperature effectively counts the number of sequences that are stable in the defined fixed backbone. The mean-field solution of subtilisin E required 8900 minutes on a single Silicon Graphics R10000 Processor running at 195 MHz and 2.1 gigabytes of physical memory. These are typical computing times and memory requirements for the entropy calculation. Other algorithms, based on the dead-end elimination and Monte Carlo algorithms, can be used to determine the energies required for the entropy calculation (Appendix B).

When solving the self-consistent equations, decreasing the temperature is analogous to decreasing the energy (increasing the fitness). As the energy is decreased, the number of sequences consistent with that energy decreases, thus decreasing the total entropy (Figure 3-1). The probabilities calculated as the temperature decreases are used to calculate the mean-field energy and entropy. The list of sequences consistent with an energy describes the tolerance of each position to amino acid substitutions, as measured by the sequence entropy. A small entropy represents the conservation of identity and a large entropy indicates mutability. When the energy is decreased, the sequence entropies of some positions drop rapidly, indicating a freezing of the amino acid identity while other positions remain highly variable (Figure 3-2). The sequence entropy captures the structural constraints on the amino acid identity at certain residues (Reidhaar-Olson and Sauer, 1988; Saven and Wolynes, 1997).

3. Results and Discussion

3.1. Correlation with Directed Evolution Experiments

To test our prediction that beneficial mutations discovered by directed evolution are biased towards structurally tolerant positions, we compared our calculations with mutations found from previous evolution experiments on subtilisin E (Chen and Arnold, 1993; You and Arnold, 1996; Zhao and Arnold, 1999) and T4 lysozyme (Pejura *et al*., 1993). Seven out of the nine mutations that improved the thermostability of subtilisin E occur at positions computed to be highly tolerant (Figure 3-3A and Table 3-1). The stabilizing mutations discovered by the evolution of T4 lysozyme also preferentially occur at the high-entropy positions (Figure 3-3B and Table 3-2). Thus, for both enzymes, the entropy predictions would aid an evolutionary search to improve thermostability, indicating that the computational method is valid independent of the specific protein or experimental protocol.

In directed evolution, it is often desired to improve properties other than stability. If the desired property is correlated with stability, then the structure-based entropy predictions will be more accurate. For instance, it has been suggested that improving thermostability is a good approach for enhancing activity at high temperatures (Giver *et al*., 1998; Zhao and Arnold, 1999). When libraries of subtilisin E mutants were screened for improved thermostability while retaining activity, some mutations improved both properties. In addition, the activity and stability are highly correlated in the screen used for T4 lysozyme; thus, the activity-improving mutations also occur at highly tolerant positions (Figure 3-6). There is a weaker correlation with improving the activity of subtilisin E in organic solvent (Figure 3-4A), implying that retention of structure is less

important. However, the mutations are still strongly biased towards the high entropy positions.

A directed evolution-type experiment was run to improve the binding of antibody 4-4-20 to fluorescein (Boder *et al*., 2000). In this experiment, yeast-displayed mutant libraries of antibodies were created and run over a binding column. The antibodies that bound tightly to the fluorescein were harder to wash from the column. Four rounds of increasing stringency (the level of required binding was increased) were performed and sets of improved mutants were isolated. Some mutations occurred only a few times in each data set and are considered neutral. Others occurred in less stringent rounds and became fixed as the stringency was increased. These mutations were considered essential for improved binding.

The average entropy for mutations discovered each round of improved stringency (excluding the neutral mutations) is compared to the experimental round in which the beneficial mutations were found (Figure 3-7). We find that as the fitness of the parent sequence increases, mutations are more concentrated at the high-entropy residues of the antibody. In addition, the standard deviation of individual entropies decreases as the fitness increases (data not shown). Together, these results indicate that, when the parent sequence is highly optimized, the beneficial mutations can be reliably found at the high-entropy positions. This correlation is an experimental verification of the dynamics of directed evolution discovered using the generic statistical model (Figure 2-3). As the parent sequence becomes more optimized, the probability that a beneficial mutation will occur at a highly coupled residue decreases dramatically.

3.2. Solvent Accessibility and Natural Diversity

The entropy profile is mapped onto the subtilisin E structure in Figure 3-5. There is a trend towards the most variable sites being on the surface and the more conserved being in the core of the protein. However, the correlation between the entropy and solvent accessibility is poor (R^2 = 0.55 for subtilisin E and 0.54 for T4 lysozyme, Figure 3-8A). The computed sequence entropies are derived from the fundamental physical features that lead to tolerance, whereas solvent accessibility is a secondary measure. The sequence entropy captures details of structural tolerance beyond solvent accessibility, including side chain packing, the coupling of backbone and side chain conformations, electrostatic interactions required by the backbone conformation, and a residue's local environment. In addition, the mean-field algorithm considers the energetic effects of all amino acid substitutions, rather than using a measure based on the single wild-type amino acid identity, as in the solvent accessibility calculation. This leads to a more accurate assessment of the tolerance of a residue for amino acid substitutions.

A comparison is made in Tables 3-1 and 3-2 between the sequence entropies and solvent accessibilities of the positions where beneficial mutations were found. Some residues with low solvent accessibility are predicted to have a high sequence entropy. Several specific residues have a high sequence entropy, but a low solvent accessibility, which demonstrate the physical principles underlying our method. For example, residue 107 in subtilisin E has an above-average sequence entropy (1.62), but a very low solvent accessibility (1%). Residue 107 is on an α-helix and the wild-type isoleucine side chain is oriented towards the center of the protein and is completely buried. However, the packing of the side chains of the surrounding residues is such that several other amino acids can

be substituted with minimal effect on the stabilization energy. After the mean-field calculation, the amino acids that are acceptable at this position (and their probabilities) are: Ile (0.42), Cys (0.23), Val (0.12), Met (0.09), Glu (0.09), Asp (0.03), Thr (0.01), Ser (0.01), and Ala (0.01). The result of the evolution experiment was an Ile $\rightarrow$ Val substitution, which increased the activity in organic solvent.

A similar example exists in the T4 lysozyme data set. Residue 151 is on an α-helix near the surface and is partially blocked from the solvent by surrounding atoms. It has an above-average site entropy (1.53) and below-average solvent accessibility (17%). The mean-field calculation reveals that the amino acids possible at this position are: Met (0.37), Leu (0.34), Cys (0.11), Glu (0.09), Gln (0.05), Asp (0.03), Ser (0.01), and Thr (0.01). The evolution experiment generated a Thr $\rightarrow$ Ser substitution. Typically, the positions with high entropies (greater that one standard deviation above the mean) and below average solvent accessibilities (< 24% exposed) are close to the surface and their side chains are partially buried.

We also compared the sequence entropies with the diversity accumulated during natural evolution, calculated from a sequence alignment (Figure 3-8B). The sequence alignment entropy was determined from the sequences of subtilisins SSII, S41, S39, BPN', E, Carlsberg, and thermitase (Siezen and Leunissen, 1997). The amino acid probabilities $p(i_a)$ are calculated as the fraction of aligned sequences where amino acid a exists at position i. We find that the calculated entropies correlate poorly with the natural amino acid variability. Because the natural sequence variability among subtilisins is great, the correlation worsens as more sequences are compared.

That the site entropy can predict the positions where mutations occur in *in vitro*, but not in natural evolution, is interesting. This disparity is due to a combination of two effects, both related to the limited number of mutants that can be screened. First, the theory that we present relies on the assumption that the number of mutants screened is relatively small. The analog of this in nature is unclear; however, it is expected that many more mutants have been attempted in nature than can be currently analyzed in the laboratory. Second, long periods of neutral evolution have eroded the information in the sequence alignment. Multiple mutations can be made to achieve a punctuated fitness improvement over long time periods via the accumulation of neutral mutations, which eventually discover beneficial combinations (Fontana and Shuster, 1998). However, the probability of finding a good multiple mutant during *in vitro* evolution is small due to the sampling limitation of the experiment (analogous to a time limitation).

3.3. Combinatorial versus Site-directed Mutagenesis Data

The sequence entropies condense the energetic information from the computational design calculation in a way that is useful to guide directed evolution. It is important to emphasize that our algorithm describes the positions where mutations will be discovered with the intention of optimizing directed evolution as a search algorithm. The probability that beneficial mutants are found increases when the high entropy positions are targeted and low-entropy sites are neglected. Non-combinatorial experiments, such as rational design strategies of alanine-scanning mutagenesis, will generally not correlate with the entropy prediction.

The requirement for a combinatorial component to the experiment is demonstrated by the example probabilities given above for residue 107 in subtilisin E and residue 151 in T4 lysozyme. In both examples, the amino acid substitution found by the evolution experiment does not correspond with the highest probability case determined by the calculation. Once the algorithm determines the positions where substitutions do not disrupt the structure, evolutionary experiments can determine the specific mutations that generate the greatest fitness improvements.

3.4. Correlation with Site Saturation Experiments

Exhaustive datasets have been experimentally generated to test the functional tolerance of all the residues of T4 lysozyme (Rennell *et al*., 1991), β-lactamase (Huang *et al*., 1996), and λ repressor (Reidhaar-Olson and Sauer, 1988; Reidhaar-Olson and Sauer, 1990). The tolerance data for each of these examples was linked to a selection, thus allowing for more variants to be screened and reducing the problem of obtaining adequate sampling. Figure 3-9 compares each of these datasets with the structural entropy calculated using our methodology. In general, the results between the theory and experiments correlate well. There are several difficulties in using the functional datasets to test and refine the entropy calculation. Functional diversity is not exactly comparable to structural diversity and differences between the two measures often reveal functionally important residues or regions surrounding the active site. In addition, to reduce the materials required for the experiments, sets of residues are often mutated simultaneously, which allows for the possibility of generating compensating mutations. Individually mutating residues could result in a different measure of tolerances.

4. Computationally Focused Mutagenesis

4.1. Focusing Strategies

The information from the sequence entropy calculations can be incorporated in several experimental methods. First, site saturation mutagenesis can be applied at positions that are predicted to be the most tolerant. The beneficial mutants can then be recombined using DNA shuffling (Stemmer, 1994) to compound the fitness improvement. As a second method, a portion of the gene that is determined to have an above-average total tolerance (such as residues 240 to 255 in subtilisin E) can be targeted using regional combinatorial mutagenesis. The choice of experimental approach is determined by the accuracy of the entropy profile. If the correlation between the screened property and stability is high, then site saturation mutagenesis is appropriate. However, if the correlation is weaker, a combinatorial search of a region that is predicted to be able to withstand the additional diversity is better.

The experiment can also combine mutagenesis with recombination, a method conceptually similar to family shuffling, in which homologous genes are recombined (Crameri *et al*., 1998). In family shuffling, the sequences have previously survived natural selection; thus, the inherent diversity is less likely to have a deleterious effect on the structure and function. In our approach, the calculated entropy profile predicts the positions that are essential to maintain the structure, allowing the tolerant sites to be mutated *en masse* to produce a family of artificially divergent sequences. Recombining these sequences could generate a mutant library with large sets of mutations that are calculated to retain structural integrity.

4.2. Saturation Mutagenesis Experiments with β-lactamase

The antibiotic-degrading enzyme TEM-1 β-lactamase was chosen as a model system to test the entropy calculation as a predictive method. The structure of the TEM-1 variant is available (Jelsch *et al*., 1993) and the mean-field and DEE-entropy calculations were run and the residues were ranked by their entropy. Many studies have been performed that evolve β-lactamase to improve its antibiotic resistance or to compare naturally occurring β-lactamase variants. As a result, there is much information on which residues can be mutated to confer improved activity in degrading various antibiotics. To avoid selecting residues where mutations have been found previously, we removed these residues from our list. The top ten remaining residues were then selected for mutagenesis (Figure 3-10).

The TEM-1 gene was obtained from the pSTBlue-1 vector offered by Stratagene. Each of the ten chosen residues was individually mutated to all twenty amino acids, using the standard Quickchange saturation mutagenesis protocol. For each mutated residue, a forward and reverse primer is synthesized with the bases of the mutated residue randomized. The libraries were then transformed into XL1-BLUE (10^6) competent cells. Each library was screened for mutants that have a higher activity in degrading the antibiotic moxalactam. To rapidly screen for this property, agar plates are made with following exponentially increasing concentrations of moxalactam: 0.45, 0.9, 1.8, and 3.6 μg/ml. Aliquots of the cell libraries are spread on the plates and allowed to grow for 24 hours. More active hybrids will grow on plates with greater concentrations of moxalactam. The activity is measured as the minimum inhibitory concentration (MIC), in

other words, the lowest concentration of ampicillin that kills the cells. Wild-type TEM-1 has a MIC of 0.38 towards moxalactam.

Improvements were found at seven out of ten of the high entropy residues targeted (Table 3-3). The largest improvement we obtained from a single mutant is 4-fold (residue 198). Interestingly, while this residue has a high entropy, it is partially buried and would not be selected based on its solvent exposure. All of the beneficial mutations retained wild-type ampicillin resistance. As a negative control, we performed the same experiment at two low entropy residues. Residue 229 was chosen because it has a high solvent accessibility, but a low entropy. Residue 268 was chosen because it also has a low entropy, but has a high sequence variability when compared to a sequence alignment of naturally occurring β-lactamase variants. No improvements were found at either of these residues. These experiments demonstrate the success of using the entropy calculation as a method to assess the structural impact of mutations.

The beneficial mutations were found individually by screening libraries created by randomizing a target residue. The beneficial mutations that are found by this process can be combined onto a single gene through a myriad of methods. Currently, we are individually combining the mutations onto a single gene by using the Quickchange protocol (described above) with primers that correspond with each single mutation. This can be achieved by various other methods, including using *in vitro* recombination to fragment and recombine the best mutants that are found at each residue. This method is also advantageous when multiple improvements are found at a single residue. The optimal combination of mutations can be determined using *in vitro* recombination and

screening the library for the combination of mutations that results in the largest improvement.

5. Conclusions

Because beneficial mutations are found at high entropy sites, we propose that mutagenesis should be preferentially applied to these regions. An alternative approach is to make specific mutations at a highly coupled set of residues, a strategy that has been successful in improving the stability of small proteins (Dahiyat and Mayo, 1997; Malakaukas and Mayo, 1998). However, we are interested in improving properties such as activity, where the exact fitness contributions cannot be accurately computed. Experimentally incorporating a sufficiently high mutation rate to reliably discover highly coupled mutants requires a screening effort larger than is practically feasible. Our algorithm provides a methodology by which enzymes can be computationally pre-screened, thus reducing the required experimental effort. By computationally calculating the entropy of each residue and using this information to guide an experimental evolutionary search, the most powerful aspects each technique are combined as a new approach to protein design.

Table 3-1. Site Entropies and Solvent Accessibility of Subtilisin E

residue	site entropy	% exposed[a]
9	2.55	56
14	2.50	34
48	2.09	20
60	0.00	0
76	2.45	46
97	0.06	19
103	2.48	61
107	1.62	1
118	2.37	79
131	2.43	37
156	2.19	53
161	2.69	92
166	0.96	8
181	0.36	23
182	1.81	52
188	2.50	88
194	2.59	71
206	1.94	40
218	2.54	50
255	2.54	41

[a] The percent surface area of the side chain accessible by solvent. The surface areas were calculated using the Lee and Richards definition of solvent accessible surface area using 1.4 Å as the radius of water (Lee and Richards, 1971). The average solvent accessibility is 24% and the standard deviation is 26%.

Table 3-2. Site Entropies and Solvent Accessibility of T4 Lysozyme

residue	site entropy	% exposed[a]
14	2.59	47
16	2.02	53
22	1.66	19
26	1.03	2
40	2.54	80
41	1.91	34
93	2.52	81
113	2.54	69
116	2.50	51
119	2.11	54
147	2.10	50
151	1.53	17
153	0.55	0
163	2.49	63

[a] The percent surface area of the side chain accessible by solvent. The surface areas were calculated using the Lee and Richards definition of solvent accessible surface area using 1.4 Å as the radius of water (Lee and Richards, 1971). The average solvent accessibility is 24% and the standard deviation is 26%.

Table 3-3. Saturation of High-Entropy Residues of β-lactamase

Residue	Improvement[a]	Mutations
39	2-fold	Q → R
		Q → N
57	none	
90	2-fold	Q → S
99	2-fold	Q → R
114	none	T → K
		T → A
		T → N
140	2-fold	
158	2-fold	H → Y
198	4-fold	L → I
227	2-fold	A → D
273	none	

a. The improvement represents the improvement in MIC over wild-type TEM-1. The notation '2-fold' refers to a MIC of 0.9 μg/ml and '4-fold' refers to a MIC of 1.8 μg/ml.

Figure 3-1:

A plot is shown of sequence entropy versus stabilization energy. The entropy is the log of the number of sequences, $S = \ln \Omega$. At $E = 0$, all sequences are possible and the entropy is at a maximum. At zero entropy, only a single sequence, the global optimum, remains. A critical energy (marked by the arrow) represents a threshold, below which sequences are stable in the defined structural context (blue) and the sequences above this threshold are unstable (red).

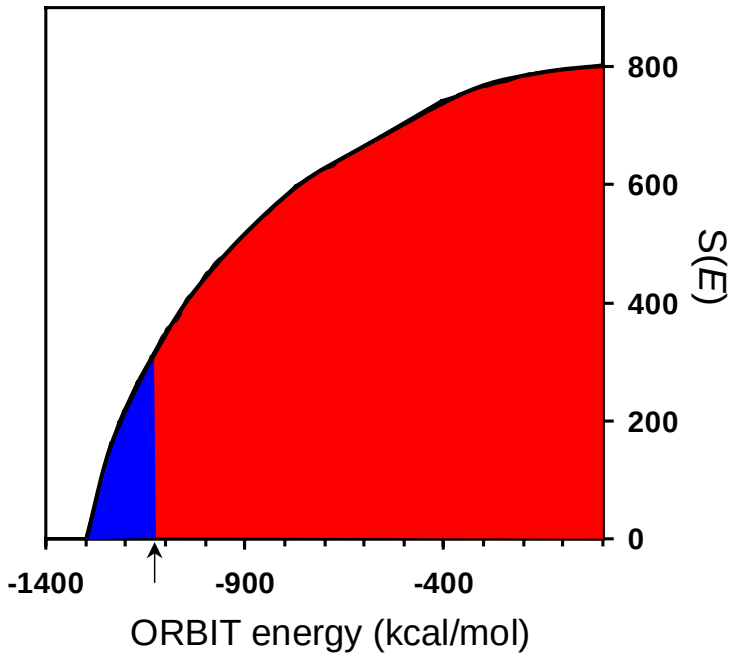

800
600
400
200
0
S(E)
-1400
-900
-400
ORBIT energy (kcal/mol)

Figure 3-2:

The sequence entropy of each residue is plotted as a function of increasing fitness (from light to dark blue). The data is for the simple model of fitness used in Chapter 2 (Equation 2-3). The correlate of increasing fitness is decreasing stabilization energy, in other words, $F = -E$. If all amino acids are equally likely, then $s_i = \ln A \approx 3.0$. As the fitness increases, the number of possible amino acid substitutions at each residue decreases, thus decreasing the total sequence entropy. At high fitness, some residues remain tolerant to substitution, while others become fixed in a single amino acid identity.

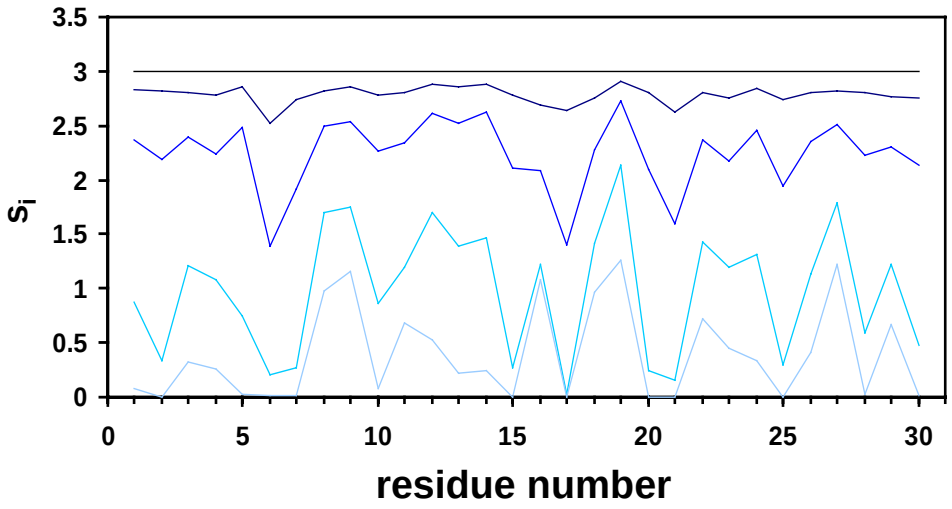

3.5
3
2.5
2
1.5
1
0.5
0
S_i
0
5
10
15
20
25
30
residue number

Figure 3-3:

The predicted sequence entropy profile (black line) and solvent accessibility (red line) for subtilisin E. If all amino acids are equally likely, then $s_i = \ln A \approx 3.0$. The solvent accessibility is the percent side chain surface area exposed, as calculated by the Lee and Richards method with a solvent radius of 1.4 Å (Lee and Richards, 1971).

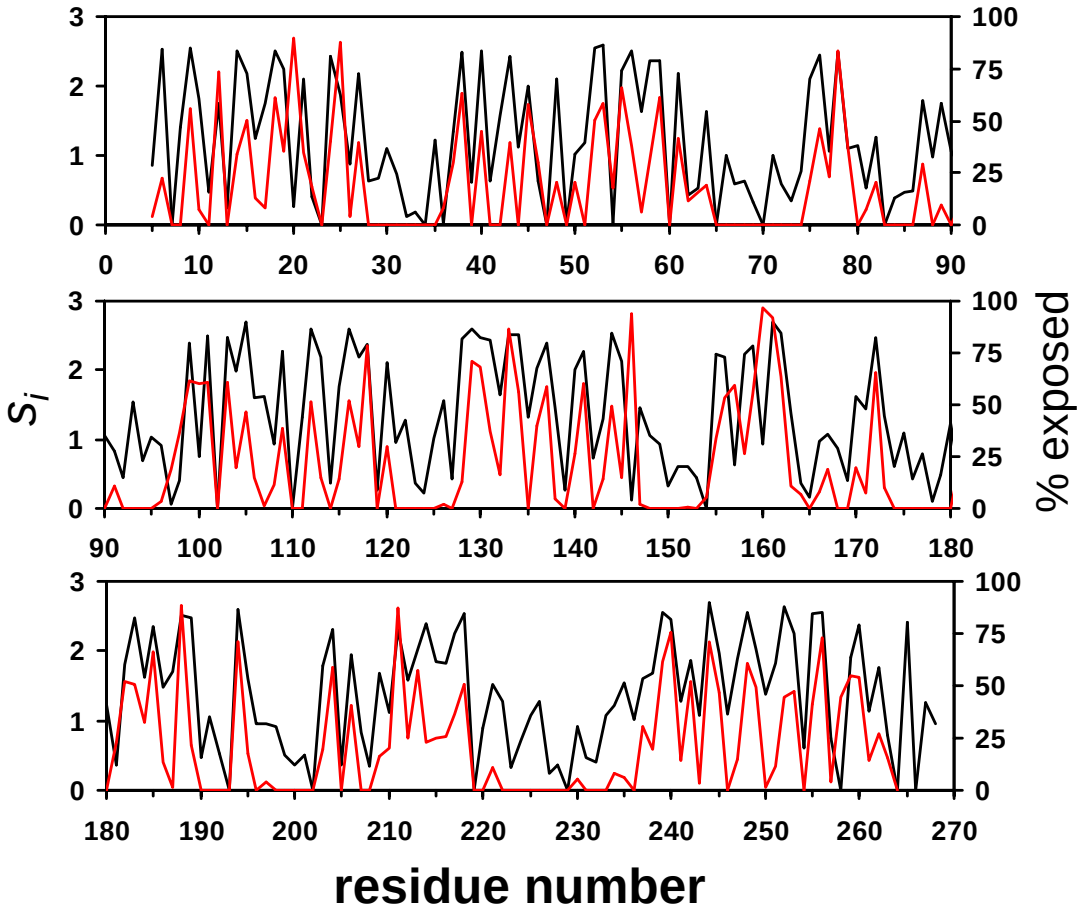
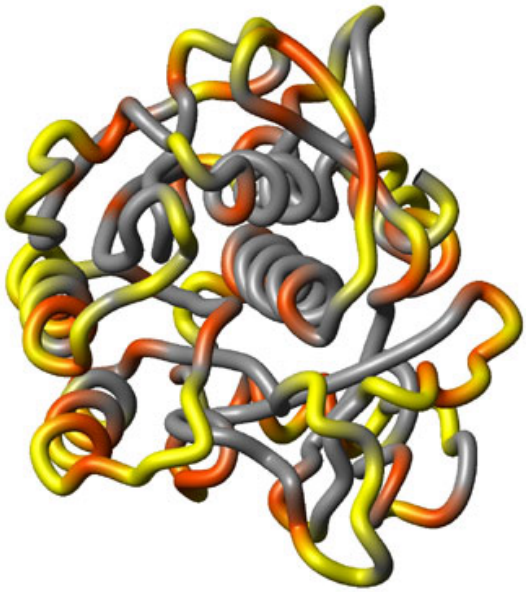

S_i
% exposed
residue number
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
0
1
2
3
0
25
50
75
100

Figure 3-4:

The entropy of each residue mapped onto the structure of subtilisin E **(A)** and T4 lysozyme **(B)**. The yellow residues are the most variable sites ($2.16 < s < 3.00$, greater than one standard deviation above the mean), the red residues are moderately variable ($1.31 < s < 2.16$, between the mean and one standard deviation), and the gray residues have below average variability ($s < 1.31$). Site saturation experiments should be directed at yellow positions whereas the contiguous yellow-red regions lend themselves to cassette mutagenesis. This figure was generated using MolMol (Koradi *et al*., 1996).



(A)

(B)

Figure 3-5 (A):

The probability distribution of site entropies $p(s_i)$ for subtilisin E. The bar indicates the mean and standard deviation of the distribution. The fraction of residues with zero entropy is 0.078, as indicated by the arrow. The site entropies of positions where experimental directed evolution found positive mutations are indicated by the lines. These beneficial mutations were found by screening ~10^3 mutants generated with an average mutation rate of 2 – 3 nucleotide substitutions. (top) Mutations made when the screen was to improve thermostability while retaining activity (Zhao and Arnold, 1999). From left to right, the positions (entropies) are 181 (0.36), 166 (0.96), 118 (2.37), 76 (2.45), 14 (2.50), 218 (2.54), 9 (2.55), 194 (2.59), and 161 (2.69). (bottom) Mutations made when the screen was to improve activity towards s-AAPF-*p*Na in the organic solvent dimethyl formamide (Chen and Arnold, 1993; You and Arnold, 1994). From left to right, the positions (entropies) are 60 (0.0), 97 (0.06), 181 (0.36), 107 (1.62), 182 (1.81), 206 (1.94), 48 (2.09), 156 (2.19), 131 (2.43), 188 (2.50), 103 (2.48), 218 (2.54), 255 (2.54). Note that residues 181 and 218 are common to both data sets (different amino acid substitutions were made at residue 181, whereas the same substitution was made at 218).

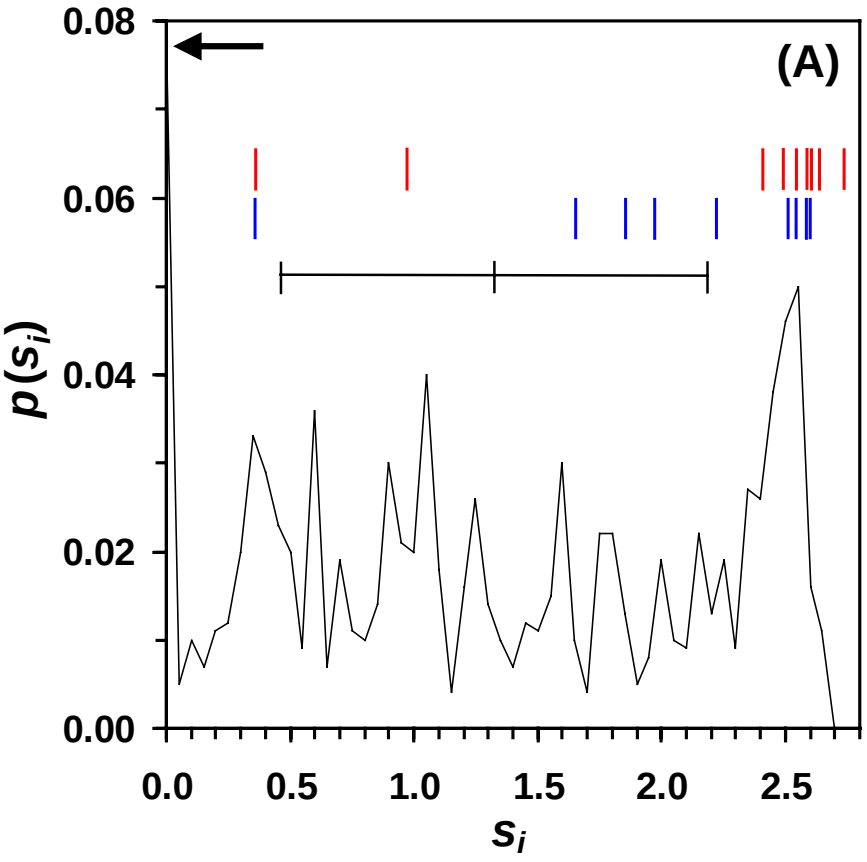

(A)
0.08
0.06
0.04
0.02
0.00
p(s_i)
0.0
0.5
1.0
1.5
2.0
2.5
s_i

Figure 3-5 (B):

The probability distribution of site entropies $p(s_i)$ for T4 lysozyme. The bar indicates the mean and standard deviation of the distribution. The fraction of residues with zero entropy is 0.039, as indicated by the arrow. The site entropies of positions where experimental directed evolution found positive mutations are indicated by the lines (Pjura *et al*., 1993). These beneficial mutations were found by screening ~10^3 mutants generated with an average mutation rate of 2–3 nucleotide substitutions. The red bars indicate mutations that improved stability, blue bars indicate mutations that improved activity, and purple bars indicate mutations that improved both properties. From left to right, the positions (entropies) are 153 (0.55), 26 (1.03), 151 (1.53), 22 (1.66), 41 (1.91), 16 (2.02), 147 (2.10), 119 (2.11), 163 (2.49), 116 (2.50), 93 (2.52), 113 (2.54), 40 (2.54), and 14 (2.59).

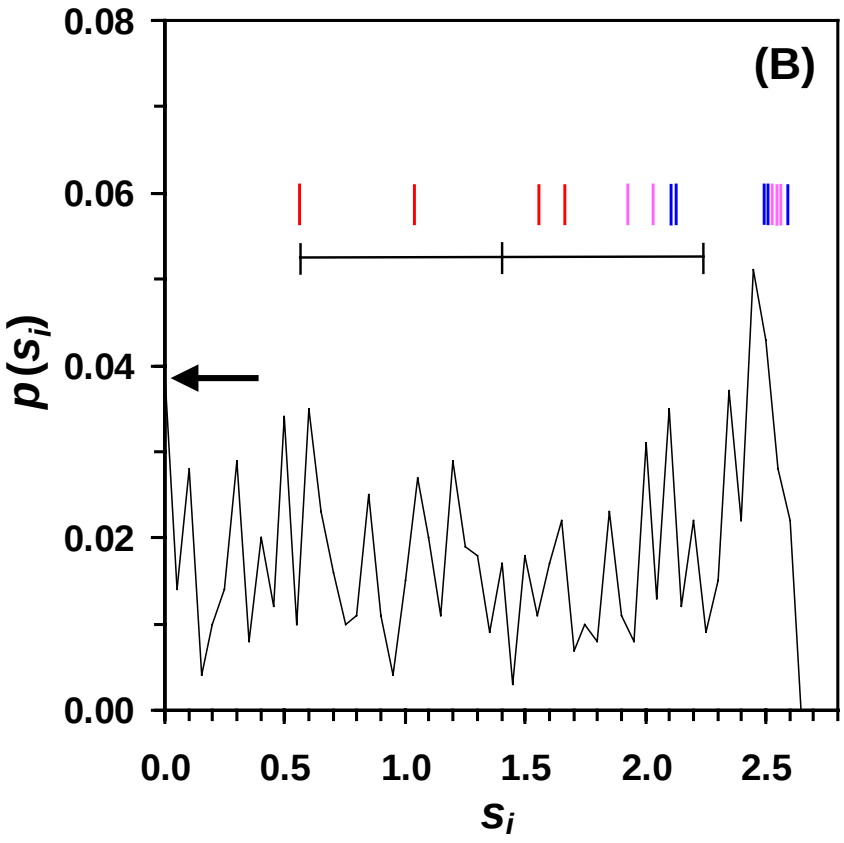

(B)
0.08
0.06
0.04
0.02
0.00
$p(s_i)$
0.0
0.5
1.0
1.5
2.0
2.5
s_i

Figure 3-6:

The percent of improvement in activity is plotted against the entropy at which the mutation was found for the T4 lysozyme dataset (Pjura *et al*., 1993). The largest activity improvements occur at the highest entropy positions. The degree to which a mutation stabilizes T4 lysozyme does not correlate with the entropy of the site where the stabilizing mutation was found (data not shown).

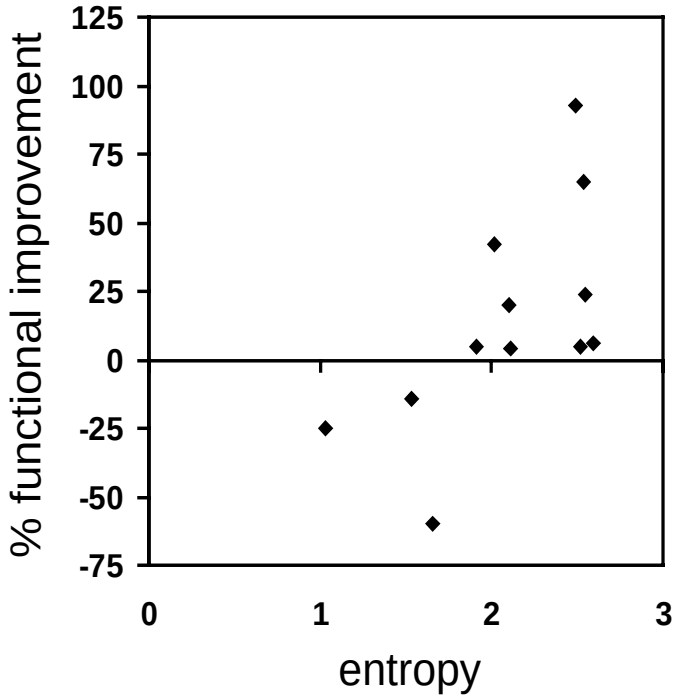

% functional improvement
125
100
75
50
25
0
-25
-50
-75
0
1
2
3
entropy

Figure 3-7:

The entropy calculation is compared to results from the directed evolution of antibody 4-4-4-20 (Boder *et al*., 2000). The average sequence entropy of residues where beneficial mutations were found to improve binding to flourescein is plotted versus the experiment in which it occurred. After each round, the stringency of selection was increased to reflect the increase in affinity obtained in the previous round. As the fitness of the parents increases, beneficial mutations became more biased towards the high entropy residues. This is predicted using the simplified statistical model of interacting residues introduced in Chapter 2 (Figure 2-2).

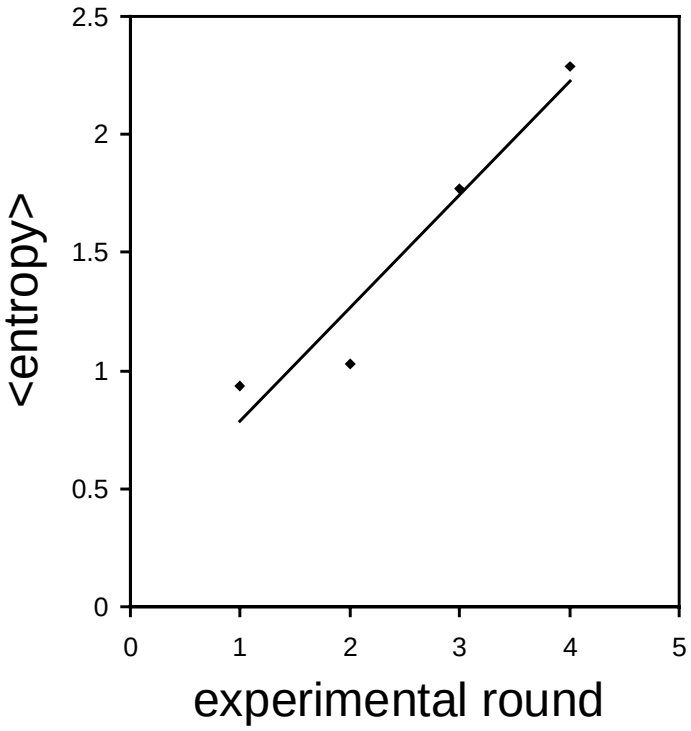
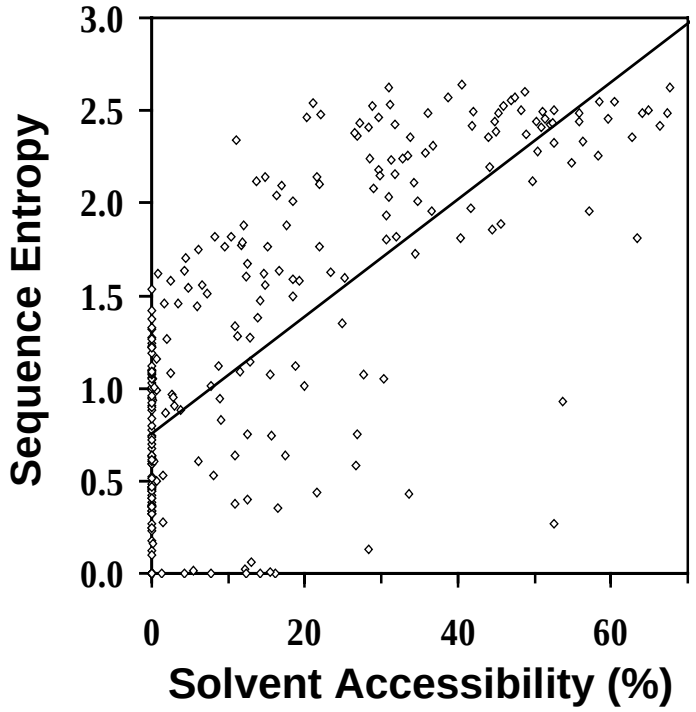
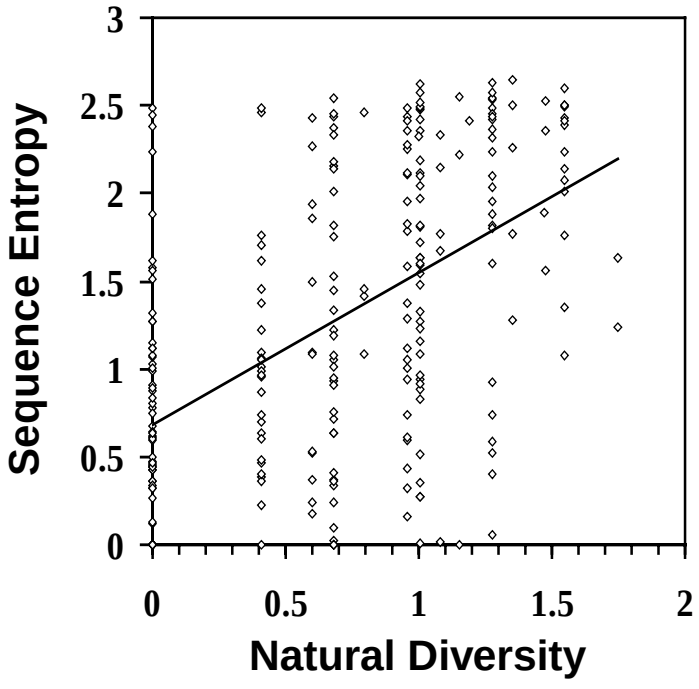

2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
<entropy>
experimental round

Figure 3-8:

The entropy of each residue of subtilisin E is plotted versus the fraction of the side-chain that is surface exposed **(A)** and the variation that is observed in a natural alignment of subtilisins **(B)**. The best fit line is shown for both data sets. The solvent accessibility is the percent side chain surface area exposed, as calculated by the Lee and Richards method with a solvent radius of 1.4 Å (Lee and Richards, 1971). The natural diversity is presented as a entropy which is calculated using a sequence alignment. The alignment is used to calculate the probabilities of amino acids existing at a particular residue. The probabilities are then converted to entropies using Equation (3-2).


Sequence Entropy
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
20
40
60
Solvent Accessibility (%)

(A)


Sequence Entropy
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
Natural Diversity

(B)

Figure 3-9 (A):

The site entropy (black line) is compared with the functional diversity (red line) for T4 lysozyme mutants (Rennell *et al*., 1991). The functional diversity of each residue is a count of all the amino acid substitutions that retained wild-type activity (a rating of ++). Those positions that are structurally tolerant and functionally intolerant tend to occur near the active site. The site entropy was calculated via the mean-field algorithm and the calculation was stopped at T = 300 K. The data is identical to that which was used to create Figure 3-5B.

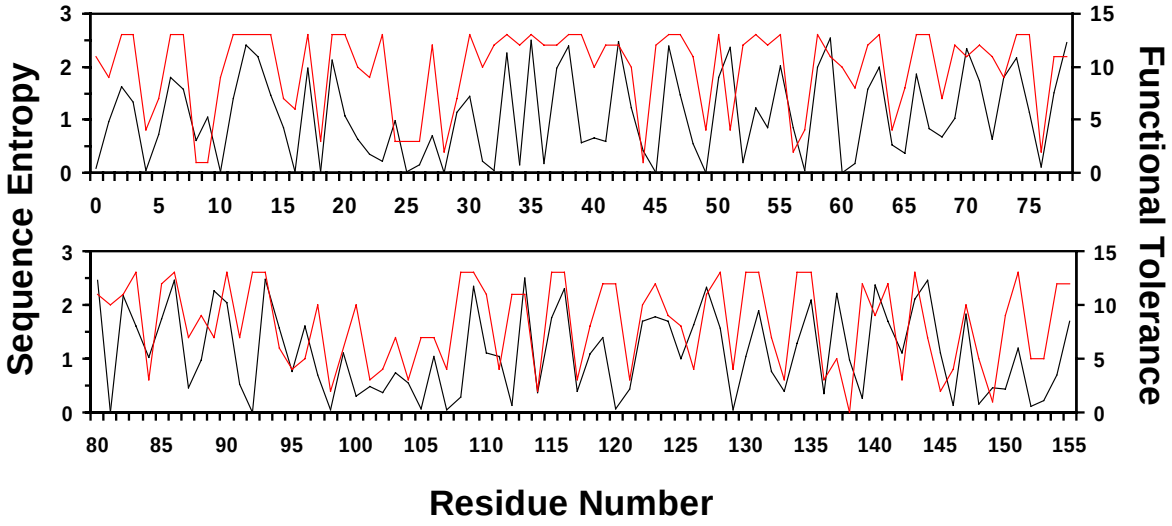

Sequence Entropy
Functional Tolerance
Residue Number
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
0
1
2
3
0
5
10
15

Figure 3-9 (B):

The site entropy (black line) is compared to the functional tolerance (red line) of TEM-1 β-lactamase (Huang *et al*., 1996). The functional tolerance is defined as the number of amino acid mutations at a residue that conferred wild-type activity towards the degradation of ampicillin (measured as a minimum inhibitory concentration). The amino acid substitutions were performed in sets of 3-6 contiguous residues simultaneously. This could produce very different results from an experiment where substitutions are made at each residue independently. The entropy was calculated using the DEE-entropy algorithm (Appendix B) using $\beta = 0.6$. The break in the profiles at 238 reflects a numbering convention in β-lactamases.

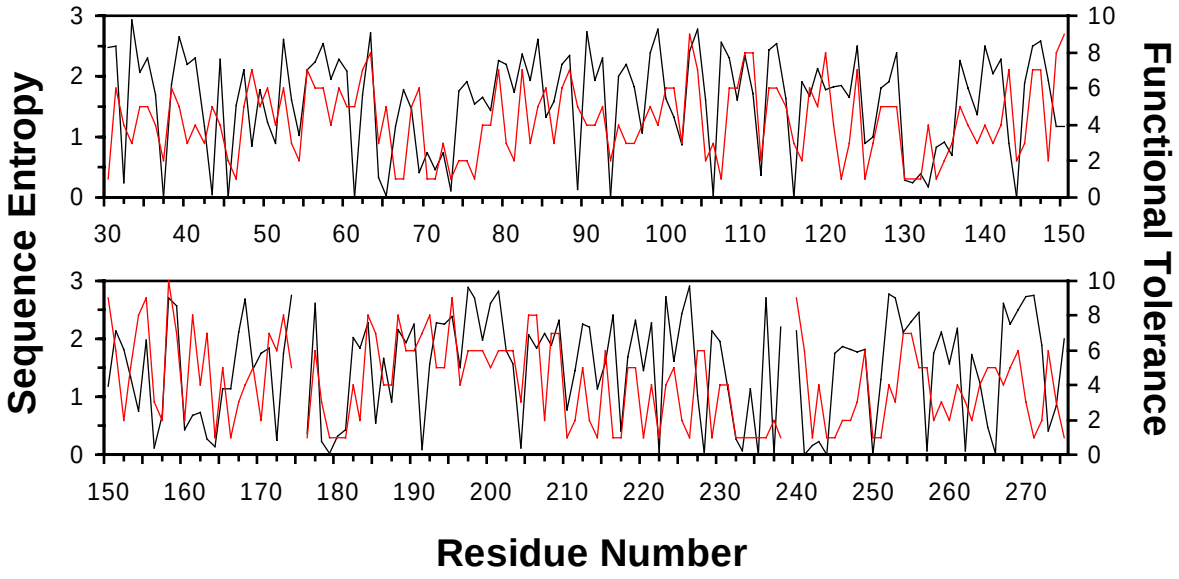

Sequence Entropy
Functional Tolerance
Residue Number
0
1
2
3
0
2
4
6
8
10
30
40
50
60
70
80
90
100
110
120
130
140
150
150
160
170
180
190
200
210
220
230
240
250
260
270

Figure 3-9 (C):

The site entropy (black line) is compared to the functional tolerance (red line) of two helices in λ repressor (Reidhaar-Olson and Sauer, 1988; Reidhaar-Olson and Sauer, 1990). The functionally acceptable mutants were discovered using a selection on which variants with 5-10% of wild-type activity survived. The entropy was calculated using the mean-field algorithm with a final temperature of T = 300 K.

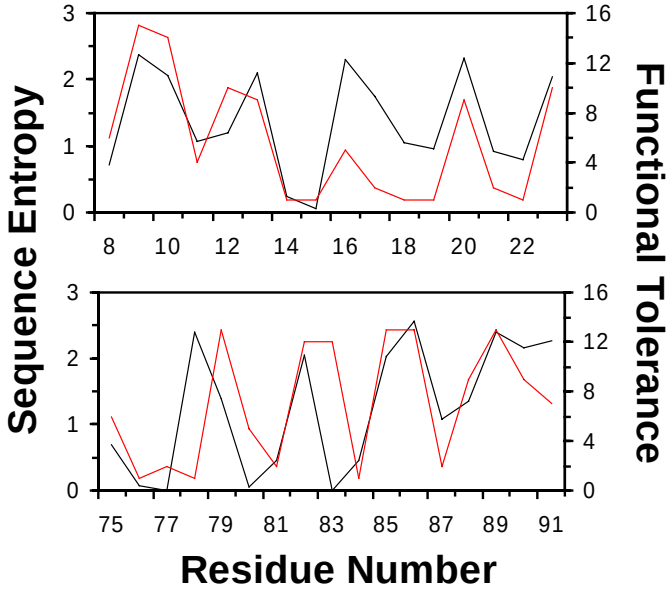
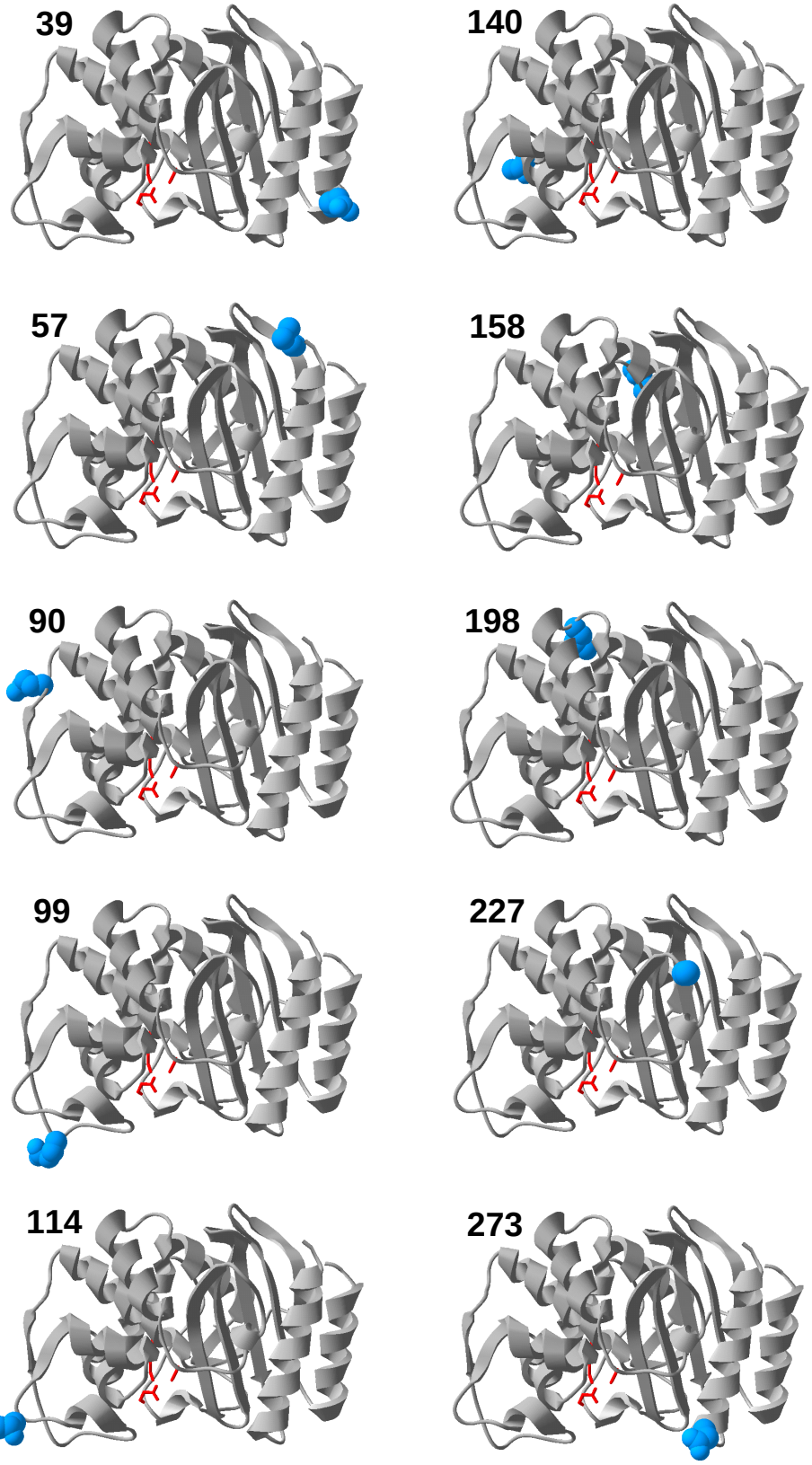

Sequence Entropy
Functional Tolerance
Residue Number
0
1
2
3
4
8
12
16
8
10
12
14
16
18
20
22
75
77
79
81
83
85
87
89
91

Figure 3-10:

The ten high-entropy β-lactamase residues chosen for site saturation are shown in blue. The amino acid identity and the improvement found at each residue are listed in Table 3-3. The three active site residues are shown in red.


39
140
57
158
90
198
99
227
114
273

Chapter 4

Tolerance of the CDRs of Antibody D1.3

The research presented in this chapter was done in collaboration with K. Dane Wittrup and Brenda Kellogg at the Massachusetts Institute of Technology. All experiments described herein were performed in the Wittrup laboratory.

Abstract

The tolerances of the complementarity determining regions (CDRs) of the antibody D1.3 interacting with hen egg white lysozyme (HEL) are calculated using mean-field theory and a structural model. The sequence entropy is recorded as a measure of the number of amino acids that can be introduced at a residue without disrupting the structural stability or binding interactions. Mutations discovered during the affinity maturation of antibody D1.3 are biased towards residues that have a high calculated sequence entropy. To test the predictive ability of the entropy calculation, a library was created by simultaneously mutagenizing four residues on the heavy chain that are computed to have high sequence entropy (V_H30, V_H56, V_H61, V_H62). While the library was enriched with functional mutants, the best mutant only has a modest 2-fold gain in affinity. In this chapter, we discuss potential improvements to our targeting strategy and use the ORBIT protein design tools to analyze alternative targeting strategies.

1. Introduction

Directed evolution often discovers beneficial mutations in surprising places. Rather than directly modifying the active sites of enzymes or the binding region of antibodies, mutations often occur far from these areas and affect catalysis or binding through subtle, long-range interactions (Spiller *et al*., 1999; Boder *et al*., 2000). Accumulating these mutations over rounds of selection has been very successful approach for improving industrially germane properties of enzymes (Petrounia and Arnold, 2000). However, the improvement of antibodies for pharmaceutical applications poses a difficult dilemma. As with enzymes, directed evolution experiments have yielded novel mutations throughout the frame of the antibody (Chen *et al*., 1999; Linden *et al*., 2000; Boder *et al*., 2000). While these mutations lead to improved binding properties, they can ultimately pose problems for using the modified sequences as pharmaceuticals. The immune system is particularly effective at removing these artificial antibodies by recognizing the amino acid mutations in the frame as foreign and inducing a rapid immune response. To circumvent this problem, mutations can be directed towards the complementary determining regions (CDRs), thus mimicking natural diversity.

Limiting the regions of diversity to the CDRs poses a conceptual difficulty for directed evolution. The binding region is more constrained than the scaffold of the protein so mutations in this region are more likely going to have deleterious, non-additive effects (Mackan and Perelson, 1989; Kauffman and Weinberger, 1989; Brown *et al*., 1996). Finding beneficial mutations in highly interacting regions is difficult for random mutagenesis methods when the mutation rate and screening capacity are small. Discovering mutations at interacting residues requires multiple mutations to collectively

generate beneficial effects. As the number of required mutations increases, the combinatorial possibilities grow exponentially and it becomes increasingly unlikely that beneficial combinations will be discovered.

Limiting mutagenesis to the CDR regions reduces the size of the search space, but it remains impossible to sample all of the possible combinations of mutations. It is useful to have a guide as to which residues should be mutagenized to maximize the probability of discovering affinity-improving mutations. In addition, a benefit of saturating a few residues is that the restrictions imposed by the genetic code can be avoided and all amino acids can be sampled at the chosen positions.

In this manuscript, we calculate the tolerance of each CDR residue of D1.3 complexed with HEL. The tolerance is determined by calculating the energetic effects of all amino acid substitutions at each position through a mean-field calculation, as described previously (Voigt *et al*., 2001). A library is produced experimentally by targeting four high entropy residues for simultaneous saturation mutagenesis. The resulting library is screened and found to be particularly tolerant to mutagenesis. Further, we test the idea that these predictions could be useful as a guide for saturation mutagenesis experiments, where the most tolerant residues should be targeted first. When the library of high-entropy positions were aggressively screened for improvements in HEL-binding, only a modest 2-fold improved mutant was found.

Several alternative targeting methods have been proposed, including the residues where improvements had been found previously by somatic mutagenesis or probing rounds of random mutagenesis (Miyazaki and Arnold, 1999; England *et al*., 1999). In addition, it has been proposed that there are hotspots in the DNA sequence encoding the

CDRs where improvements can be found independent of the specific antibody-antigen system (Chowdhury and Pastan, 1999). These strategies require different forms of information: the entropy calculation requires a structure, the somatic mutagenesis requires the germline sequence, and the consensus strategy requires an initial experimental step of mutagenesis and screening. Each of these strategies is examined using the ORBIT (Optimization of Rotamers by Iterative Techniques) protein design tools and the results are compared with the entropy calculation. We predict that the residues mutated by somatic mutagenesis are particularly tolerant, whereas the other sets of residues are more constrained by their environments.

2. Computational Methods

Various computational tools are applied to calculate the energetic effect of mutations, both on the stability of the antibody scaffold as well as the effect on the interactions between the antibody and HEL. First, a high-resolution crystal structure is used to obtain the backbone structure. At each residue, all amino acids are inserted and their flexibility is discretized into a set of conformationally distinct rotamers. The energetic interactions between all pairs of rotamers are then calculated using a force field. Finally, a single rotamer sequence is obtained by minimizing the energy function. Alternatively, mean-field theory is used to calculate the tolerance (sequence entropy) of each residue to amino acid substitutions.

This computational strategy is derived from a set of tools, collectively referred to as ORBIT, which has been used to solve the inverse folding problem (Dahiyat and Mayo, 1997; Malkauskas and Mayo, 1998). The goal of inverse folding is to design an amino

acid sequence that will fold into a target three-dimensional structure. The energy function is highly reliant on obtaining the correct packing by monitoring van der Waals interactions, as well as through the enforcement of a hydrophobic core and hydrophilic surface (binary patterning) (Marshall and Mayo, 2001). Additional crude H-bonding and electrostatic terms are included. This approach has been successful in designing thermophilic sequences for a variety of structures. However, it is unclear how well this energetic model will transfer to calculating the interactions between proteins. For example, electrostatics may dominate protein-protein interactions while side chain packing and hydrophobic burial may be less important.

2.1. D1.3 –HEL Interaction Thermodynamics

The D1.3–HEL complex has several characteristics that make it particularly suitable for protein engineering by directed evolution. First, the interface between the proteins is large and most of the residues have been found to be tolerant to amino acid substitutions (Ito *et al*., 1992; Fields *et al*., 1996; Braden *et al*., 1998). Alanine-scanning mutagenesis studies have demonstrated that alanines are not tolerated at only four of the contact residues, implying that these residues are critical for binding (England *et al*., 1997; Dall'Acqua *et al*., 1998). These essential residues tend to be hydrophobic amino acids on the V_H CDR2 and CDR3 and the V_L CDR1 and CDR3 variable loops. The energetic effect of amino acid substitutions drops as mutations are made further from this region (Pons *et al*., 1999). If the interactions were highly coupled and subtly propagated across many residues, as in the case of many antibody-antigen interfaces (Dall'Acqua *et*

al., 1996; Kobayashi *et al*., 1999), then this would be more difficult for directed evolution to discover beneficial mutants.

Several types of interactions contribute to the binding between D1.3 and HEL. An important component is the shape complimentarity between the interfaces (Novotny *et al*., 1989; Hawkins *et al*., 1993). The maximization of van der Waals interactions is therefore important for binding (Braden *et al*., 1996). Hydrogen bonding and favorable electrostatic contacts across the interface contributes strongly to the interaction energy (Fields *et al*., 1996). The hydrophobic effect also drives the binding, where interfacial residues that do not bury hydrophobic area have to compensate through a complementary electrostatic interaction (a salt bridge or hydrogen bond). Finally, the affinity can be improved by reducing the entropy of the side chains that participate in antibody-protein binding (Bhat *et al*., 1994; Fields *et al*., 1996). Mutations in the CDR regions can have a deleterious effect on the association between the V_H and V_L fragments (Yasui *et al*., 1994). Therefore, it is important when mutating the CDRs, that mutations not disrupt the stability of the V_H-V_L complex.

Beyond directly improving the interactions between D1.3 and HEL, mutations can alter the affinity by stabilizing the binding conformation of D1.3. The energetic benefit from shape complementarity is sufficiently strong to induce small rearrangements in the backbone to improve fit (Braden *et al*., 1996). However, there is an energetic cost to bend the backbone, so the affinity is decreased when this occurs. By stabilizing the antigen-bound form of the structure, the binding energy can be improved (Braden *et al*., 1996). Mariuzza and co-workers demonstrated this effect by comparing the binding of D1.3 with HEL and turkey egg white lysozyme (TEL). While the amino acid sequences of the two

lysozymes where nearly identical, a small conformational change was required to bind TEL, having the effect of reducing the binding by two orders of magnitude. This reduction is dominated by a slow association rate, implying that there is an energetic barrier created by the need for the free D1.3 to attain the correct conformation for binding. Thus, minimizing the conformational variability has the potential to improve binding.

2.2. Calculating the Interaction Energy

The calculations are based on the high-resolution crystal structure of the HEL-D1.3 complex (Bhat *et al.*, 1994) (Figure 4-1). A subset of residues' amino acid side chains are fixed in their wild-type conformations and the remaining residues are allowed to vary in identity and conformation. For example, the energy of an amino acid rotamer sequence $\{R\}$ threaded onto the CDRs is

$$E\{R\} = \sum_{i}^{N_{CDR}} E(i_r) + \sum_{i}^{N_{CDR}} \sum_{j>i}^{N_{CDR}} E(i_r, j_s) + \sum_{i}^{N_{CDR}} \sum_{j}^{N_{D1.3}} E(i_r, j_s) + \sum_{i}^{N_{CDR}} \sum_{j}^{N_{HEL}} E(i_r, j_s), \qquad (4\text{-}1)$$

where N_{CDR} is the number of CDR residues, $N_{D1.3}$ is the number of non-CDR residues in antibody D1.3, and N_{HEL} is the number of residues in hen egg-white lysozyme. The interaction energies between non-CDR and HEL residues do not vary in the calculation and are therefore not considered in Equation (4-1). The single-body term $E(i_r)$ captures the interaction between rotamer i_r with the carbon backbone. The energy between rotomers i_r and j_s, $E(i_r,j_s)$, is composed of van der Waals, electrostatic, and hydrogen-bonding terms (Gorden *et al.*, 1999). An additional pairwise energy term is included to calculate the solvation effect of burying hydrophobic surface area (Street and Mayo, 1998).

In our calculations, the backbone remains fixed. Therefore, its internal energy contribution is not relevant to the optimization procedure. This is often a limitation in inverse folding algorithms, but there is some evidence that the backbone of D1.3 does not significantly deviate during mutagenesis. Using the crystal structures of several mutant D1.3 antibodies bound to HEL, Mariuzza and co-workers showed that the effect of amino acid substitutions is dominated by local rearrangements of the side chains, rather than shifts in the backbone structure (Fields *et al*., 1996). In addition, there is minimal structural rearrangement upon binding to HEL (Bhat *et al*., 1990; Freire, 1999).

A limitation of our model is the inability to explicitly incorporate water molecules. Water molecules often mediate important interactions in antibody-antigen binding, both to propagate hydrogen-bonding and to improve shape complementarity (Bhat *et al*., 1994; Fields *et al*., 1996; England *et al*., 1997; Li *et al*., 2000). However, the ability for our model to capture improvements in the shape complimentarity, buried hydrophobic area, and hydrogen bonding should facilitate the discovery of alternate binding mechanisms. It has been proposed that direct hydrogen bonds between D1.3 and lysozyme are more energetically favorable than hydrogen bonds that are mediated by water molecules (Fields *et al*., 1996).

2.3. Rotamer Libraries

The amino acid conformations are discretized into a set of rotamers. The rotamer libraries used in this chapter contain conformations specific to the ϕ- and ψ-angles of each residue ("backbone-dependent") with several modifications that have been previously described (Dunbrack and Karplus, 1993; Dunbrack and Karplus, 1994;

Dahiyat *et al*., 1997). We use three variations on the backbone-dependent library. The e0 library is as described by Dunbrack and Karplus. The e2 library expands the χ1- and χ2-angles of all rotamers by one standard deviation. The a2h1p0 expands the rotamers corresponding to aromatic amino acids by two standard deviations, the rotamers corresponding to hydrophobic amino acids by one standard deviation, and the polar rotamers corresponding to polar amino acids are not expanded. In all of the libraries, the wild-type rotamer conformation at each residue is included. The e0 library contains the least number of rotamers while the e2 library contains the most rotamers. The a2h1p0 library is typically used to design the core of proteins. Increasing the number of rotamers has a detrimental effect on the calculations: dead-end-elimination is less likely to converge and the mean-field calculation becomes constrained by memory requirements.

2.4. Sequence Design of CDR residues

To examine the ORBIT's ability to optimize surface residues in the context of interacting with a second protein, we ran a full sequence design of the 61 CDR residues of HEL (Figure 4-1). In this calculation, the non-CDR and HEL residues are fixed in their wild-type amino acid side chain conformation. The remaining CDR residues are allowed to mutate to 17 amino acids (Met, Cys, and Pro are excluded) using the a2h1p0 library. All rotamers that exhibit one-body or pairwise energies above 100 kcal/mol are pruned from the rotamer list. Then, the dead-end elimination (DEE) algorithm is used to converge on the global optimum energy conformation of rotamers (Desmet *et al*., 1992; Goldstein, 1994; Gordon and Mayo, 1998; Pierce *et al*., 2000).

The solution to which DEE converges is shown in Table 4-1 and compared with the wild-type conformation in Figure 4-2. Only 31% of the amino acids are conserved, most of which are on the V_L CDR3 and V_H CDR1 loops. Mutations from polar amino acids to aromatics occur frequently. This typically occurs to facilitate the filling of void spaces that were previously occupied water, some of which were mediating H-bonding. Those residues that are participating in the different design strategies are shown in Table 4-2 ("big"). The consensus residues are the most conserved (3/4) whereas the hotspot and high-entropy residues were less conserved (2/4 and 1/4) and none of the somatic residues are conserved.

A second DEE calculation was run that focuses on the changes incurred by saturating the sets of targeted residues in the background of the wild-type CDR amino acid sequence (Figure 4-3). For this calculation, all of the residues except for the four designed positions are fixed in their wild-type identity and conformation. The designed residues are allowed to vary using the e2 rotamer library and the minimum conformation is obtained using DEE. The results of this calculation tend to be more conservative than the full sequence design (Table 4-2). All of the consensus residues remain unmutated and the somatic and high-entropy residues are mutated to amino acids with properties more similar to wild-type.

2.5. Structural Tolerance of CDR Residues

To calculate the structural tolerance of the CDR residues, the non-CDR and HEL amino acid side chains are fixed in the wild-type conformation and the amino acid identity and side chain conformation of the CDR residues are varied using the mean-field

algorithm. Since the non-CDR residues and the HEL residues are confined to a single rotamer, the mean-field treatment yields the energy of rotamer *r* at position *i*,

$$E_{mf}(i_r) = \sum_{j \neq i}^{N_{CDR}} \sum_{r}^{R_i} \sum_{s}^{R_j} E(i_r, j_s) p(i_r) p(j_s) + \sum_{j}^{N_{D1.3}} \sum_{r}^{R_i} E(i_r, j_s) p(i_r) + \sum_{j}^{N_{HEL}} \sum_{r}^{R_i} E(i_r, j_s) p(i_r) \quad (4\text{-}2)$$

where R_i is the total number of rotamers at residue *i*, and $p(i_r)$ is the probability that rotamer *r* exists at residue *i*. The probabilities can subsequently be calculated by

$$p(i_r) = \frac{e^{-\beta E_{mf}(i_r)}}{\sum_{s}^{R_i} e^{-\beta E_{mf}(i_s)}}, \quad (4\text{-}3)$$

where β is the Boltzman temperature. As the temperature is lowered, the probabilities become more skewed towards a few dominating amino acids. A threshold temperature defines a mean-field energy below which sequences are stable in the D1.3-HEL complex whereas sequences above this energy are unstable (Figure 3-1).

Before the mean-field minimization algorithm is run, all of the pairwise energies between rotamers required by Equation (4-2) are calculated. The e0 rotamer library is used and all twenty amino acids are allowed at each residue. From this initial list, all of those rotamers that interact with the protein backbone with energies greater than 5 kcal/mol are eliminated from the calculation. An average of 121 rotamers per residue survive this step, corresponding to 3.2×10^4 one-body energies and 5.1×10^8 pairwise energies. Next, the mean-field algorithm is started by initializing all of the rotamer probabilities $p(i_r)$ to $1/R_i$. A high initial temperature is set (100,000 K) and then the temperature is lowered in increments of 100 K until the final temperature of 500 K is reached. After each temperature decrease, Equations (4-2) and (4-3) are iterated until self-consistency is achieved.

After the mean-field minimization, the amino acid probabilities can be defined as the sum of the rotamer probabilities for that amino acid at a given position. The tolerance of each CDR residue can be described by the sequence entropy

$$s_i = -\sum_a^A p(i_a) \ln p(i_a) \ , \qquad (4\text{-}4)$$

where $p(i_a)$ is the probability of amino acid a at residue i. A residue with a high entropy is tolerant to amino acid substitution, whereas a low-entropy residue is intolerant. The entropy of each CDR residue is shown in Figure 4-4 and is mapped onto the structure in Figure 4-5. As expected, the center of the D1.3-HEL interface tends to be less tolerant than the surrounding residues (Figure 4-6). While there are some characteristics that are shared between the interface and the core of a protein, the extent of this analogy is unclear. To study the effect of solvation on the tolerance, the entropy of each residue is calculated with and without an energetic term that accounts for the burial of hydrophobic surface area. The results of this comparison are shown in Figure 4-7.

3. Results and Discussion

3.1. Somatic Mutagenesis

The primary immune response, representing the result of recombining the V, D, and J genes, yields antibodies with generally low affinities. It is the secondary response, involving point mutation and selection, referred to as somatic hypermutation or affinity maturation, which generates antibodies with the required physiological affinity (Neuberger and Milstein, 1995). The primary response generates a B-cell clone, which is then subjected to mutagenesis concentrated in the variable region. The mutagenesis rate has been estimated to be ~10^{-3}–10^{-4} per generation (Berek and Milstein, 1987; Neuberger

and Milstein, 1995). This process leads to the production of offspring B-cells, which are then selected for improved binding. This process is repeated until adequate affinity is obtained. The mechanism of somatic mutation is very similar to directed evolution. A low mutation rate is employed, and a finite number of mutants are "screened." Due to these limitations, it is possible that somatic mutagenesis follows the same trend towards mutating discovering beneficial mutations at residues with high sequence entropy, as was predicted for directed evolution (Voigt *et al*., 2001).

The structural effects of somatic mutations have been explored. It has been observed that the replacement of non-contact residues occurs frequently (Chien *et al*., 1989; Sharon, 1990; Bhat *et al*., 1994; Patten *et al*., 1996). A study of the somatic mutagenesis of the NQ10/12.5 antibody raised against a hapten (2-phenlyoxazolone) demonstrated that the contact residues were rarely mutated (Spinelli and Alzari, 1994). Further, by comparing the structures of germline and affinity-matured antibodies, it has been demonstrated that most somatic mutations do not significantly rearrange the binding hole in antibody-hapten complexes (Orencia *et al*., 2000). Many of the mutations in this system were found to occur far from the binding site and the mechanism by which affinity was approved was attributed to long-range effects. These studies demonstrate that the structure of the binding site, as well as the specific pattern of antibody-antigen binding interactions are often preserved by somatic mutations (Spinelli and Alzari, 1994; Orencia *et al*., 2000).

Five somatic mutations occur during the affinity maturation of D1.3 from germline. Bedoulle and co-workers determined the contribution of each mutation to the overall 60-fold increase in affinity (England *et al*., 1999). In addition, the mutations were

shown to be additive. Each mutation independently improved the interaction with lysozyme and the order that these mutations were made was inconsequential. Four non-silent somatic mutations in the CDR regions were identified: V_L 50 N→Y, V_L 51 A→T, V_L 52 K→T, and V_H 56 S→N. The amino acid substitutions at residue V_L 50 and V_H 56 most contribute to the improvement in HEL binding (England *et al*., 1999). Three out of four of these mutations occur at high-entropy positions (Figure 4-4). These residues form the basis for the somatic targeting strategy.

3.2. Mutagenesis of Selected Residues

Based on each of the design strategies, libraries were experimentally constructed corresponding to the simultaneous mutagenesis of all four targeted residues. To create the library, a forward primer is designed with a NN(G/T) codon for each residue that is being mutagenized. A reverse primer is also designed to overlap with each forward primer. After PCR, sets of DNA fragments are generated for each target residue, which are then being pieced together using the SOEing procedure to produce the full-length gene (Horton, 1995). The gene was then inserted into the pCT303 display vector via homologous recombination in yeast. The randomization of the NN(G/T) codons were confirmed by sequencing.

The antibody libraries were displayed on yeast, and the mutants were analyzed by fluorescence-activated flow cytometry (FACS). Between two to four conservative rounds of sorting were initially performed where those clones with affinities near wild-type were isolated. This enriched library was then subjected to several rounds of aggressive sorting. The aggressive sorting is based on the off rate of the D1.3-HEL interaction, which is

measured by allowing the mutant antibodies to bind a labeled HEL and then incubating with unlabeled HEL. The top 0.1% of the clones was isolated after each sorting round. When the population converges, the remaining clone(s) is (are) analyzed as the highest affinity mutants in the library.

The libraries corresponding to the four design strategies were created and analyzed (Table 4-3). The fraction of each library with affinities above 100nM and the highest-affinity mutant discovered are reported. The library based on the sequence entropy calculation is very enriched in functional mutants (13.3%), but the best mutant found in the library had only a 2-fold increase in affinity. The somatic and hotspot libraries have 3.4% and 0.4% functional mutants, respectively. The library that has surprising characteristics is the one constructed via the consensus strategy. While these positions are buried in the core of the antibody and D1.3-HEL interface and are predicted to have low sequence entropy, they are found to be tolerant to substitutions (4.8% are functional). Further, in this library, a mutant with 20-fold improvement in affinity was found - the best of any of the libraries.

Our inability to calculate the tolerance of the consensus residues may reflect several difficulties in our algorithm. It is possible that mutations at these positions may induce structural rearrangements or have some other affect that is not explicitly described in our energy function. The loops of the CDR region may have more flexibility than the typical structural targets of inverse folding algorithms. A related problem may be in understanding the difference between interactions that stabilize a protein structure (in particular in the core of the protein) and those that are important for improving affinity.

The original motivation for the entropy algorithm is to be to able predict a set of residues that may be good targets individually for saturation mutagenesis. There may be a set of different constraints on mutagenizing multiple interacting residues simultaneously. By targeting several residues, the possibility for compensating mutations arises and the entropy calculation becomes less accurate. To overcome this limitation, we are developing of algorithms that can sort through sets of multiple residues to discover the optimal one for mutagenesis. While this will be a useful tool for protein engineering, it still is unlikely to explain the tolerance of the consensus residues, as the partial and full sequence designs failed to make amino acid substitutions at these residues (Table 4-2).

While we have observed that beneficial mutations tend to occur at residues with high sequence entropy (Chapter 3), it may not be the best strategy to optimize this value when targeting residues. For example, while there are high entropy residues close to the center of the binding interface of D1.3, when the entropy is maximized, the residues that are chosen are far from this region (Figure 4-6). Somatic mutagenesis found beneficial mutations at the high-entropy residues that are relatively close to the interface. In designing a targeting strategy, it may be required to maximize multiple constraints. Chowdhury and Pastan took this approach when they formulated the hotspot strategy (Chowdhury and Pastan, 1999). Many residues were predicted to be hotspots, but they chose those that were closest to the binding interface.

4. Conclusions

Four design strategies are analyzed in this chapter using the mean-field entropy algorithm and other ORBIT design tools. Those residues that have the highest sequence

entropy are chosen to undergo simultaneous mutagenesis. The resulting library was significantly more enriched in functional mutants than the other design strategies. However, when compared to the other libraries, the smallest improvement in affinity was found. Improvements could be made to this strategy by assessing the difficulty in calculating protein-protein interactions, the differences in single- and multiple-residue mutagenesis, and introducing multiple optimization constraints.

Table 4-1: Full ORBIT CDR sequence design

Light Chain																	
CDR1		24	25	26	27	28	29	30	31	32	33	34					
	wt	R	A	S	G	N	I	H	N	Y	L	A					
	D	Y	-	T	N	D	-	Y	H	F	I	-					
CDR2		50	51	52	53	54	55	56									
	wt	Y	T	T	T	L	A	D									
	D	F	D	W	Q	R	E	W									
CDR3		89	90	91	92	93	94	95	96	97							
	wt	Q	H	F	W	S	T	P	R	T							
	D	E	E	-	-	-	W	-	E	Y							
Heavy Chain																	
CDR1		26	27	28	29	30	31	32	33	34	35						
	wt	G	F	S	L	Y	G	Y	G	V	N						
	D	-	-	N	-	-	E	-	A	-	A						
CDR2		50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65
	wt	M	I	W	G	D	G	N	T	D	Y	N	S	A	L	K	S
	D	-	-	-	-	E	A	R	R	Y	L	-	Q	W	Y	R	N
CDR3		98	99	100	101	102	103	104	105								
	wt	E	R	D	Y	R	L	D	Y								
	D	L	F	V	F	F	A	-	-								

Table 4-2: ORBIT design of targeted residues

	Consensus				Hotspot			
	L34	L91	L93	H103	L25	L26	L33	L34
wt	A	F	S	L	A	S	L	A
big	-	-	-	A	-	T	I	-
small	-	-	-	-	-	R	-	T

	Somatic				Entropy			
	L50	L51	L52	H56	H30	H56	H61	H62
wt	Y	T	T	N	Y	N	S	A
big	F	D	W	R	-	R	Q	W
small	F	S	S	Q	-	Q	Q	N

Table 4-3: Comparison of Experimental and Computational Results

	Consensus	Hotspot	Somatic	Entropy
% functional[a]	4.8	0.4	3.4	13.3
maximum[b]	0.1	0.6	-	1.1
<entropy>[c]	0.3	0.9	1.7	2.3
ΔE_{max}[d]	-5.4	-8.0	-10.6	-10.1

a. The percent of the library that is experimentally determined to be functional, with 100nM or better binding as a criterion for functionality

b. The K_d of the best mutant in the library, in units of nM. Wild-type D1.3 has K_d = 2.4 nM.

c. The average entropy of the four residues targeted in each meathod, as determined using the mean-field algorithm (Figure 4-4).

d. The change in energy between the wild-type and GMEC sequence, as determined using the ORBIT partial sequence design. The units are in kcal/mol.

Figure 4-1:

The structure of antibody D1.3 bound to HEL (Bhat *et al*., 1994). The blue and green structures are the heavy and light chains and the black structure is the bound HEL. The CDR regions are shown in red.

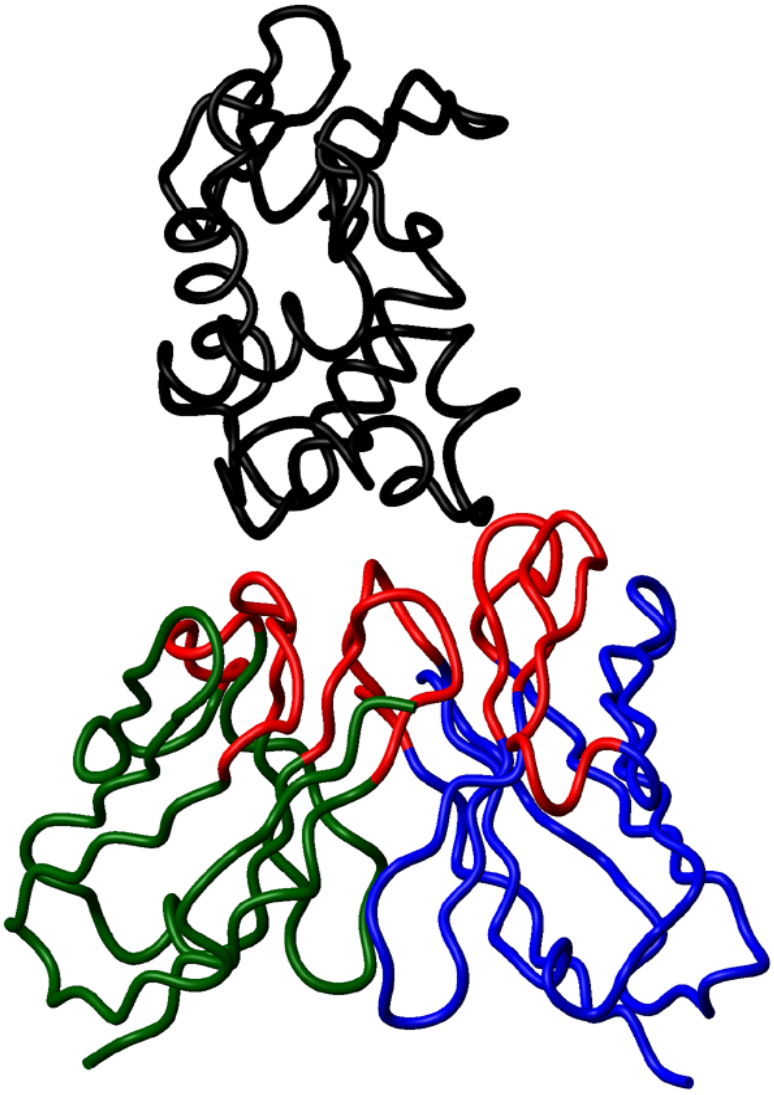


Figure 4-2:

The wild-type (blue) and designed (red) amino acid sequences for the CDR region of D1.3, as shown from the perspective of the bound HEL. The full sequence design was performed by allowing all of the CDR residues to vary simultaneously in amino acid identity while holding the non-CDR residues in D1.3 and HEL in their wild-type identity and conformation. The energy is then minimized using dead-end elimination. The sequence obtained from the full design tends to be more tightly packed, notably by aromatics. The core of the binding region is more highly conserved in the calculation than the surrounding residues. The amino acids chosen by the full sequence design are listed in Table 1.

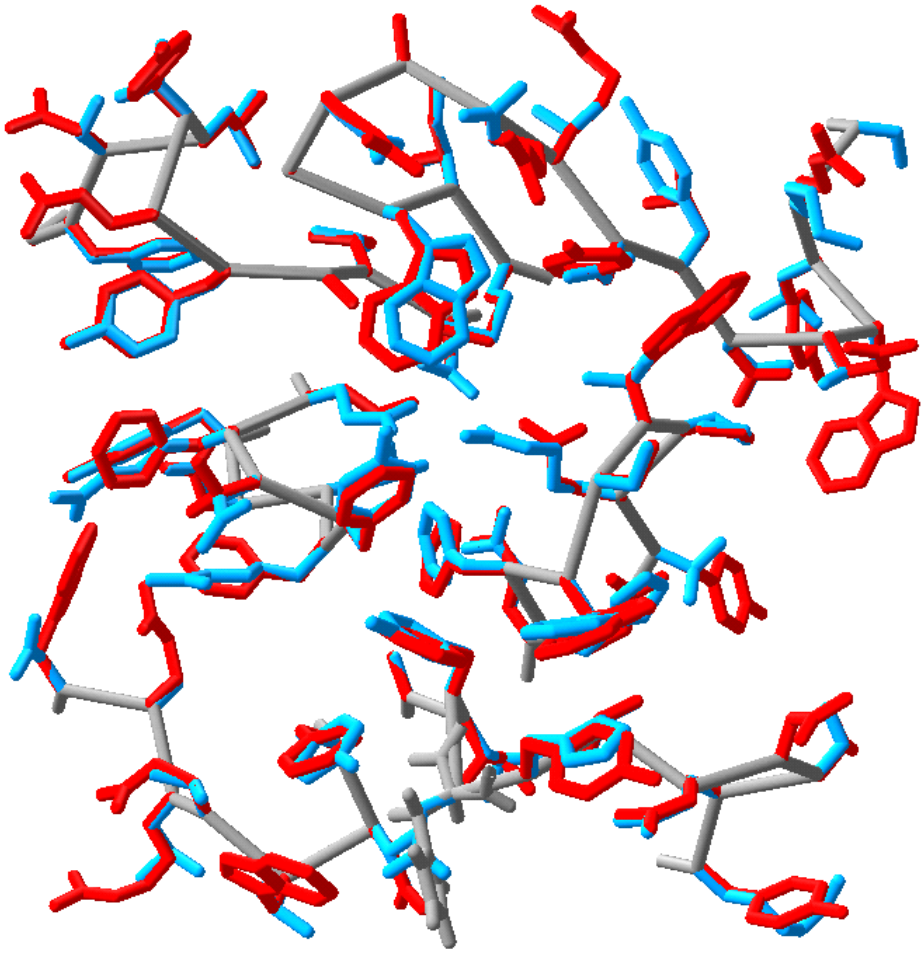
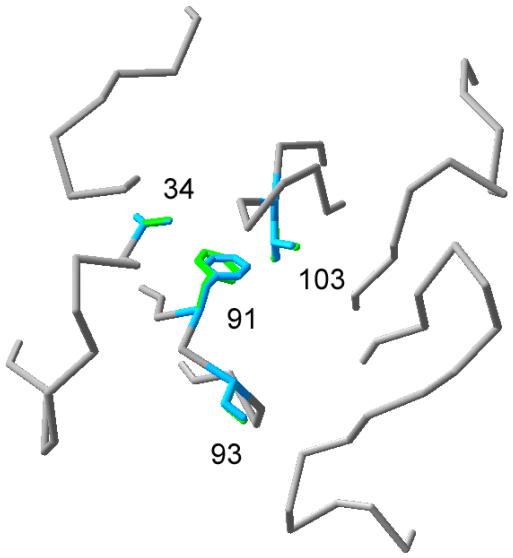
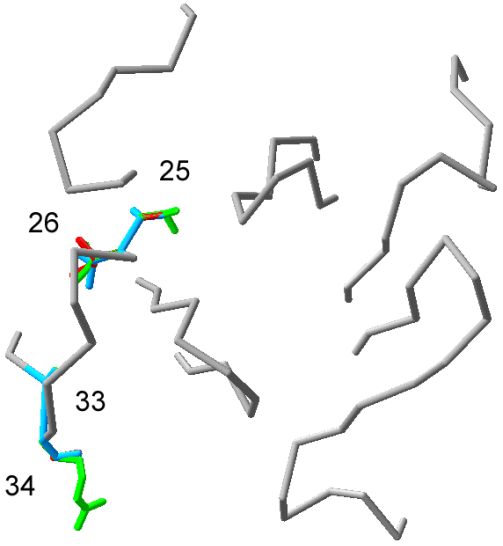


Figure 4-3:

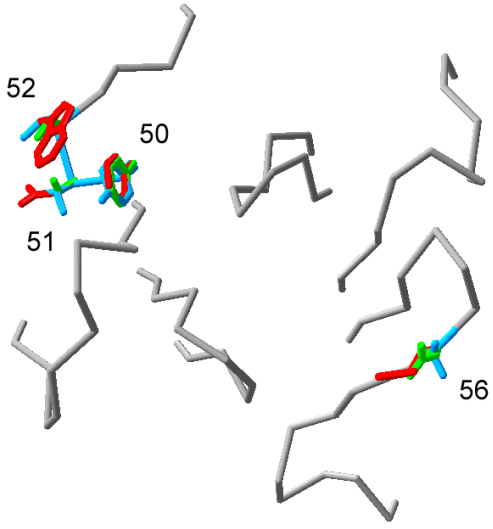
The wild-type amino acid identities (blue) are compared with those identified by the full sequence design (red) and the limited sequence design (green). The heavy and light chain CDRs are shown from the perspective of bound HEL. The limited sequence design was performed by allowing the set of four chosen residues to vary in amino acid identity while holding the remaining residues of D1.3 and HEL in their wild-type identity and conformation. The energy was then minimized using dead-end elimination. This calculation was performed on the residues chosen by the **(A)** consensus, **(B)** hotspot, **(C)** somatic, and **(D)** sequence entropy strategies. The amino acid identities converged on by this calculation are listed in Table 2.



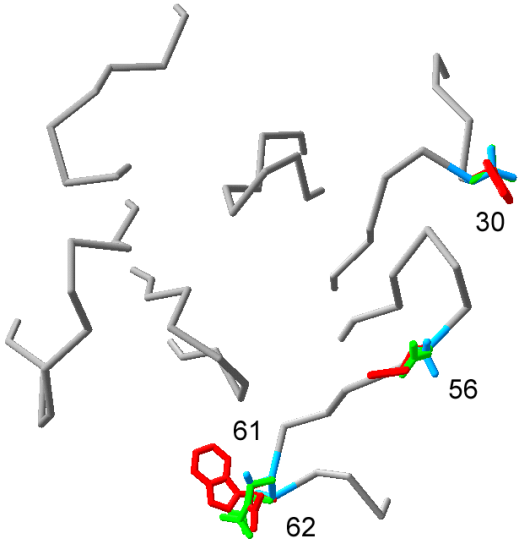
(A)



(B)



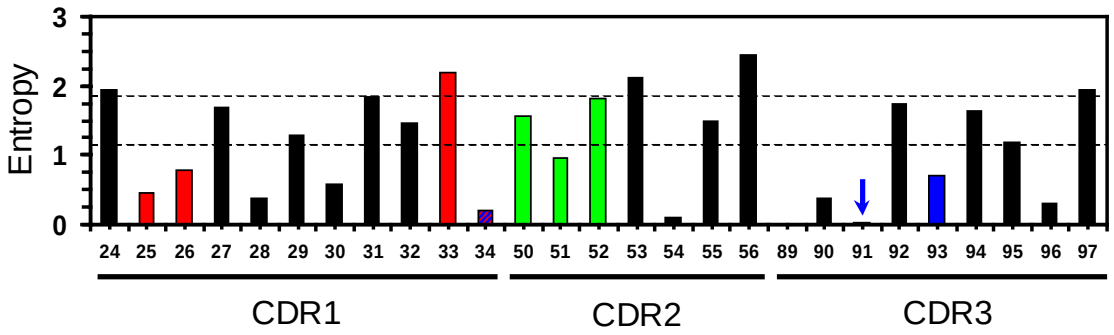
(C)



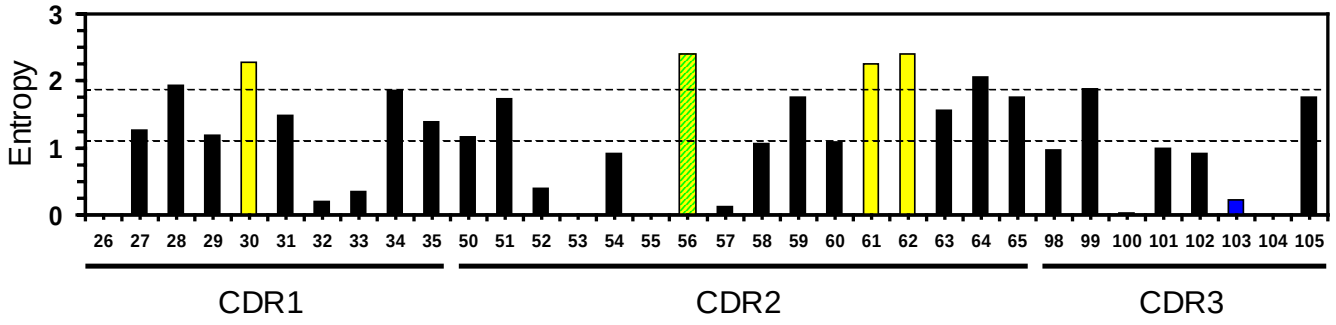
(D)

Figure 4-4:

The entropy of each CDR residue in the light **(A)** and heavy **(B)** chains are shown. The dotted lines mark the mean entropy and one standard deviation above the mean. The residues corresponding to the four design strategies are color-coded: (blue) consensus, (green) somatic, (red) hotspot, and (yellow) sequence entropy. Residues L34 and H56 are shared between two strategies.


Entropy
3
2
1
0
24 25 26 27 28 29 30 31 32 33 34 50 51 52 53 54 55 56 89 90 91 92 93 94 95 96 97
CDR1
CDR2
CDR3

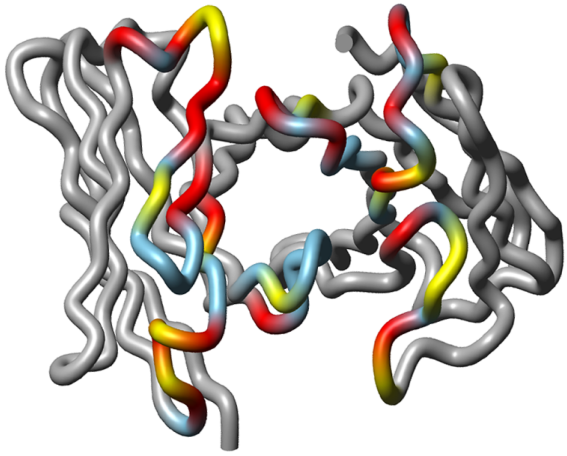
(A)


Entropy
3
2
1
0
26 27 28 29 30 31 32 33 34 35 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 98 99 100 101 102 103 104 105
CDR1
CDR2
CDR3

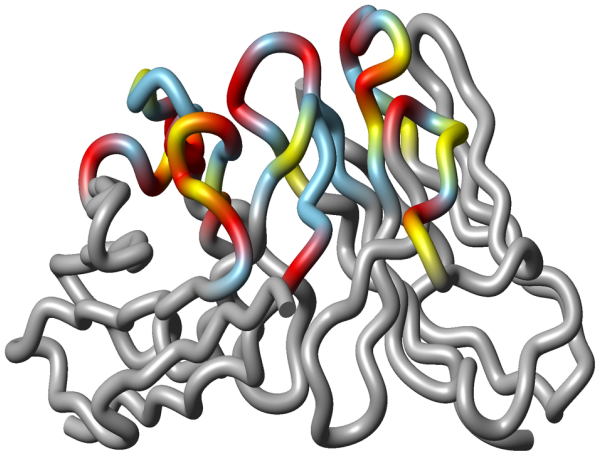
(B)

Figure 4-5:

The entropy of each CDR residue is mapped onto the D1.3-HEL structure (Bhat *et al.*, 1994). Two views are shown: **(A)** from the perspective of HEL (the center of the image is the center of the D1.3-HEL interface) and **(B)** a side perspective. The high entropy residues are shown in yellow (greater than one standard deviation above the mean) and the above average entropy residues are shown in red (greater than the mean). The low-entropy CDR residues are blue. There is a trend for conserved sites to be near the center of the binding interface (Figure 4-7).



(A)



(B)

Figure 4-6:

The sequence entropy is plotted against the distance from the center of the D1.3-HEL interface. The distance is measured from the C_β of V_H Tyr101, the approximate center of the interface. Residues further from the center tend to have higher sequence entropies, but the overall correlation is weak. When residues are chosen that maximize the sequence entropy, these tend to be a peripheral locations, as are the four positions chosen by the entropy method (red points). It may be advantageous to introduce a second constraint when optimizing the library, such as choosing high entropy residues that are participating in the interaction (marked with the black box).

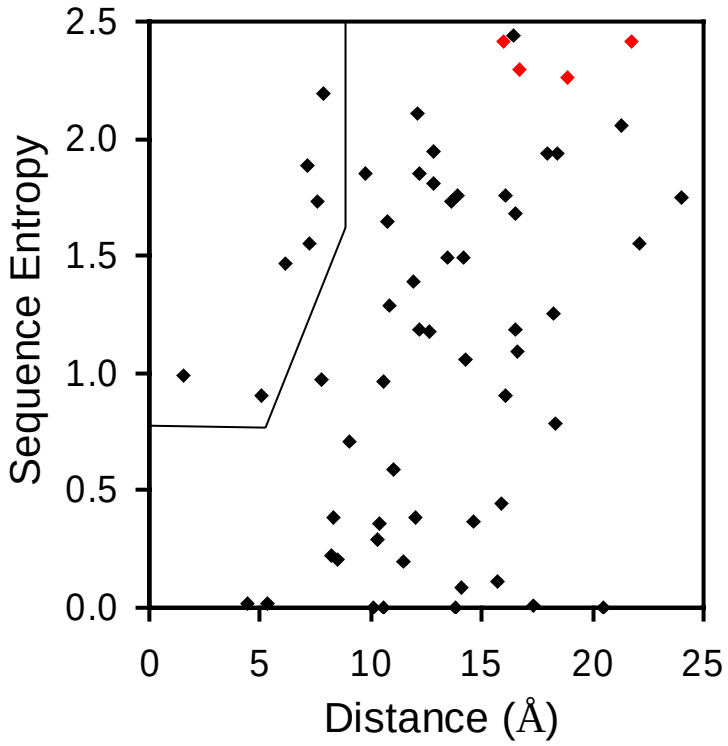
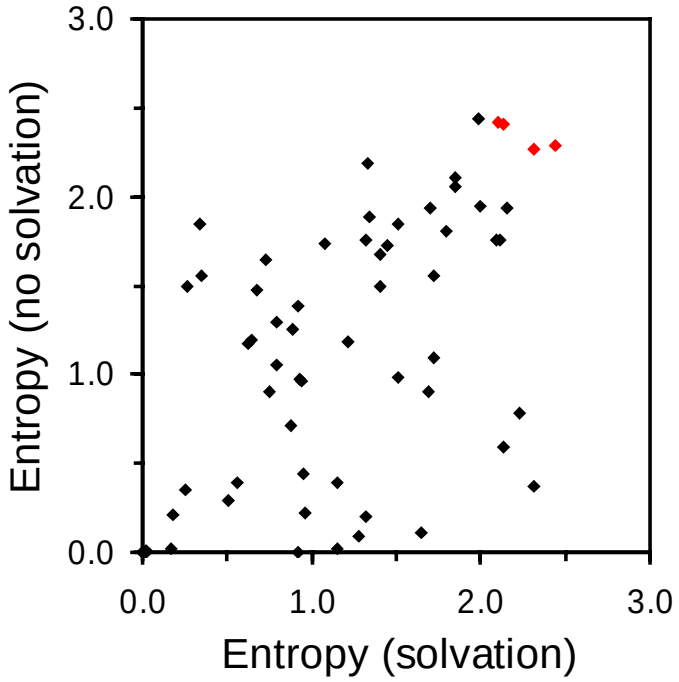

Sequence Entropy
2.5
2.0
1.5
1.0
0.5
0.0
0
5
10
15
20
25
Distance (Å)

Figure 4-7:

The sequence entropy of the CDR residues calculated with and without an energetic term for solvation. The solvation term is a large contribution to the energy, so to equate the average entropy for each dataset, the data with solvation is shown for T = 800 K and the data without solvation is shown for T = 500 K. The residues chosen for mutagenesis (30, 56, 61, 62) are indicated by the red points. It is noteworthy that they are predicted to have a high entropy using both energy expressions.


3.0
2.0
1.0
0.0
Entropy (no solvation)
0.0
1.0
2.0
3.0
Entropy (solvation)

Chapter 5

Protein Building Blocks Preserved by Recombination

Portions of this chapter are reproduced from:
Voigt, C. A., Martinez, C., Wang, Z-G., Mayo, S. L., and Arnold, F. H. (2002). Recombination Preserves Protein Building Blocks, Nature Struct. Biol. 9, 553-558

Abstract

Borrowing concepts from the schema theory of genetic algorithms, we have developed a computational algorithm to identify the fragments of proteins, or schemas, that can be recombined without disturbing the integrity of the three-dimensional structure. When recombination leaves these schemas undisturbed, the hybrid proteins are more likely to be folded and functional. Crossovers found by screening libraries of several randomly shuffled proteins for functional hybrids strongly correlate with those predicted by this approach. Experimental results in the construction of hybrids of two β-lactamases sharing 40% amino acid identity demonstrate a threshold in the amount of schema disruption that the hybrid protein can tolerate. To the extent that introns function to promote recombination within proteins, natural selection would serve to bias their locations to schema boundaries.

1. Introduction

In vitro recombination is a powerful tool for the tuning and optimization of proteins. It promotes the combination of traits from multiple parents onto a single offspring, thus exploiting information obtained in previous rounds of selection (Holland, 1975; Stemmer, 1994; Crameri *et al*., 1998). Recombination plays a key role in the natural evolution of proteins, notably in the generation of diverse antibodies, synthases, and proteases (Ostermeier and Benkovic, 2000). In these examples, crossovers occur at well-defined domain boundaries. The role of recombination in evolution is less well understood when the domain structure of a protein is not obvious. Here, we introduce a computational algorithm to divide a protein structure into pieces that can be swapped by recombination and compare the predictions with data generated by *in vitro* recombination experiments.

Ever since the first protein structures were elucidated, researchers have attempted to divide their otherwise complicated topologies into well-defined domains, defined variously as secondary structure units, structural elements that fold independently, or clusters of residues close in geometric space (Rossman and Liljas, 1974; Crippen, 1978; Rose, 1979; Gō, 1981; Zehfus and Rose, 1986; Holm and Sander, 1994; Panchenko *et al*., 1996; Tsai *et al*., 2000). An operationally relevant domain definition is a protein fragment that can be swapped among related structures. The locations of certain types of introns were shown to occur at structural domain boundaries, suggesting that larger proteins are composed of smaller domains discovered earlier in evolution and pieced together by gene duplication and recombination (Gō, 1981; Gō, 1983; De Souza *et al*., 1996; Gilbert *et al*., 1997). Using *in vitro* recombination experiments to observe that a crossover is

acceptable, rather than inferring it from the existence of introns, provides a direct approach to understanding how domains can be interchanged to create new, functional proteins.

2. Materials and Methods

2.1. The SCHEMA Algorithm

It has been previously suggested that the optimal recombination points allow swapping of structural domains (Ranganathan *et al*., 1999; Bogarad and Deem, 1999; Ostermeier and Benkovic, 2000; Riechmann and Winter, 2000). The difficulty has been to identify what these smaller building blocks look like. Research in computer science has demonstrated that the optimal crossover locations in genetic algorithms correspond to those that retain and combine clusters of bits that interact favorably (a "schema") (Holland, 1975; Forrest and Mitchell, 1993; Mitchell, 1996). Solutions in which recombination divides a schema such that an offspring inherits fractions of it from different parents are generally less fit. To identify the equivalent of schema in proteins, we introduce a computational algorithm, SCHEMA, which can predict fragments that must be inherited from the same parent. The schemas will therefore be the building blocks from which novel proteins can be assembled by recombination.

SCHEMA works by calculating the interactions between residues and then determining the number of interactions that are disrupted in the creation of a hybrid protein. A disruption occurs when an interaction is broken due to different amino acids being inherited from each parent (Figure 5-1). In the simplest implementation, two residues are considered interacting if any of their atoms (excluding hydrogen atoms) are

within a cutoff distance d_c = 4.5 Å, corresponding to 5–8 interactions per residue. Ideally, an algorithm would search all possible crossover combinations and determine the associated disruption for each (Appendix E). Analyzing multiple crossovers by this method leads to combinatorial difficulties, both in the calculation and the visualization of the data. SCHEMA overcomes this limitation by scanning the protein structure with a defined window size. Calculating how many interactions are disrupted when a crossover is made generates a schema profile *S*; if *S* is large for residue *i*, then the residue is involved in a more compact schema. Crossovers that correspond to minima of the schema profile preserve the maximum number of internal interactions and are therefore favored.

2.2. Calculating the Schema Profile

The schema disruption of a hybrid protein is the number of interactions that are broken when a certain pattern of fragments is inherited from each of the parents. If a hybrid protein is constructed from two parents where fragment(s) α is (are) inherited from parent A and fragment(s) β is (are) inherited from parent B, then the disruption $E_{\alpha\beta}$ of this hybrid can be calculated by

$$E_{\alpha\beta} = \sum_{i\in\alpha}\sum_{j\in\beta} c_{ij} P_{ij} \ , \qquad (5\text{-}1)$$

where c_{ij} = 1 if residues *i* and *j* are within distance d_c, otherwise c_{ij} = 0. Equation (5-1) calculates the exact disruption caused by a particular hybrid construction (e.g., Table 5-1 and Figure 5-6). The probabilities P_{ij} account for the fact that there is no disruption if the amino acid identities of the residue pair *i*, *j* in the set of potential hybrids are the same as in any of the parents (see Section 2.3).

Equation (5-1) can be used to calculate the disruption of any particular hybrid construction. However, when analyzing data from *in vitro* recombination experiments, the number of possible hybrid combinations prohibits the calculation of the disruption of all possible hybrids and the condensation of this information into a useful format. In order to compare recombination results with the schema disruption theory, we have developed an algorithm that searches for the most likely regions for crossovers to be non-disruptive.

The inputs into the SCHEMA program are the coordinates of the three-dimensional structure and an alignment of the parental sequences. The structure of only one parent is required under the assumption that for *in vitro* recombination to be successful the parents must share similar structures. A window of residues w is defined and the number of internal interactions within this window is counted. In choosing the window size, the assumption is made that the probability that two or more crossovers occurring in the window is very small. The window is then slid along the protein structure and a profile is generated where the schema profile of each residue in the window is incremented by the amount of disruption created by a crossover in that region. The numerical value of the schema profile function S at residue i is defined by

$$S_i = \frac{1}{\sqrt{w}} \sum_{j=i-w+1}^{i} \sum_{k=j}^{j+w-2} \sum_{l=k+1}^{j+w-1} c_{kl} P_{kl} \, . \qquad (5\text{-}2)$$

If a residue has a large S_i, then it is likely to be participating in a compact schema. A low S_i indicates that a crossover is likely to be tolerated at that position. In other words, the crossovers at regions of high S_i are more likely to create fragments with large $E_{\alpha\beta}$. For all of the calculations presented in this chapter, the parameters are d_c = 4.5 Å and w = 14

residues. The topology of the profiles is robust to the specific values of these parameters (Section 3.3).

2.3. Calculating the Probabilities

Equations (5-1) and (5-2) require the probability P_{ij} that a hybrid protein will have a combination of amino acids not present in the parents. To calculate this probability, a list is generated of all possible amino acid combinations that can occur in the hybrids. From this set, those combinations are removed that are present in the parents. To get the final probability, the number of unique combinations is divided by the total number of combinations, $p(p\text{-}1)$, where p is the number of parents. An example of this calculation is shown in Figure 5-2.

2.4. Sequence Alignments

Sequence alignments were performed using the BLAST algorithm with the BLOSUM 62 similarity matrix and open gap/extension gap penalties of 11/1. For the data sets in this study, the sequence identity between the parents is greater than 60%, reducing the ambiguity of the alignment. For the β-lactamase TEM-1/PSE-4 system (40% identity), the availability of both structures made a structural alignment possible (using the SwissProt software package).

2.5. Recombination Data Sets

2.5.1. Cephalosporinase

The schema profile was calculated based on the structure of cephalosporinase (Figure 5-3A) (Lobkovsky *et al*., 1993). In a random recombination experiment, several crossovers led to improved moxalactam antibiotic resistance (Crameri *et al*., 1998). Further, an experiment was performed by Levesque and co-workers where a fragment was taken from the β-lactamase TEM-1 gene and inserted into the PSE-4 gene[21]. The resulting hybrid protein was found to have wild-type activity towards various antibiotics.

2.5.2 Subtilisin Families

Minshull and co-workers recombined a set of 26 subtilisin genes by DNA shuffling and screened the recombinant mutants for improved thermostabilitity, high and low pH activity, and activity in organic solvent (Ness *et al*., 1999). When aligned, the 26 genes fall into four well-defined families. Within each family, the genes have approximately 99% sequence identity. Crossovers between parents that have this high degree of sequence identity are impossible to analyze by schema disruption. However, the sequence identity between parents from different families ranged from 76 to 90%. It is possible, then, to compare the crossovers between families with the schema profile. In the experiments, crossovers were allowed in the region between residues 60 and 224. The remaining portions of the sequence (1-60, 224-269) were taken from the Savinase gene. The structure of Savinase was used to calculate the schema profile (Figure 5-3B) (Betzel *et al*., 1992). Nearly all of the sequences of the 26 parental genes are unavailable. To overcome this, we ran a BLAST search, and selected a *Bacillus halodurus* serine protease

(SwissProt P41363), which is 65% identical to the Savinase sequence. The probabilities required by Equations 5-1 and 5-2 were estimated based on an alignment of these two sequences.

2.5.3. Cytochromes P450

A recombination experiment was performed on two P450c17 genes (rat and human), sharing 68% sequence identity, and a variety of functional hybrid proteins were discovered (Brock and Waterman, 2000). The structure of c17 is unknown, however, a structure of a homologous mammalian membrane-bound P450 2C5 has been solved (Williams *et al*., 2000). The equivalent locations for the crossovers were determined by aligning the parental sequences used in the experiment with the 2C5 sequence.

2.5.4. Glycinamide Ribonucleotide Transformylase

Benkovic and co-workers recombined PurN and GART glycinamide ribonucleotide transformylase, and functional hybrid proteins were selected (Ostermeier *et al*., 1999; Lutz *et al*., 2001). In this experiment, recombination was restricted to occur between amino acid positions 50 and 150. The schema profile was calculated from the structure of PurN (Almassy *et al*., 1992).

2.6. Hybrid Gene Construction

The oligonucleotide fragments corresponding to the peptide schemas were made via PCR amplification, where the primers at either end contain a short piece of DNA that overlaps with preceding gene fragment (Horton, 1995). This overlap ensures that the

fragments will re-anneal to produce a full-length gene. The promoter for PSE-4 in the PMON vector (Sanschagrin *et al*., 2000) is used for the 'A' fragments and the promoter for TEM-1 in the PSTBlue-1 vector (Novagen) is used for the 'B' fragments. The PCR protocol is to initially heat at 95 ºC, then perform 25 cycles of heating at 94 ºC for 45 seconds, cooling at 52 ºC for 45 seconds, and extending at 72 ºC for 1 minute. The fragment is then gel purified and concentrated either through ethanol precipitation (for fragments less than 100 bp) or using a Zymoclean-5 gel extraction kit (for fragments > 100 bp). Once the oligonucleotide fragments are isolated, they are re-annealed to create a complete gene fragment through a second PCR amplification step. The forward and reverse primers have the sequences for the restriction sites of EcoRI and HindIII, respectively, so that the complete genes can be inserted into the PMON vector that has been modified to contain these restriction sites. The times and temperatures are identical to the previous amplification round. A pre-PCR step can be used to improve the purity of the amplified genes. This PCR protocol is 25 iterations of 95 ºC for 30 seconds, 5 ºC for 30 seconds, and 72 ºC for 2 minutes. A final extension of 10 minutes at 72 ºC is done after the cycles are complete. The fragments are purified using the Zymoclean-5 gel extraction kit. Finally, the fragments are ligated into the PMON vector, which has kanamycin resistance. The vectors containing the hybrid genes are transformed into XL1-BLUE super competent ($>10^9$) cells and grown on plates that contain 10 µg/ml kanamycin. Colonies are isolated and the vector is extracted and sequenced. Some of the recombinant genes contained point mutations after the construction process (approximately 0.06% nucleotide changes per gene). The PSE-4 gene and the PMON vector were provided by Roger C. Levesque (Université Laval, Québec, Canada).

2.7. MIC Screening

Each hybrid β-lactamase is tested for its activity towards the degradation of the antibiotic ampicillin. To rapidly screen for this property, agar plates are made with following exponentially increasing concentrations of ampicillin: 10, 20, 40, 80, 160, 320, 640, and 1280 μg/ml. Aliquots of transformed cells are spread on the plates and allowed to grow at 37 ºC for 24 hours. More active hybrids will grow on plates with greater concentrations of ampicillin. The activity is measured as the minimum inhibitory concentration (MIC), in other words, the lowest concentration of ampicillin that kills the cells. The XL1-BLUE cells naturally have a MIC of 10, so β-lactamase activity cannot be measured below this point. The wild-type TEM-1 and PSE-4 enzymes have MICs > 2560 μg/ml.

3. Results and Discussion

3.1. Correlation with *In Vitro* Recombination Experiments

The SCHEMA calculation was tested against five experiments in which the genetic information from several parents was recombined to create random libraries of hybrid proteins. In each experiment, a subset of the crossovers survives the screen or selection by retaining (or improving) function. In Figure 5-3, we compare the locations of the functional crossovers with the calculated schema profiles for functional hybrids of cephalosporinases (Crameri *et al*., 1998; Sanschagrin, *et al*., 2000), subtilisins (Ness *et al*., 1999), cytochrome P450s (Brock and Waterman, 2000), and glycinamide-ribonucleotide transformylases (Ostermeier *et al*., 1999; Lutz *et al*., 2001). Nearly all of the observed crossovers appear in regions corresponding to minima in the schema

profiles. The recombination techniques used in these experiments vary significantly, demonstrating the robustness of the predictions.

We find that the window size that best predicts the locations of crossovers in selected libraries is fourteen, which results in domain sizes of approximately twenty to thirty residues. Typically, three types of schema are observed: (i) bundles of alpha-helixes, (ii) an alpha-helix combined with beta-strands, and (iii) beta-strands connected by a hairpin turn. While the algorithm finds these schemas relatively often, there are numerous interesting exceptions. For example, crossovers are frequently predicted to occur in the center of alpha-helixes. In addition, there are schema that are composed of complicated topologies with little discernable secondary structure.

The regions where crossovers are predicted to be deleterious are also noteworthy. For example, crossovers in loops can be highly disruptive if they divide interacting units of secondary structure. A common motif that demonstrates this effect is a single α-helix that is connected by a loop to a β-strand. A single crossover in the loop will disrupt interactions between these secondary structural elements. By the same reasoning, recombining isolated units of secondary structure can be disruptive.

3.2 Single-crossover Recombination Experiments

Several experimental techniques have been proposed that can recombine two parents to create a library where each hybrid is restricted to having a single crossover. This strategy has been applied to recombine P450s (Sieber *et al*., 2001) and glycinamide-ribonucleotide transformylases (Ostermeier *et al*., 1999; Lutz *et al*., 2001). The transformylase experiment resulted in many hybrids that have crossovers in the center of

the sequence, but the P450 experiment resulted in crossovers restricted to the N- and C-termini. The disruption of all possible single crossovers can be easily calculated and plotted (Figure 5-4). All of the single crossovers that led to functional hybrids in the tranformylase experiment are found to occur at points that minimize the disruption. This profile is unusual because the minima predicted by the single-crossover and schema disruption profiles are nearly identical (Figure 5-3D). Usually, single crossovers are highly disruptive in the middle of the primary amino acid sequence. Due to this effect, it is likely that most functional crossovers discovered by these techniques will occur near the termini of the sequence.

3.3 Designing β-lactamase Hybrids

Although there is good agreement between the schema profile and the positions of crossovers found during *in vitro* recombination experiments, this agreement does not tell us the degree to which the total amount of schema disruption can be tolerated in a given hybrid. To test this aspect, we recombined two β-lactamases (TEM-1 and PSE-4) that share only 40% amino acid sequence identity, but have highly similar structures (Jelsch *et al*., 1993; Sanschagrin *et al*., 2000; Lim *et al*., 2001). The calculated schema profile of β-lactamase (Figure 5-5) was used to divide the structure into schemas (Figure 5-6) and then the degree to which the schemas are interacting was calculated (Figure 5-7). We then designed hybrids such that they have increasing disruption (Figure 5-8), but there is no correlation with the size of the recombined fragment or with the number of effective mutations corresponding with the recombination event (Table 5-1). This series of hybrid β-lactamases was then experimentally constructed by piecing together DNA fragments of

TEM-1 and PSE-4 by PCR (Horton, 1995; Sanschagrin et al., 2000) (see Section 2.6). In addition, we constructed the sequence mirrors of several hybrids. For example, for a two-crossover hybrid (three fragments), we constructed the hybrid in which the first fragment is from PSE-4 (labeled 'A') as well as that having the first fragment from TEM-1 (labeled 'B').

We tested each hybrid protein for activity by measuring the minimum concentration of ampicillin required to inhibit cell growth (MIC). Wild-type TEM-1 and PSE-4 are highly active towards ampicillin (MIC > 2560 μg/ml) and have similar activities towards various β-lactam substrates (Sanschagrin *et al*., 2000). The MIC value is a complex combination of various parameters, including expression, stability and activity (Palzkill and Botstein, 1992; Huang *et al*., 1996). Here, we are using the observation of resistance merely as a measure that the hybrid β-lactamase is folded and functional and not to precisely rank the individual activities of hybrid enzymes. Measuring the MIC of each hybrid, we found a sharp transition in disruption, beyond which hybrids are non-functional (Figure 5-9). This transition does not correlate with the number of mutations that effectively occur when the hybrid is constructed (Table 5-1). The transition divides the graph into two regions: tolerated (non-disruptive) and "dead" (highly disruptive). The region just before the transition may be the optimal level of disruption to target in creating libraries of hybrids. In this way, diversity is maximized while the fraction of the library that is non-functional or unfolded is minimized.

The eight hybrids that show activity (1A to 5B) have interesting characteristics. Many have at least one crossover at a buried position. Additionally, a crossover occurs in the middle of a helix for two hybrids (2A, 2B) and at the end of a β-strand in hybrid 1A.

Finally, four of the hybrids (2A, 2B, 3A, 3B) have crossovers near the active site. Notably, several hybrids that were determined to be non-functional (7A, 8A) have crossovers that occur in a loop on the surface with only a few residues being recombined at the termini. Crossovers in loops are often considered to be non-disruptive, yet our algorithm correctly identified them to be strongly disruptive in this context. Finally, we constructed two hybrids (4A and 7A) that only differ only by twelve residues near the N-terminus. Hybrid 4A was found to be functional, whereas hybrid 7A was found to be non-functional. This distinction would be hard to predict based solely on visualizing the differences mapped onto the three-dimensional structure (Figure 5-7).

3.3 Characteristics of the Schema Profile

3.3.1. Parameter Sensitivity

The form of the schema profile is robust with regard to the model parameters. In the SCHEMA algorithm, the only model parameters are the size of the window w and the distance used to determine if residues are interacting d_c. For all of the data sets presented in this chapter, we use the parameters w = 14 residues and d_c = 4.5 Å. These parameters can be varied without losing the general topology of the schema profile. In Figure 5-10, the schema profile of the β-lactamase TEM-1/PSE-4 system is calculated using various values of w (6 to 18) and d_c (3.5 to 5.5). Within these parameter ranges, the topology of the schema profile remains remarkably robust.

3.3.2. The Probability Matrix

The parental sequences are required to calculate the schema disruption. The probabilities used in Equation (5-1) and described in Section 2.3 are calculated based on an alignment of these sequences. These probabilities are required to obtain the proper trend in the schema disruption profile as the sequence identity of the parents goes to 100%. Even if two residues are interacting in the structure, if their amino acid identities are the same in all of the parents, then it is impossible for recombination to cause a disruption (Figure 5-2). Therefore, as the identity of the sequences increases, the number of crossovers that are consistent with maintaining structural integrity will also increase. As some point, the schema profile will no longer resemble the underlying structural motifs, but rather will reflect the amino acid differences between the parental sequences. To demonstrate the effect, the schema disruption profiles with and without the probability matrix are shown in Figure 5-11. For these examples, there is sufficiently little sequence identity that the minima of the profiles remain similar.

3.3.3. Correlation with other Domain Algorithms

Many algorithms have been proposed to divide protein structures into domains (Rossman and Liljas, 1974; Crippen, 1978; Rose, 1979; Go, 1981; Zehfus and Rose, 1986; Holm and Sander, 1994; Panchenko *et al*., 1996; Tsai *et al*., 2000). Algorithms have been developed to identify folding units, intron locations, and evolutionary motifs. In general, there are several difficulties in using these algorithms to identify schema. First, they cannot be used to quantitatively assess the ability for a given fragment to be recombined, as is done in Figures 5-8 and 5-9. In addition, they are unable to scale the

suitability of a fragment for recombination by the sequence identity shared between the parents (Sections 2.3 and 3.3.2). In this section, we compare results obtained from two two widely used domain-finding algorithms with the schema disruption profile.

Gō proposed a domain-finding algorithm based on the contact map for a protein structure (Gō, 1981; Gō, 1983). Two residues are considered to be in contact if their C_α atoms are within a cutoff distance. Gō found that plotting the contact map for a protein structure and visually subdividing the map into regions that preserve the maximum number of residue contacts could identify domains. This process can be automated by counting the number of residues that are outside of the cutoff distance for residue *i*

$$C_i = \sum_{j=1,\, j\neq i}^{N} c_{ij} \ , \tag{5-3}$$

where N is the number of residues and c_{ij} is equal to 1 if residues i and j are within a cutoff distance and is equal to 0 if they are outside this distance. The cutoff distances are typically large (15 to 30 Å). Residues for which C_i is at a minimum represent domain boundaries. The profile of C for cephalosporinase is shown in Figure 5-12A and compared with the schema profile. While the two domain definitions are consistent, the schema profile fits the recombination data sets more accurately. In addition, Equation (5-3) cannot be scaled by sequence identity nor be used to determine the disruption caused by a specific fragment.

Other domain algorithms have been developed that split the protein structure into small building blocks which are subsequently pieced back together to form domains. The current state-of-the-art version of these methods was developed by Nussinov and co-workers (Tsai *et al*., 2000). This algorithm looks for compact units of protein structure based on a scoring function that includes terms that maximize the amount of buried

surface area and the degree of isolation for a domain. An advantage of the Nussinov algorithm is that it can predict non-contiguous domains, created by multiple breaks in the primary sequence. This algorithm is available for use on the Internet at http://protein3d.ncicrf.gov/tsai/anatomy.html. The output for the cephalosporinase structure is shown in Figure 5-12B. While the output is generally consistent with the schema profile, there are several important differences. First, the Nussinov algorithm requires that the domain be larger than a minimum size and that it be isolated from the remainder of the structure. In contrast, we observe that as the fragment size decreases, it is more likely to be accepted during shuffling. Second, the Nussinov algorithm's scoring system is inappropriate for determining regions of acceptable crossovers. These are typically the areas between compact domains throughout which crossovers can occur. Finally, the Nussinov algorithm cannot be scaled by the degree of sequence identity that is shared between the parents.

3.4. The Natural Selection of Intron Locations

Gō discovered a correlation between the location of introns and isolated geometrical domains, a correlation that has held for a wide range of proteins (Gō, 1981; Gō, 1983). This correlation has been interpreted as evidence for the "introns-early" theory of evolution, which states that the first large proteins were constructed from smaller domains through recombination and gene duplication (De Souza *et al*., 1996; Gilbert *et al*., 1997). The merging of genes resulted in the separation of the coding DNA by regions of non-coding DNA (introns). Over evolutionary time, the introns disappeared where they were no longer necessary or were disadvantageous, for example, in the

restricted genome sizes of prokaryotes. Proponents of this theory have argued that if introns appeared late in evolution, their locations would appear random with respect to structural domains (De Souza *et al*., 1996; Gilbert *et al*., 1997). Our results indicate that the correlation between introns and domains could occur as a result of natural selection, even if the introns appeared late.

Of the many proposed functions of introns, one is that they facilitate the swapping of exons (De Souza *et al*., 1996; Gilbert *et al*., 1997). If the probability of a crossover is equal across the gene, then a long region of non-coding DNA will bias the crossovers towards a specific region of the fully spliced gene. Cycles of recombination and selection can bias the location of introns if the ability of an intron to promote shuffling contributes to an organism's fitness. If, in a population of these organisms, introns were randomly distributed throughout the gene, then there would be a selective advantage to those individuals whose introns appeared in regions that are the most likely to result in successful shuffling events. We have observed this directly in *in vitro* recombination experiments. When crossovers are randomly distributed throughout the gene, the subset that preserve the schema is also the most likely to result in folded, functional hybrids. Therefore, if introns promote recombination, they will most likely reside in low-disruption regions after selection.

4. Conclusions

We have demonstrated that crossovers that lead to folded and functional hybrid proteins occur at positions that minimize the number of disrupted interactions. It is noteworthy that our very simple model of interacting residues can capture this effect. An

important application of the results presented here will be to accelerate molecular optimization by laboratory evolution methods through the use of computational tools (Voigt *et al*., 2000; Voigt *et al*., 2001a; Voigt *et al*., 2001b; Bolon *et al*., 2002). Combinatorial libraries with targeted crossovers can dramatically improve an evolutionary search by significantly reducing the number of mutants that must be screened to obtain specific functional changes. The elucidation and experimental verification of evolutionary dynamics will allow the design of a new generation of evolutionary methods that maximize our ability to discover novel biological molecules.

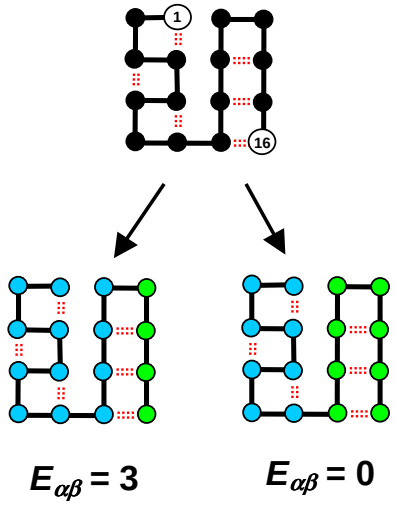
Table 5-1. Designed TEM-1/PSE-4 Hybrid β-lactamases

	Crossover 1		Crossover 2[a]				
Hybrid[b]	#	Context[c]	#	Context[c]	*m*[d]	E_{ab}	MIC
1A[e]	163	loop, surface	179	strand, core	7	6	2560[f]
2A	189	helix, core	216	loop, surface, as	18	15	1280
2B	189	helix, core	216	loop, surface, as	18	15	40
3A	130	loop, core, as	163	loop, surface	13	21	20
3B	130	loop, core, as	163	loop, surface	13	21	320
4A	65	loop, surface			42	25	320
5A	70	loop, core, as	216	loop, surface, as	83	26	320
5B	70	loop, core, as	216	loop, surface, as	83	26	20
6A	70	loop, core, as	130	loop, core, as	41	27	10[g]
6B	70	loop, core, as	130	loop, core, as	41	27	10[g]
7A	53	loop, surface			42	33	10[g]
8A	254	loop, surface			23	37	10[g]

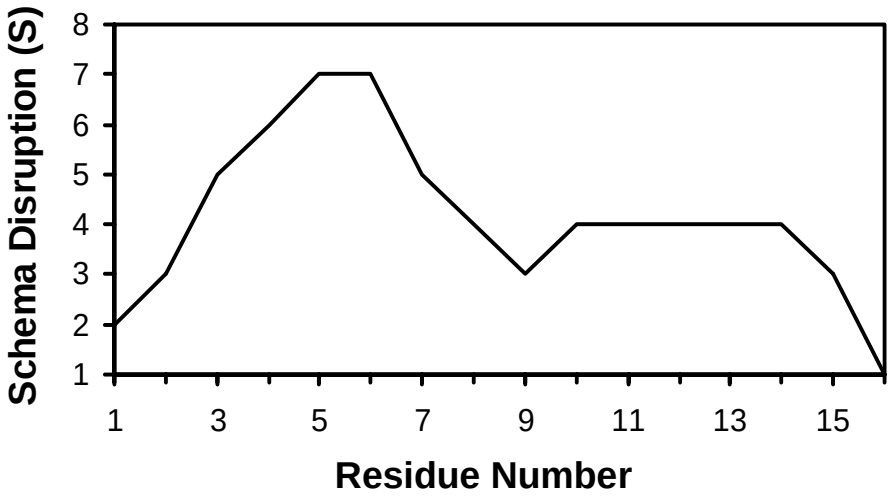
a. This portion is left blank if the hybrid protein only has a single crossover.
b. The letter in the name indicates the parent that composes the first portion of the gene, where 'A' is PSE-4 and 'B' is TEM-1. For the double crossover mutants, an 'A' indicates a gene structure of A-B-A and 'B' indicates B-A-B.
c. The context of the side chain of the residue where the cut occurs. The notation "as" indicates that the crossover occurs near the active site.
d. The number of mutations that occur when the smaller fragment of one parent is inserted into the larger context of the remaining parent.
e. This hybrid has been previously constructed by Levesque and co-workers (Sanshagrin *et al*., 2000)
f. Wild-type activity of both PSE-4 and TEM-1
g. The MIC of XL1-BLUE cells. No β-lactamase activity is observed.

Figure 5-1:

(A) An illustration of schema disruption. Black lines in the structure represent peptide bonds and the red dots are interactions between amino acid side chains. Two hybrid proteins are shown. When the last four residues come from one parent and the remaining residues come from the other parent, three interactions are disrupted. When the last eight residues come from the same parent, then there is no disruption. According to our schema theory, achieving folded hybrid proteins is more likely when the fewest interactions are disrupted. **(B)** The schema profile of the structure in (A) calculated with a window size $w = 6$.


1
16
$E_{\alpha\beta} = 3$
$E_{\alpha\beta} = 0$

(A)

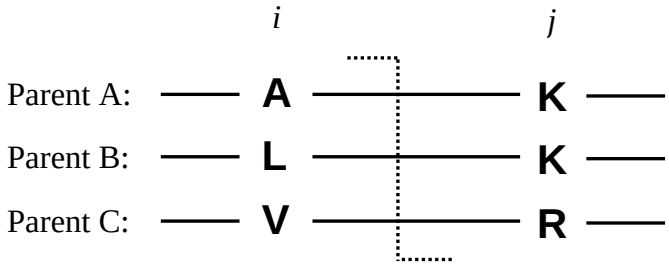

Schema Disruption (S)
8
7
6
5
4
3
2
1
1
3
5
7
9
11
13
15
Residue Number

(B)

Figure 5-2:

An example of the calculation of the probabilitities P_{ij} required by Equations (5-1) and (5-2). In this example, there are three parents with different amino acid combinations at residues *i* and *j*. P_{ij} is the probability that the hybrid proteins will have a combination of amino acids that is not present in any of the parents. Considering a crossover that divides these residues (dashed line), there are six possible hybrid proteins. Of this set, two hybrids have the same pair of amino acids present in the parental set (boxed). The probability that this crossover will result in a disrupted interaction is then $P_{ij} = 4/6$.

Three Parents:



Possible Hybrids after Recombination:

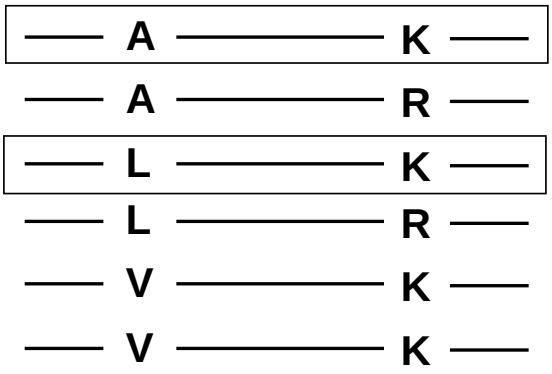
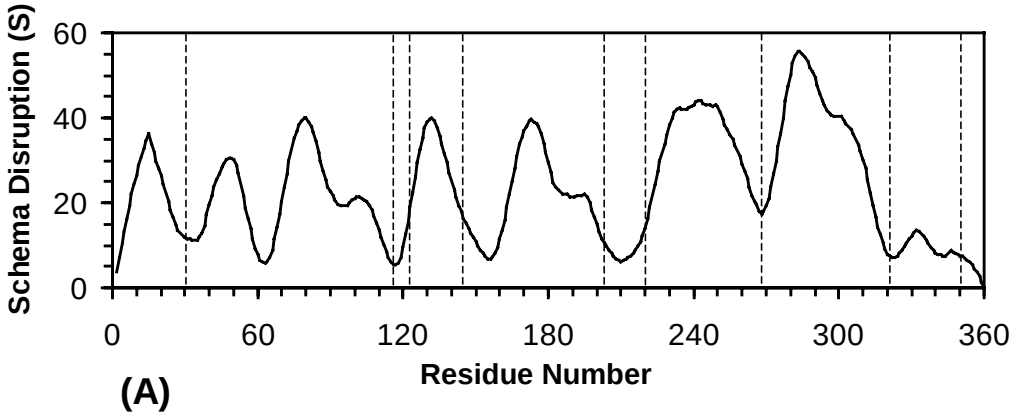
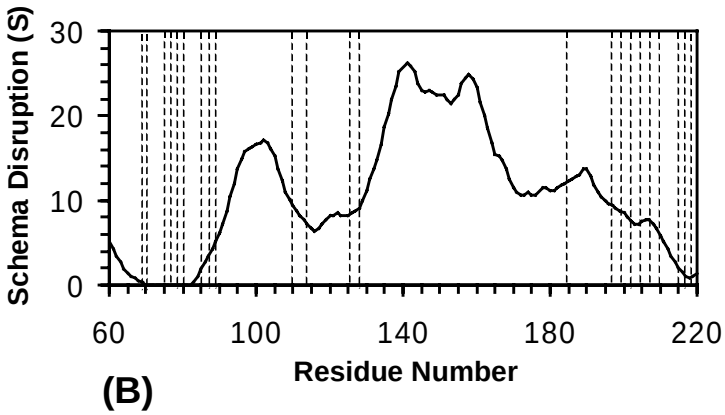


Figure 5-3:

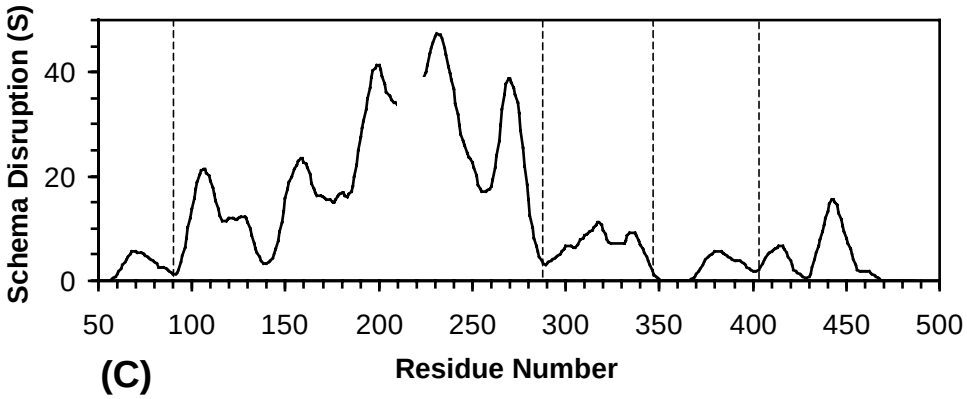
The schema disruption profile is compared with various *in vitro* recombination data sets. Each hatched line indicates where a recombination event resulted in a functional hybrid protein. All of the calculations were run using Equation (5-2) with a window size of 14 residues and d_c = 4.5 angstroms. **(A)** The schema profile as determined from the cephalosporinase structure compared with the experimentally observed crossover points by DNA shuffling (Crameri *et al*., 1998) and rational design (Sanschagrin *et al*., 2000). **(B)** A comparison of the schema profile of Savinase with the crossovers that led to the improvement of the properties of subtilisin (Ness *et al*., 1999). The crossovers between subtilisin families that led to improvements in thermostability, activity at high or low pH, or stability in organic solvent, are indicated. **(C)** A schema disruption calculation of the P450 2C5 structure, based on the sequences of rat and bacterial c17 (Brock and Waterman, 2000). The dashed lines indicate where single crossover recombination events led to folded hybrids. Note that residues 212-222 are missing from the structure, represented by a break in the schema profile. **(D)** The schema profile for the sequence independent recombination of PurN and GART glycinamide ribonucleotide transformylase (Ostermeier *et al*., 1999; Lutz *et al*., 2000). Recombination was only allowed to occur between residues 50 and 150. The single crossovers that led to functional hybrid proteins are indicated.


Schema Disruption (S)
60
40
20
0
0
60
120
180
240
300
360
Residue Number

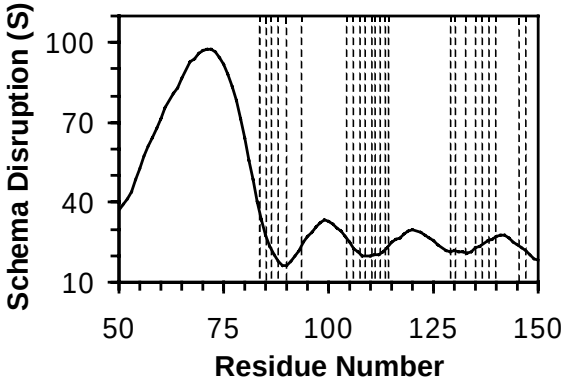
(A)


Schema Disruption (S)
30
20
10
0
60
100
140
180
220
Residue Number

(B)


Schema Disruption (S)
40
20
0
50
100
150
200
250
300
350
400
450
500
Residue Number

(C)


Schema Disruption (S)
100
70
40
10
50
75
100
125
150
Residue Number

(D)

Figure 5-4:

The single-crossover schema disruption profile is shown for glycinamide ribonucleotide transformylase. The stars indicate where single crossovers led to functional hybrid proteins (Ostermeier *et al*., 1999; Lutz *et al*., 2000). Note that the minima correlate with the schema calculation for this example (Figure 5-3D).

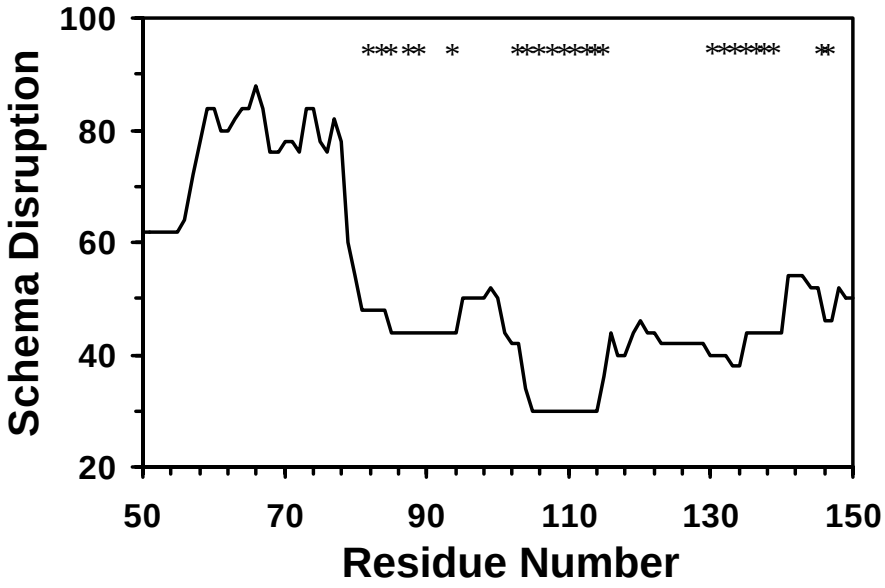

Schema Disruption
100
80
60
40
20
50
70
90
110
130
150
Residue Number

Figure 5-5:

Schema disruption profile for recombination of β-lactamase TEM-1 and PSE-4. Nearly identical results are obtained when the calculation is run on the TEM-1 (gray line) and PSE-4 (black line) structures. The orange and purple regions mark the basins of large minima. Crossovers are predicted to acceptable throughout these basins.

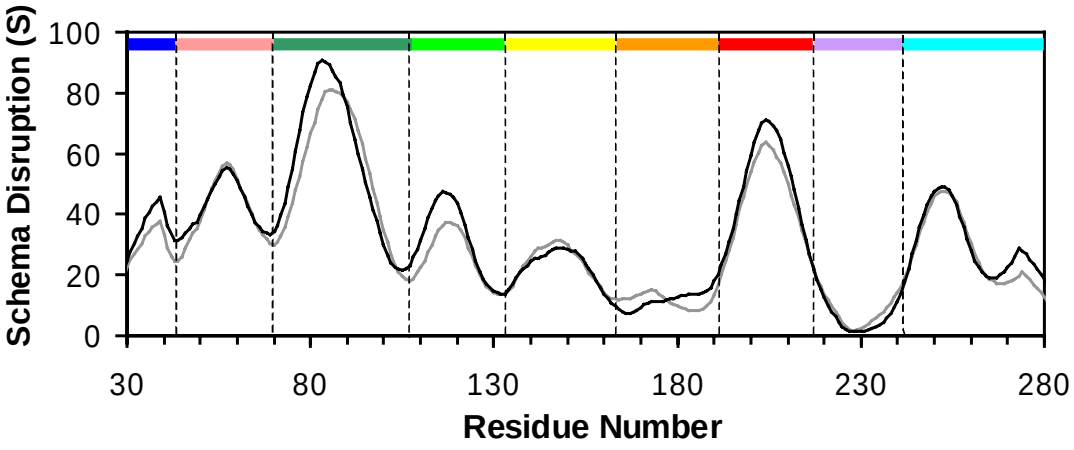

Schema Disruption (S)
100
80
60
40
20
0
30
80
130
180
230
280
Residue Number

Figure 5-6:

The predicted schema mapped onto the three dimensional structure of TEM-1 β-lactamase. This figure was generated using MolMol (Koradi *et al*., 1996).

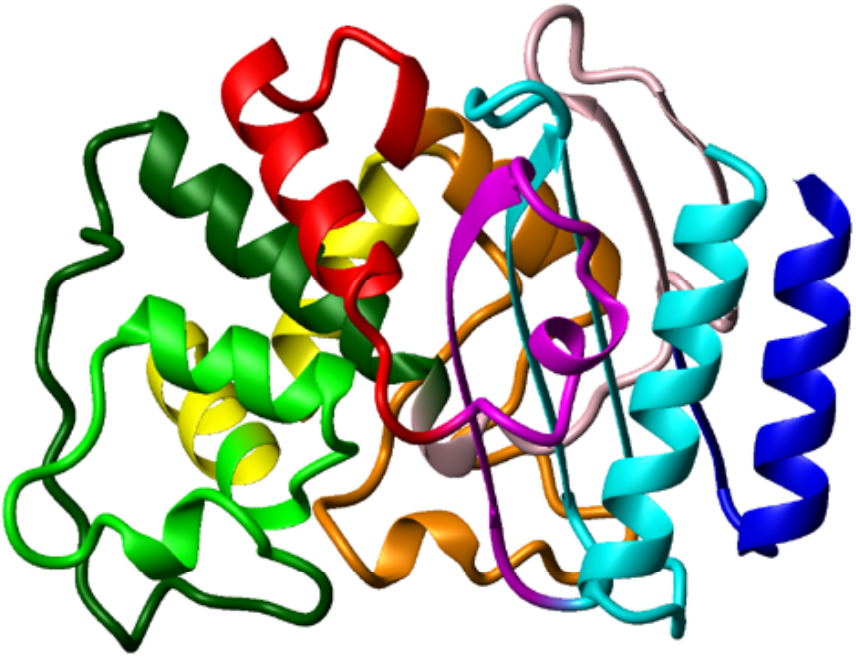


Figure 5-7:

The number of interactions between schemas averaged between the PSE-4 and TEM-1 structures. The thickness of each line is proportional to the number of interactions between two schemas, as calculated using Equation 5-1. The thickest lines represent highly interacting schemas for which there are greater than eight interactions, the medium line for between five and eight interactions, and the thin line for between two and four interactions. Note that the magenta and orange fragments are not true schemas; rather, they represent extended minima in the schema profile (Figure 5-5). This figure was generated using MolMol (Koradi *et al*., 1996).

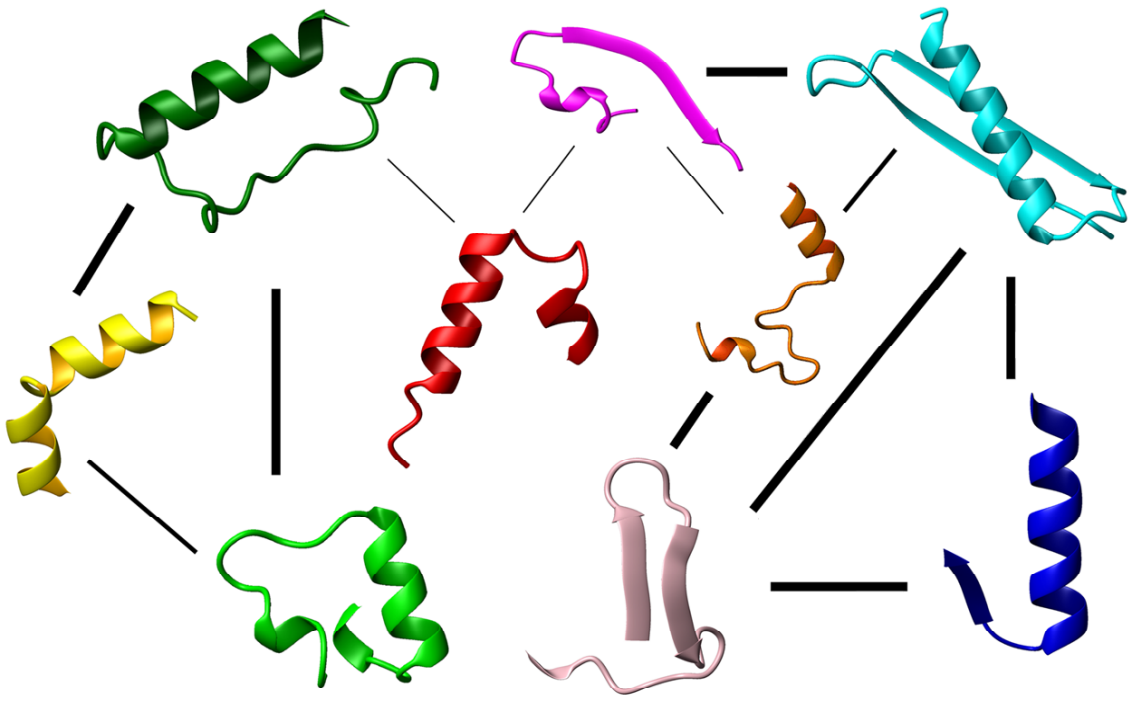


Figure 5-8:

The designed hybrids of β-lactamase TEM-1 (red) and PSE-4 (blue) mapped onto the TEM-1 structure, shown in order of increasing disruption (see Table 5-1).

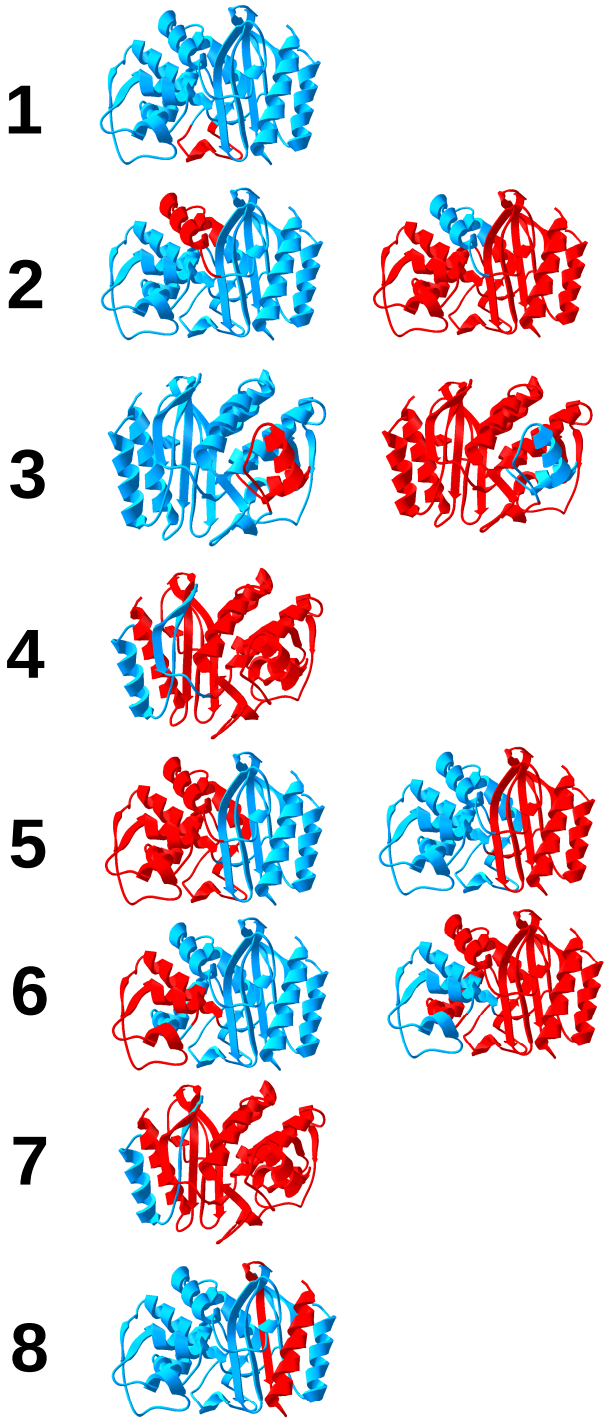

1
2
3
4
5
6
7
8

Figure 5-9:

The activities of hybrid β-lactamases are shown as a function of their disruption (Equation 5-1). The lower line marks the point at which the MIC represents the background antibiotic resistance of the cells. Activity is lost at $E_{\alpha\beta} \approx 27$. Below the transition, the recombination events are non-disruptive. Above the transition, the hybrids are non-functional. The region just before the transition may be the optimal target for library creation, where diversity is maximized without disturbing the stability or enzymatic activity. The color of the points indicates the parent of the first fragment: red is PSE-4 (‘A’), blue is TEM-1 (‘B’), and purple indicates that the points overlap.

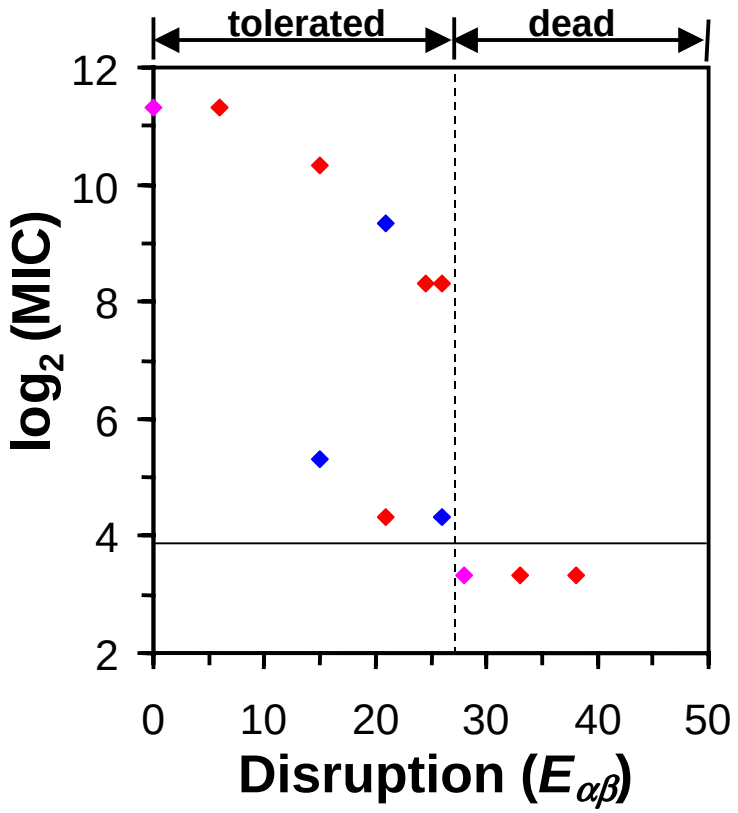
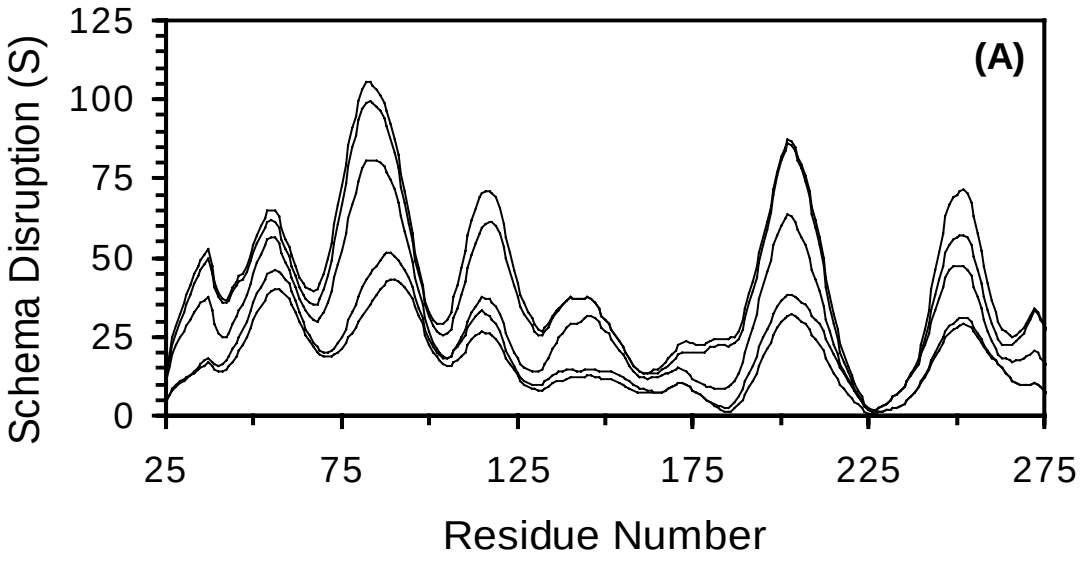

tolerated
dead
12
10
8
6
4
2
log2 (MIC)
0
10
20
30
40
50
Disruption (Eαβ)

Figure 5-10:

(A) The sensitivity of the schema profile to variation in d_c. The profile for the β-lactamase TEM-1/PSE-4 system was generated for different values of these parameters. The curves (from bottom to top) for w = 14 and d_c = 3.5, 4.0, 4.5, 5.0, 5.5 angstroms. The sensitivity of the schema profile to variation in w. **(B)** The profile for the β-lactamase TEM-1/PSE-4 system was generated for different values of these parameters. The curves (from bottom to top) for d_c = 4.5 and w = 6, 10, 14, 18 residues.


(A)
Schema Disruption (S)
125
100
75
50
25
0
25
75
125
175
225
275
Residue Number

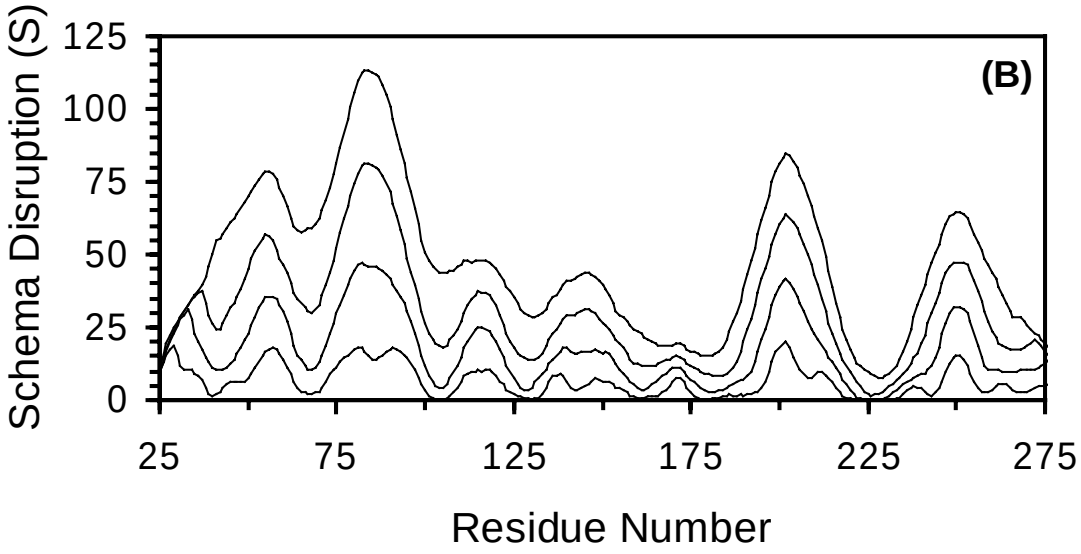
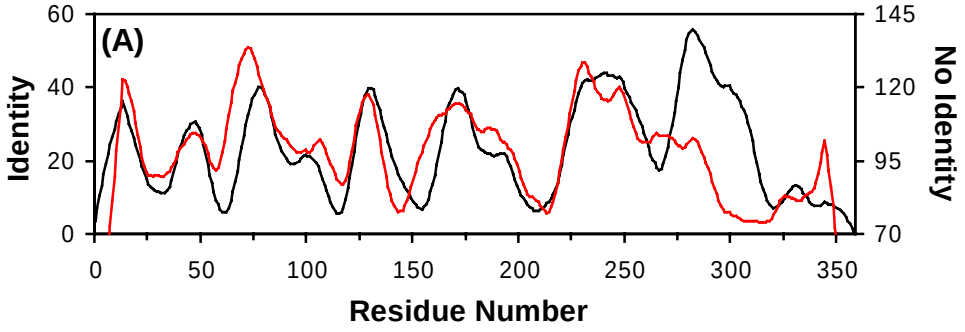
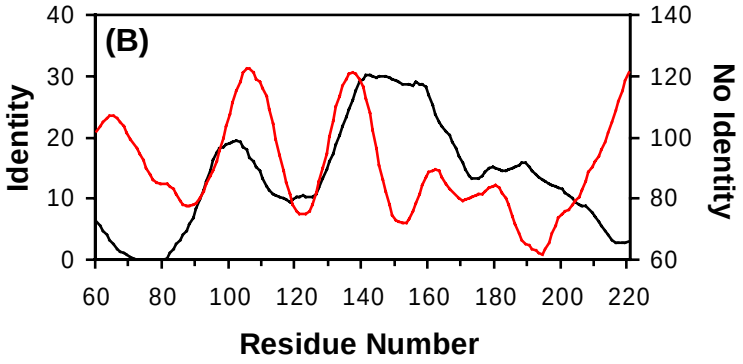
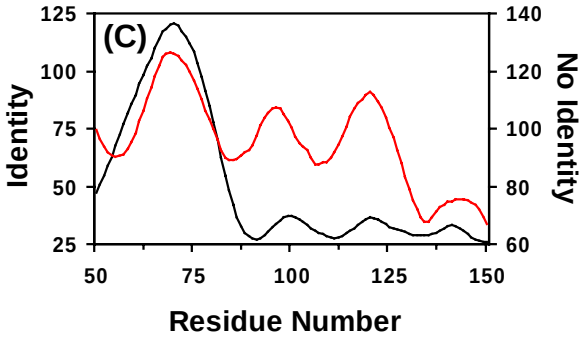

(B)
Schema Disruption (S)
125
100
75
50
25
0
25
75
125
175
225
275
Residue Number

Figure 5-11:

The effect of the probability matrix on the schema disruption profile is demonstrated for **(A)** cephalosporinase, **(B)** subtilisin, **(C)** transformylase, and **(D)** β-lactamase. Each profile is shown both with (black line) and without (red line) the identity matrix. Notably, the minima are similar in the two profiles. As the sequence identity shared between the parents increases, the similarity between the two curves will diminish.


(A)
Identity
No Identity
Residue Number
0
20
40
60
70
95
120
145
0
50
100
150
200
250
300
350


(B)
Identity
No Identity
Residue Number
0
10
20
30
40
60
80
100
120
140
60
80
100
120
140
160
180
200
220


(C)
Identity
No Identity
Residue Number
25
50
75
100
125
60
80
100
120
140
50
75
100
125
150

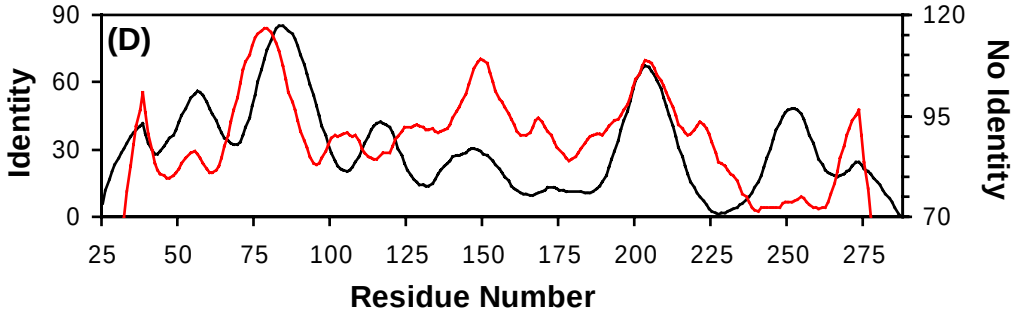
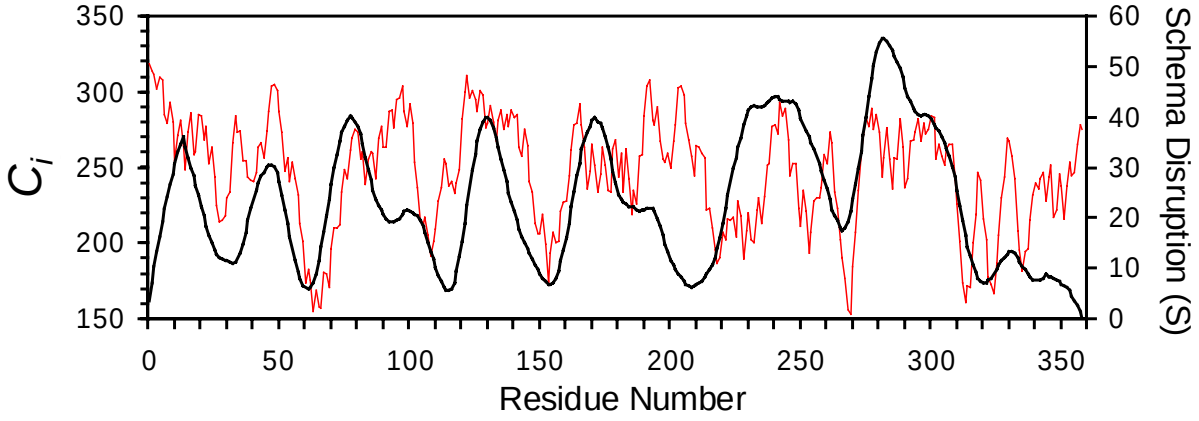
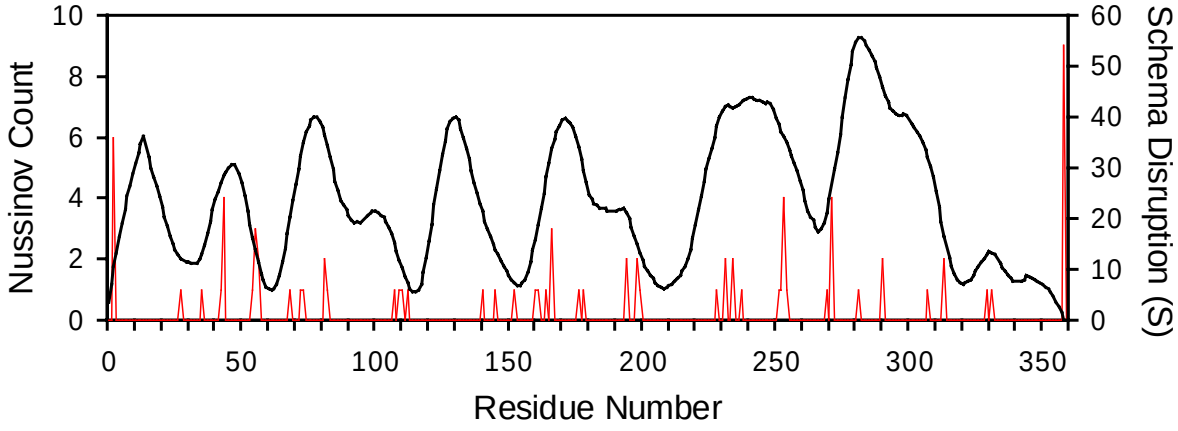

(D)
Identity
No Identity
Residue Number
0
30
60
90
70
95
120
25
50
75
100
125
150
175
200
225
250
275

Figure 5-12:

Two domain-finding algorithms are compared with the schema disruption profile for the cephalosporinase structure. **(A)** The schema profile (black) is compared to a Gō-like algorithm (red) for discovering domain boundaries. The Gō algorithm tracks the number of C_α's that are within a cutoff radius of each residue. Here, the radius is set to be 20 Å. **(B)** The schema profile (black) is compared to the Nussinov algorithm (Tsai *et al*., 2000). The output of the Nussinov algorithm is a list of optimal protein subunits. To visualize this, we have incremented the end points of each subunit by one (red spikes). Larger spikes indicate that that residue is the end point for more than one subunit.


C_i
Schema Disruption (S)
Residue Number
350
300
250
200
150
60
50
40
30
20
10
0
0
50
100
150
200
250
300
350

(A)


Nussinov Count
Schema Disruption (S)
Residue Number
10
8
6
4
2
0
60
50
40
30
20
10
0
0
50
100
150
200
250
300
350

(B)

Chapter 6

Comparing Search Algorithms in Protein Sequence Design

Portions of this chapter are reproduced from:
Voigt, C. A., Gordon, D. B., and Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. J. Mol. Biol. 299, 789-803.

Abstract

Finding the minimum energy amino acid side chain conformation is a fundamental problem in both homology modeling and protein design. To address this issue, numerous computational algorithms have been proposed. However, there have been few quantitative comparisons between methods and there is very little general understanding of the types of problems that are appropriate for each algorithm. Here, we study four common search techniques - Monte Carlo (MC) and Monte Carlo plus quench (MCQ), genetic algorithms (GA), self-consistent mean field (SCMF), and dead-end elimination (DEE). Both SCMF and DEE are deterministic, and if DEE converges, it is guaranteed that its solution is the global minimum energy conformation (GMEC). This provides a means to compare the accuracy of SCMF and the stochastic methods. For the side chain placement calculations, we find that DEE rapidly converges to the GMEC in all the test cases. The other algorithms converge on significantly incorrect solutions; the average fraction of incorrect rotamers for SCMF is 0.12, GA 0.09, and MCQ 0.05. For the protein design calculations, design positions are progressively added to the side chain placement calculation until the time required for DEE diverges sharply. As the complexity of the problem increases, the accuracy of each method is determined so that the results can be extrapolated into the region where DEE is no longer tractable. We find that both SCMF and MCQ perform reasonably well on core calculations (fraction amino acids incorrect:

SCMF 0.07, MCQ 0.04), but fail considerably on the boundary (SCMF 0.28, MCQ 0.32) and surface calculations (SCMF 0.37, MCQ 0.44).

1. Introduction

A goal of computational protein design is to compute an amino acid sequence *de novo* that will fold into a defined backbone structure (Street and Mayo, 1999). This is a difficult task as protein stability results from the sum of many subtle and highly coupled interactions. By applying a quantitative, generalized approach, computational protein design has proven successful for cro protein (Desjarlais and Handel, 1995), gcn4 (Dahiyat and Mayo, 1996; Dahiyat *et al*., 1997a), protein G (Dahiyat and Mayo, 1997b; Su and Mayo, 1997; Malakaukas and Mayo, 1998), ubiquitin (Lazar *et al*., 1997), zinc finger domain (Dahiyat and Mayo, 1997a), and engrailed homeodomain (Marshall *et al*., 1999). Protein design has also been successful in designing alpha-helical peptides that form right-handed supercoils (Harbury *et al*., 1998). The trend towards designing sequences for larger and flexible backbones has been facilitated by a greater understanding of the forces responsible for protein stability as well as improvements in methods to search for the minimum energy conformation.

A combination of important techniques constitutes the protein design algorithm. First, the flexibility of each amino acid is course-grained into a discrete set of statistically significant conformations called rotamers (Ponder and Richards, 1987; Dunbrack and Karplus, 1993; Dunbrack and Karplus, 1994). While this drastically reduces the search space and makes the problem computationally tractable, the combinatorial complexity remains enormous. As an illustration, the number of side chain conformations for a

protein of n residues, a = 20 amino acids, and r rotamers per amino acid, is $(r \times a)^n$. For even a moderately sized enzyme, n is approximately 200–400, creating an immense set of possible solutions.

Second, the interactions between amino acids pertinent to stability have to be identified and their essence captured by a series of equations that together constitute the force field. This description represents the balance of forces responsible for protein stability including van der Waals interactions, hydrogen bonds, salt bridges, and hydrophobic-polar interactions (Gordon *et al*., 1999). The energy term consists of three contributions: backbone/backbone, rotamer/backbone, and rotamer/rotamer. Because the backbone remains fixed in most protein design algorithms, it is not relevant to the optimization procedure. Therefore, the energy of a sequence folded into a defined structure can be expressed as

$$E = \sum_{i=1}^{N} E(i_r) + \sum_{i=1}^{N-1} \sum_{j>i}^{N} E(i_r, j_s), \tag{6-1}$$

where $E(i_r)$ is the rotamer/backbone energy for rotamer i_r of residue i, $E(i_r, j_s)$ is the rotamer/rotamer energy of rotamers i_r and j_s of residues i and j, respectively, and N is the total number of residues. By assuming that the energy between rotamers is pairwise as in Equation (6-1), certain non-additive energy contributions cannot be treated exactly, such as a surface area based solvation potential (Street and Mayo, 1998).

When the rotamer description is combined with the force field, a discrete sequence-rotamer energy landscape is created in which each point represents a rotamer combination and an assigned energy. The final computational task in the protein design algorithm is to search this space for the global minimum energy conformation (GMEC) (Desjarlais and Clarke, 1998). Because the number of points in the landscape increases

exponentially with protein size, the time to find the minimum scales unfavorably. In addition, the landscape contains a large number of local minima created by the high degree of side chain coupling in the system. These effects compound, producing a hard search problem.

As a problem related to protein sequence design, homology modeling aims to align the sequence of an unknown structure with a sequence where the structure is known (Schiffer *et al*., 1990; Lee and Subbiah, 1991; Tuffery *et al*., 1991; Laughton, 1994; Lee, 1996; Sánchez and Šali, 1997). As the information in the sequence-structure database grows, it is increasingly observed that newly solved structures share structural motifs with proteins already in the database. Homology modeling consists of three central steps. First, a match is found between the sequence of the unknown structure and a sequence in the database of known structures. Then, the sequence is threaded onto the known backbone. Finally, the side chains are arranged onto the backbone based on an energy expression (Lee and Subbiah, 1991; Laughton, 1994). There are two issues in positioning the side chains correctly: the accuracy of the force field and rotamer descriptions, and the ability to find the minimum energy arrangement. Finding the GMEC of side-chain descriptions has led to the proposal of many search algorithms (Desjarlais and Clarke, 1998). Here, we are interested in comparing the different proposed techniques for energy minimization and are not concerned whether the GMEC of the energy landscape actually coincides with the proper (experimentally determined) side-chain conformation.

There are two general classes of search algorithms: stochastic and deterministic. Stochastic algorithms, such as Monte Carlo (MC) (Metropolis *et al*., 1953) and Genetic Algorithms (GA) (Holland, 1993), rely on probabilistic trajectories where the outcome is

determined by the initial conditions as well as the random number generator seed. Confirming that a solution is the GMEC is impossible, as there is always a degree of uncertainty. In contrast, deterministic methods will repeat the same solutions given the set of parameters used. Both dead-end elimination (DEE) (Desmet *et al.*, 1992) and self-consistent mean-field (SCMF) (Koehl and Delarue, 1994; Lee, 1994) are deterministic; however, they often do not converge to the same solution. If DEE converges, it is the GMEC, whereas this is not necessarily true for SCMF. The issues in comparing search algorithms include weighing the accuracy of the solution with the computational time required. Recently, the common algorithms used in protein sequence design have been reviewed (Desjarlais and Clarke, 1998). However, the tradeoffs of choosing one method over another are not well understood and there have been no comprehensive comparisons of methodologies. An understanding of the strengths and weaknesses of each algorithm is required so: (1) the algorithm best suited to the design problem can be utilized and (2) if an algorithm is chosen that does not give the GMEC, the expected accuracy of the solution can be estimated. In this paper, we compare four common approaches; MC, GA, SCMF, and DEE, for both side chain placement and protein design calculations.

2. Description of Search Algorithms

2.1. Monte Carlo

As one of the simplest stochastic search techniques, Monte Carlo (Metropolis *et al.*, 1953) often performs well on difficult energy landscapes. MC has been previously applied to problems relating to protein sequence design (Holm and Sander, 1992; Hellinga and Richards, 1994; Godzik, 1995; Sasai, 1995; Dahiyat and Mayo, 1996).

Initially, the rotamers for a sequence are picked at random. Then, a rotamer substitution is made at a randomly picked residue in the sequence. Rotamers of different amino acids are treated equally so a rotamer substitution can be either the same amino acid or a new one. A new energy E_{new} is calculated and if this energy is lower than the previous energy E_{old}, the move is accepted. If the energy is higher, the move is accepted with the Boltzman probability $p = \exp(-\beta(E_{new} - E_{old}))$, $\beta=1/kT$, where k is Boltzman's constant. The role of the temperature T is to overcome the multiple local minima in the energy landscape by allowing the trajectory to surmount energy barriers. To strengthen this effect, an initial temperature is selected and annealed. The temperature is then cyclically raised and lowered over the course of a single run between a designated high and low temperature (for the calculations performed here, the high and low temperatures were set to 4000 and 150 K, respectively). The optimization can be run for any number of cycles with each cycle containing a number of substitution attempts. Here, the optimization is run for 20 cycles of 10^6 substitution attempts for each test case. The number of cycles is arbitrarily set at 20 to generate computational times comparable to SCMF and DEE. MC can be run for longer periods to theoretically produce better solutions. In our hands, the number of cycles has been typically set at 1000.

At the end of the run, the energy of the stored solutions may be quenched. For each residue, in random order, all possible rotamers of the amino acid in the solution are attempted. If a new rotamer is lower in energy, it is kept; if not, it is rejected. The quench step produces a large improvement in the solution while adding trivially to the time of the run. This step assures that there are no single-rotamer changes that will improve the

energy. For the side chain placement calculations, results are presented for both the MC procedure alone as well as the MC with the quenching step (MCQ).

2.2. Genetic Algorithm

Genetic algorithms seek to optimize a population of solutions using biologically inspired operators (Holland, 1993). GA's have been applied to a wide range of problems including protein structure prediction (Pedersen and Moult, 1996) and design (Jones, 1994; Desjarlais and Handel, 1995; Lazar *et al.*, 1997). The advantages of a GA are that the population dynamics can overcome energy barriers by making moves in sequence space that are larger than the moves typically used in MC. In addition, beneficial mutations can be combined onto a single sequence, increasing the number of paths that circumvent local minima. As a disadvantage, GA's tend not to work well on highly coupled systems where crossover disruption is problematic as is expected for side chain systems. Further, residues that are close in sequence are not necessarily close structurally making it difficult for the algorithm to find clean crossover points.

While the specific implementation of GA's varies tremendously in literature, there are several common characteristics. In order to study the effectiveness of this approach on the protein design problem, we tried several different algorithms and chose a relatively universal description of a GA that produced the best results. The implementation of our algorithm is slightly different from that described by Desjarlais and Handel (Desjarlais and Handel, 1995). They include an inversion operator and utilize a different selection scheme. It is not expected that these differences would significantly change our results.

First, a population of $M = 50$ random sequences is initialized. Then, mutations are applied at rate $p_M = 0.016$, producing a Poisson distribution of mutations with an average of one per sequence. The new energies of the mutants are determined and ranked. The top C of these mutants are chosen for recombination where C represents the recombination rate. Here, the optimal value is found to be $C = 10$. For each pair, the strings are recombined by comparing each residue and if the rotamers differ among the parents, the offspring will inherit either parent's rotamer with equal probability. The new population is generated using the tournament selection technique where S sequences are picked randomly from the mutant library and their energies compared. The sequence with the lowest energy is allowed to continue to the next generation. The selection process is repeated M times to produce the pool of sequences that will continue to the next round of mutation, recombination, and selection. This algorithm is repeated so the average fitness of the population improves after each cycle until equilibrium is reached.

The selection strength, represented by the number of sequences S that undergo competition, is analogous to the temperature in the MC algorithm. By starting at low S and annealing to a high S, the population distribution in sequence space is first very broad and then narrows after each generation until the population consists of a single sequence. This "heating and cooling" process is repeated to improve the probability that the population will find lower minima. At the beginning of each cycle, S is initialized at 2 and is incrementally increased to 5. The full optimization procedure consists of 10 cycles of 10^4 mutation-recombination-selection steps. Due to its size, the number of cycles was increased to 15 for the 1arb test case. Similar to the Monte Carlo algorithms, the number

of cycles was arbitrarily set to produce competitive times against the deterministic algorithms.

2.3. Self-Consistent Mean Field

Unlike the MC or GA algorithms that focus on clever search methods to evade local minima, SCMF uses a mean-field description of the rotamer interactions to alter the energy landscape (Lee, 1994; Koehl and Delarue, 1994; Koehl and Delarue, 1995; Vásquez, 1995; Koehl and Delarue, 1996). SCMF is deterministic in that given a set of run parameters, the algorithm will always converge to the same solution. Unfortunately, there is no guarantee that the minimum of the mean-field landscape corresponds with the true GMEC. The advantage of SCMF is that the computational time scales linearly with the number of residues making it possible to obtain solutions for proteins currently unattainable by other methods (Koehl and Delarue, 1996).

As derrived by Koehl and Delarue (Koehl and Delarue, 1994), the mean-field energy for rotamer i_r at residue i is

$$E_{mf}(i_r) = E(i_r) + \sum_{j=1}^{N} \sum_{s=1}^{K_j} E(i_r, j_s) V(j_s) \,, \tag{6-2}$$

where K_j is the total number of rotamers at residue j. The weight of each rotamer $V(j_s)$ (the conformational probability vector) is normalized to unity. The first term in (6-2) is the contribution due to the interaction between the rotamer and the backbone and the second term describes all the inter-rotamer pairwise interactions weighted by the probability of that rotamer existing in the GMEC. The conformational probability vector can be independently calculated by Gibb's ensemble

$$V(j_s) = \frac{1}{q_j} \mathrm{e}^{-\beta E_{mf}(j_s)}, \tag{6-3}$$

where q_j is the partition function

$$q_j = \sum_{s=1}^{K_j} e^{-\beta E_{mf}(j_s)}. \tag{6-4}$$

The effect of this procedure is to smoothen the landscape and avoid the problem of multiple local minima making it relatively simple to locate the minimum of the mean-field energy landscape.

The mean-field energy is minimized using an annealing method as described by Lee (Lee, 1994). A high initial temperature (often > 20,000 degrees K) is chosen and the probability vector $V(j_s)$ is initialized at $1/K_j$ thereby assigning equal probability to each rotamer. The purpose of annealing the temperature is to assist the convergence, reducing the total run time. The solution found by SCMF is not dependent on the specific initial temperature used.

A pair-energy threshold is applied that implements a ceiling to which higher energies are set. The success of SCMF is highly dependent on the optimization of this parameter. The optimal threshold is determined individually for each side chain placement test case and is found to vary widely between 5 and 500 kcal/mol. Qualitatively, smaller backbones tended to correspond with a smaller threshold. The time required for SCMF to converge did not differ significantly with the threshold chosen. For the sequence design calculations, this parameter was set at 500 kcal/mol due to the increase in the problem difficulty.

After initialization, the mean-field potential E_{mf} (i_r) is calculated from Equation (6-2) for each residue and rotamer. The energies are converted into probabilities using

Gibb's equations. The algorithm iterates between Equations (6-2) and (6-3) until the energy converges and self-consistency is achieved (Koehl and Delarue, 1994). A convergence criterion of 0.0001 for $V(j_s)$ was used to define self-consistency. Convergence is significantly improved if the probability vector V is updated with a "memory" of the previous step as expressed by the following

$$V_{new}(j_s) = \lambda V_{new}(j_s) + (1-\lambda)V_{old}(j_s), \tag{6-5}$$

where the optimum step size was found to be $\lambda = 0.9$ (Koehl and Delarue, 1994). The temperature is then lowered in linear increments of 100 K and the routine repeated. When the final temperature is reached (100 K), the conformational vector represents the probability of finding each rotamer at a given residue position. The best solution is determined as the collection of rotamers that have the highest probability at each position.

2.4. Dead-End Elimination

As opposed to optimizing a single solution or set of solutions by iterative improvement as done by the MC procedure or GA, dead-end elimination seeks to systematically eliminate bad rotamers and combinations of rotamers until a single solution remains. Unlike SCMF, the theoretical basis for DEE proves that, if DEE converges, the solution is the GMEC with no uncertainty (Desmet *et al*., 1992). It is a necessary criterion for DEE that the energy description is pairwise as described in equation (6-1) and the effectiveness of the search is, in part, due to the distribution of interactions that arise in a protein side-chain system (Goldstein, 1994).

DEE is fundamentally based on the following physical concept. Consider two rotamers, i_r and i_t, at residue i and the set of all other rotamer configurations $\{S\}$ at all

residues excluding i of which rotamer j_s is a member. If the pairwise energy contributed between i_r and j_s is higher than the pairwise energy between i_t and j_s for all $\{S\}$, then i_r cannot exist in the GMEC and can be eliminated. This notion is expressed mathematically by the inequality

$$E(i_r)+\sum_{j\neq i}^{N}E(i_r,j_s)>E(i_t)+\sum_{j\neq i}^{N}E(i_t,j_s)\quad\forall\{S\}. \tag{6-6}$$

If the above is true, the single rotamer i_r can be eliminated (Desmet *et al*., 1992). In this form, (6-6) is not computationally tractable because, to make an elimination, it is required that the entire sequence/rotamer space be enumerated. To simplify the problem, the bounds implied by (6-6) can be utilized:

$$E(i_r)+\sum_{j\neq i}^{N}\min_{s}E(i_r,j_s)>E(i_t)+\sum_{j\neq i}^{N}\max_{s}E(i_t,j_s). \tag{6-7}$$

Using an analogous argument, Equation (6-7) can be extended to the elimination of pairs of rotamers inconsistent with the GMEC. This is done by determining that a pair of rotamers i_r at residue i and j_s at residue j, always contribute higher energies than rotamers i_u and j_v with all possible rotamer combinations $\{L\}$. Similar to (6-7), the strict bound of this statement is given by

$$\varepsilon(i_r,j_s)+\sum_{k\neq i,j}^{N}\min_{t}\varepsilon(i_r,j_s,k_t)>\varepsilon(i_u,j_v)+\sum_{k\neq i,j}^{N}\max_{t}\varepsilon(i_u,j_v,k_t), \tag{6-8}$$

where ε is the combined energies for rotamer pairs

$$\varepsilon(i_r,j_s)=E(i_r)+E(j_s)+E(i_r,j_s) \tag{6-9}$$

and

$$\varepsilon(i_r,j_s,k_t)=E(i_r,k_t)+E(j_s,k_t). \tag{6-10}$$

This leads to the doubles elimination of the pair of rotamers i_r and j_s, but does not eliminate the individual rotamers completely as either could exist independently in the GMEC. The doubles elimination step reduces the number of possible pairs (reduces *S*) that need to be evaluated in the right-hand side of Equation (6-7) henceforth allowing more rotamers to be individually eliminated.

The singles and doubles criteria presented by Desmet *et al*. fail to discover special conditions that lead to the determination of more dead-ending rotamers. For instance, it is possible that the energy contribution of rotamer i_t is always lower than i_r without the maximum of i_t being below the minimum of i_r. To address this problem, Goldstein presented a modification of the criteria that determines if the energy profiles of two rotamers cross (Goldstein, 1994). If they do not, the higher energy rotamer can be determined to be dead-ending. The improved criterion for singles is

$$E(i_r) - E(i_t) + \sum_{j \neq i}^{N} \min_{s} \left[E(i_r, j_s) - E(i_t, j_s) \right] > 0 \qquad (6\text{-}11)$$

and likewise for doubles,

$$\varepsilon(i_r, j_s) - \varepsilon(i_u, j_v) + \sum_{k \neq i,j}^{N} \min_{t} \left[\varepsilon(i_r, j_s, k_t) - \varepsilon(i_u, j_v, k_t) \right] > 0 \quad . \qquad (6\text{-}12)$$

In computational time, the doubles calculation is significantly more expensive than the singles calculation. To accelerate the process, computationally inexpensive methods have been developed to predict the doubles calculations that will be the most productive (Gordon and Mayo, 1998). These modifications, collectively referred to as fast doubles, significantly improved the speed and effectiveness of DEE.

The probability of successfully finding the GMEC has been shown to improve by utilizing an expanded rotamer library and including an initial energy threshold (De

Maeyer *et al*., 1997). For these calculations, we use an energy cutoff of 1000 kcal/mol. Single or double rotamers that produce energies above this threshold are automatically flagged as dead-ending. These values are considered conservative and ensure that the minimum energy found is the GMEC. Parameters that are more aggressive can be used to improve the speed of DEE, but accuracy is sometimes lost.

Several additional modifications collectively enhance DEE further. Rotamers from multiple residues can be combined into so-called super-rotamers to prompt further eliminations (Desmet *et al*., 1994; Goldstein, 1994). This has the advantage of eliminating multiple rotamers in a single step. In addition, it was shown that "splitting" the conformational space between rotamers improves the efficiency of DEE (Pierce *et al*., 2000). Splitting handles the following special case. Consider rotamer i_r. If a rotamer i_{t1} contributes a lower energy than i_r for a portion of the conformational space, and a rotamer i_{t2} has a lower energy than i_r for the remaining fraction, then i_r can be eliminated. This case would not be detected by the less sensitive Desmet or Goldstein criteria. In the implementation used in this study, all of the enhancements described above were combined into a single computational approach. Because of these improvements, convergence to the GMEC in less than 1/2 hour on a single processor can now be expected for side chain placement calculations on proteins in excess of 300 residues.

3. Materials and Methods

We use a test set of 20 protein structures (Table 6-1) from the Brookhaven Protein Databank (Bernstein *et al*., 1977) that was compiled previously by Carrando and co-workers (Mendes *et al*., 1999). This set was chosen due to the high resolution of the

structures and the inclusion of a wide distribution of sizes and structure types (Table 6-1). For the side chain placement calculations, there are positions in the fixed backbone where all allowed rotamers cause steric clashes that lead to unrealistically high energies. These positions were not considered in these calculations. The effective number of residues shown in Table 6-1 is the total number of residues minus cysteines involved in disulfide bonds and residues that clash with the backbone. Both alanine and glycine are described by a single rotamer and are therefore not taken into account when comparing the accuracy of the search algorithms for the side chain placement calculations. The modeled number of residues is determined by the effective number of residues minus wild-type alanine and glycine positions.

We use the DREIDING force field parameters for the atomic radii and internal coordinate parameters (Mayo *et al*., 1990). The van der Waals energies are modeled using a 6-12 Leonard-Jones potential with an additional 0.9 scale factor applied to the atomic radii to soften the lack of flexibility implied by the fixed backbone and the rotamer descriptions (Dahiyat and Mayo, 1997b). Solvation terms were not included in order to accelerate energy matrix calculations. The rotamer library is backbone-dependent as described by Dunbrack and Karplus (Dunbrack and Karplus, 1993). The following modifications were included as previously described (Dahiyat *et al*., 1997a). χ_1 and χ_2 angle values of rotamers for all aromatic amino acids, and χ_1 angle values for all other hydrophobic amino acids were expanded ±1 standard deviation about the mean value reported in the Dunbrack and Karplus library. The χ_3 angles that were undetermined from the database statistics were assigned the following values: Arg, –60°, 60°, and 180°; Gln, –120°, –60°, 0°, 60°, 120°, and 180°; Glu, 0°, 60°, and 120°; Lys,

−60°, 60°, and 180°. The χ_4 angles that were undetermined from the database statistics were assigned the following values: Arg, −120°, −60°, 60°, 120°, and 180°; Lys, −60°, 60°, and 180°. Rotamers with combinations of χ_3 and χ_4 angles resulting in sequential g^+/g^- or g^-/g^+ angles were eliminated. Uncharged His rotamers were used.

The calculations were performed on a SGI Origin 2000 supercomputer with 32 R10000 processors running at 195 MHz. While the codes for both DEE and SCMF are written to utilize parallel capabilities, the times presented for all algorithms are based on a single processor. The complete energy matrices (all pairwise interactions, $E(i_r,j_s)$ in Equation 6-1) were computed prior to the optimization procedure. The time required for this step is independent of the search algorithm and was approximately 60 minutes on a single processor for each test case.

4. Results and Discussion

4.1. Side-Chain Placement

DEE converges to the GMEC rapidly for the entire side chain placement test set thereby providing the standard to which the solutions found by other methods can be compared. The results are shown in Table 6-2 and Figures 6-1 and 6-2. We found that MC and SCMF consistently perform the worst with the average fraction of incorrect rotamers $\langle f \rangle$ = 0.23 and 0.12 and the average difference in energy from the GMEC $\langle \Delta E \rangle$ = 5.6 and 5.9 kcal/mol, respectively. It is interesting that SCMF consistently gave solutions that have fewer incorrect rotamers, but worse energies than MC indicating that the methods are failing by different mechanisms. We believe MC does poorly because it becomes easily trapped by rotamer combinations that are relatively low in energy

whereas SCMF has difficulty converging the probability of a single rotamer to unity at certain sequence positions. The GA performed better: $\langle f \rangle$ = 0.09 and $\langle \Delta E \rangle$ = 4.3 kcal/mol. The MCQ outperformed the other methods with $\langle f \rangle$ = 0.05 and $\langle \Delta E \rangle$ = 1.3 kcal/mol. The ability for MCQ to give reasonable solutions indicates that there is no benefit to the more complex GA or SCMF methods. The 2hbg structure was the only case out of twenty where SCMF outperformed MCQ in both f and ΔE.

The relationship between the size of sequence/rotamer space and the number of incorrect rotamers predicted by the algorithms was determined (data not shown). As expected, MC showed the greatest correlation (R^2 = 0.81) because it is a sampling algorithm and as the size of the search space increases and the number of cycles remains fixed, the fraction of the space searched decreases. Both SCMF and GA do not fit as well (R^2 = 0.27 and 0.20) indicating that there are other aspects of the energy landscape that impede their search, such as the strength and distribution of coupling interactions. It has been suggested that the advantage of SCMF is that it provides solutions for problems for which DEE does not converge (Koehl and Delarue, 1996). As shown in Table 6-3, this is not necessary for side chain placement calculations as the recent improvements in DEE have allowed it to converge on solutions in times comparable to SCMF. The times for both the MC/MCQ and GA runs presented here are arbitrary because the algorithms could be run indefinitely and better solutions might be obtained. Here, we ran a fixed number of cycles, making the larger proteins appear to take longer.

The results of the side chain placement calculations strongly suggest that there is no compelling reason to use an algorithm other than DEE for side chain placement as it consistently and quickly converges to the GMEC. However, as design calculations

become more complex, there is a point beyond which DEE will not converge in a reasonable amount of time. To solve these problems, it is necessary to trade the accuracy of DEE for the speed of SCMF or MCQ.

4.2. Sequence Design

For the protein sequence design comparisons, amino acid substitutions are allowed at the designed positions while the side chains for the remaining residues are floated as in the side chain placement calculations. By specifying more positions to be designed, the difficulty of the problem can be tuned from the easier side chain placement calculation to an intractable full sequence design. Because the GA and MC methods rarely outperformed MCQ, they are not run on the more difficult design problems. While DEE performed extraordinarily well on the side chain placement calculations, the time to convergence explodes as the number of designed residues reached some threshold (Table 6-4 and Figure 6-3). In contrast, the times required by the other algorithms scale linearly with increasing problem size. This is notably true for SCMF as the time to solve even large design problems is often less than thirty minutes on a single processor. MCQ is allowed to run for the same number of cycles as the side chain placement calculations and requires between 60–120 minutes to complete. However, we observe that it is highly unlikely that either SCMF or MCQ provides a solution that is the GMEC.

Two important questions arise from this conflict. First, if DEE does not converge, which alternate method will produce the best results? Second, at the point which DEE explodes, how incorrect are the solutions given by the less accurate algorithms? Because DEE gives the GMEC, the accuracy of SCMF and MCQ can be compared as the design

problem increases in complexity. By extrapolating this result into the region where DEE explodes, the accuracy of the other algorithms can be reasonably approximated.

To determine the relationship between these questions and the specifics of the design problem, we ran tests on five structurally different proteins. Cytochrome b_{562} (256b) and hemoglobin (2hbg) are primarily α-helical proteins, amicyanin (1aac) and plastocyanin (1plc) are primarily β-sheet proteins, and ribonuclease (9rnt) is a protein that contains both α-helical and β-sheet structures. To ensure that we studied proteins where the success of the algorithms varied for the side chain placement calculations, we included 2hbg in the design test set. This represents one of the few proteins where SCMF performed better than MCQ for the side chain placement calculations.

The residues of each protein are labeled core, boundary, or surface based on solvent accessibility (Dahiyat and Mayo, 1997a). From the perspective of protein design, this partition is motivated by the need for a hydrophobic core and hydrophilic surface for stable folding. For the design calculations, the hydrophobic amino acids (A, V, L, I, F, Y, W) are considered in the core, the hydrophilic amino acids (A, S, T, D, N, E, Q, H, K, R) at the surface and the combination of both sets in the boundary. The remaining four amino acids (G, C, P, M) are omitted from these calculations. The protein core has been the target of most design efforts as this region tends to be an easier calculation due the primary dependence on the sterics of side chain packing (Lee and Levitt, 1991; Desjarlais and Handel, 1995; Dahiyat and Mayo, 1996; Dahiyat and Mayo, 1997b; Lazar *et al.*, 1997; Su and Mayo, 1997). More recently, computational protein design has expanded successfully into the boundary and surface regions (Malakaukas and Mayo, 1998;

Dahiyat *et al.*, 1997) and to complete sequence design (Dahiyat *et al.*, 1997a; Morgan and Mayo, unpublished results).

Protein length is not a good indicator of problem difficulty as the number of rotamers allowed at each residue, the specific conformation of the backbone, and the particular choice of the force field can make a problem difficult. We have found that increasing the number of design positions qualitatively makes the search problem more difficult. For each test protein, we complete three series of runs where design positions are added in sequence order from the core, boundary, or surface. To address the concern that the results are dependent on the order in which the design residues are added, we ran a separate series of runs for the core of 1aac. In these runs, we added design residues in structural clusters rather than sequence order. We find that the accuracy of SCMF and MCQ as a function of the number of design positions does not change qualitatively with the order of addition. This result can be generalized to the boundary and surface as, in these regions, residues are separated by greater distances and coupling is less likely to affect the results.

DEE tends to converge on problems containing more design residues for the surface and boundary than the core with the exception of the surface of 256b, which diverges after 8 residues are added (Table 6-4). This result is somewhat dependent on the order in which the design residues are added. There is usually a specific combination of positions that cause DEE to fail. When these residues are designed, the time explosion is observed. The apparent ability to design more positions on the surface than in the core is due to the presence of a higher fraction of deleterious combinations in the core region. This is an expected result as the core contains more coupled interactions.

For protein design, the relevant measure of accuracy is not the average fraction of incorrect rotamers as for the side chain placement comparisons. Because the sequence is designed as well as the specific rotamer conformation, it is more interesting to know the fraction of amino acids that are predicted incorrectly *a* as compared to the GMEC. The results for intermediate and hard design problems are shown for the core (Table 6-5), boundary (Table 6-6), and surface (Table 6-7). The hard design problem represents the case where the time required by DEE diverges. Tables 6-5, 6-6, and 6-7 are representative of only two test runs; the following averages are calculated from the complete trajectories (as in Figure 6-4) over the entire sequence design test set. For the core, a = 0.07 ($\langle \Delta E \rangle$ = 14.3 kcal/mol) for SCMF and a = 0.04 ($\langle \Delta E \rangle$ = 1.1 kcal/mol) for MCQ. For the boundary, a = 0.28 ($\langle \Delta E \rangle$ = 7.1 kcal/mol) for SCMF and 0.32 ($\langle \Delta E \rangle$ = 4.6 kcal/mol) for MCQ. The algorithms performed the worst on surface calculations with a = 0.37 ($\langle \Delta E \rangle$ = 15.1 kcal/mol) for SCMF and a = 0.44 ($\langle \Delta E \rangle$ = 8.7 kcal/mol) for MCQ. Similar to the side chain placement calculations, SCMF obtains a solution that is more accurate in amino acid sequence, but higher in energy than MCQ. It is unclear which is the better answer in practice, as a single bad amino acid substitution can severely disrupt structural integrity whereas combinations of mutations, which together contribute an energy comparable to the GMEC, may be more acceptable. There is no observed dependence on the secondary structure of the protein as the results for both the α-helix and β-sheet dominated backbones are qualitatively similar.

We observe that the accuracy of MCQ and SCMF drops rapidly in boundary and surface calculations. This is also related to the increase in the number of rotamers allowed at each position. MCQ fails because the size of the search space rapidly increases

while the number of cycles remained fixed, thereby allowing less space to be sampled. One explanation for the failure of SCMF is through the compounding of two mechanisms. First, the mean-field description of the energy landscape is approximate, leading to error. Second, SCMF must converge the probability of a single rotamer existing in the GMEC to close to unity. As the number of rotamers is increased at each position, the probability that SCMF cannot converge on a single rotamer also increases, leading to incorrect assignments. It is the second of these effects that causes the loss of accuracy in the surface and boundary regions due to the increase in the number of rotamers allowed at each position.

Through these results, we show that the underlying premise of SCMF is erroneous. The global minimum of the annealed mean-field landscape fails to correspond to the true global minimum. However, to solve problems in the regime where conservative DEE fails, it could be argued that this is a necessary trade-off. While our results demonstrate that mean-field theory does not accurately find the GMEC, this should not be taken as a blanket disqualification of its utility. For example, the calculation of rotomer probabilities is useful in determining the entropy (and free energy) of the sequence (Koehl and Delarue, 1996). However, we have shown it does not accurately find the GMEC of the system. Accurately finding the GMEC is an essential step in the protein design algorithm. Because approximate computational results may be artifacts, it is possible to draw erroneous conclusions about the quality of the design strategy. This is particularly a problem when the combinatorial complexity is high and there is a high density of low energy configurations. In such a case, it is possible to be

close to the true global minimum in energy and have a completely different amino acid sequence.

In this study, we use extremely conservative DEE parameters (1000 kcal/mol threshold for automatic determination of dead-ending rotamers and pairs of rotamers). This was done to ensure that the solution obtained is the GMEC. Most practical design calculations are performed using more aggressive parameters (20 kcal/mol threshold for automatic determination of dead-ending rotamers and 1000 kcal/mol threshold for pairs of rotamers) or highly aggressive parameters (-20 kcal/mol threshold for automatic determination of dead-ending rotamers and -20 kcal/mol threshold for pairs of rotamers). A negative value indicates that the threshold is taken from the minimum rotamer energy at each residue position rather than zero (De Maeyer *et al*., 1997). To test the accuracy of DEE with the moderately and highly aggressive parameters, calculations were performed on the most difficult design problems. In the best case, DEE converged to the same solution up to 15 times more quickly. However, the effect on convergence time was generally unpredictable.

As another option, the quality of solution produced by MCQ can theoretically be improved by running for longer time periods. In our hands, MCQ is typically run for 1000 cycles with each cycle consisting of 10^6 moves, requiring 3,000 – 10,000 minutes. We ran MCQ on three difficult design problems: 30 surface design positions of 2hbg, 24 boundary positions of 1plc, and 17 core positions of 256b. For each case, the number of incorrect amino acids for 20 cycles versus 1000 cycles is 15/10, 8/2, and 2/0. The additional time clearly improved the results for MCQ. It has been suggested that an improvement in SCMF can be achieved by initializing the rotamer probability vector

randomly and running the convergence algorithm for each random initialization (Mendes *et al*., 1999). This has the effect of extending the run time of SCMF. We implemented this algorithm and found that it never improved the solution. Most likely, the improvement that was observed by Mendes *et al*. was due to their lack of use of temperature annealing. Increasing the run length of SCMF was not found to improve the solution and the aggressiveness had been previously optimized through the energy threshold. Both aggressive DEE and MCQ comprehensively produce better results on the more difficult design problems.

Currently, the SCMF and DEE algorithms can only be applied to pairwise energy functions (Equation 6-1). Higher-order terms may be important in determining the stabilization energy of a sequence. In particular, buried surface area is a higher-order contribution to energy. In their present form, the stochastic methods can easily incorporate higher-order energy terms. If the incorporation of higher than two-body terms is desired, a new trade-off is created. The reduction of the force field to the pairwise form that is required for the deterministic methods must be weighed with the inaccuracy of the stochastic search methods.

Of the four search algorithms we studied, there are extensive variations in literature. Our lab uses the MC, MCQ, SCMF, and DEE algorithms and the specific formulations presented in this paper represent the methods that we have found previously to be the most successful. The exception is the genetic algorithm, which was programmed solely for this study. We tried many versions and found that the algorithm used here is the most reliable over the entire test set.

In this paper, we study each algorithm as a stand-alone search technique. An alternative is to create hybrid algorithms that combine elements from different techniques. For instance, the addition of a Monte Carlo quench step to the GA and SCMF algorithms improves the solution. In the case of the GA, the quench step does not improve the algorithm beyond what is attainable with MCQ. In addition, the combination of DEE and MC can potentially improve the search. Finally, a new branch-and-terminate technique has been proposed which extends the capabilities of DEE (Gorden and Mayo, 1999).

5. Conclusions

We have shown that DEE is the most appropriate search algorithm for side chain placement as it consistently and rapidly converges to the GMEC for a full range of structure sizes and types. However, for the design calculations, there is a point beyond which DEE fails to converge and it becomes necessary to utilize a less accurate method to obtain a solution. We find that the accuracy and speed of SCMF and MCQ are comparable for the design calculations. Both methods give reasonable solutions in the core, but fail considerably in the boundary and surface regions. The advantage of MCQ relative to SCMF is that, because it is a stochastic method, it can be run for longer periods of time so better solutions might be obtained. In contrast, the answer provided by SCMF is the only solution that it will provide and therefore does not take advantage of the increasing capability of computer hardware and software. Experimentally, the utility of an imprecise answer is unclear. Because the energy landscape constructed by the force field is not the actual landscape, an answer that is close to the theoretical GMEC may be

adequate to provide a folded, stable structure. Nevertheless, it is clearly important to understand that a solution provided by SCMF or MCQ could be off by more than 20 kcal/mol in computed energy and produces sequences that have 40% disparate amino acids from the optimum for the given force field.

Table 6-1. Overview of the 20 protein test set

	No. residues[a]				Secondary structure			Solvent access[c]		
PDB	total	effect	model	log c[b]	turn	beta	alpha	core	bound	surf
1aac	105	104	85	86	46	58	0	30	31	43
1cbn	46	40	31	31	13	6	21			
256b	106	106	88	101	16	0	89	28	34	44
1isu	62	62	43	46	45	12	9			
5rxn	54	54	48	48	41	13	0			
1arb	263	250	191	193	119	98	33			
2hbg	147	147	98	110	24	0	123	51	45	51
1bpi	58	52	40	49	25	14	13			
1igd	61	61	51	57	14	30	17			
1cex	213	186	166	158	55	33	98			
1xnb	185	176	144	157	41	125	10			
1plc	99	99	83	86	55	44	0	27	27	45
1ptx	64	56	51	50	31	14	11			
2erl	40	32	24	31	8	0	24			
2end	137	136	118	138	62	4	70			
1amm	174	173	157	174	78	81	14			
1whi	122	120	99	111	49	57	14			
9rnt	104	103	86	87	49	35	19	36	29	38
2ihl	129	119	100	102	42	20	57			
1ctj	89	87	63	65	46	0	51			

[a] The effective and modeled residues are as described in Section 3.

[b] The number of configurations is the total number of points in rotamer space for the homology calculations. The number of configurations increases dramatically for the design calculations (data not shown).

[c] The designation of a residue as core, boundary or surface was done in the following manner. A solvent-accessible surface was generated using the Connolly algorithm with a probe radius of 8.0 Å, a dot density of 10 $Å^{-2}$, and a C_α radius of 1.95 Å. A residue was classified as a core position if the distance from its C_α, along its C_α-C_β vector, to the solvent-accessible surface was greater than 5.0 Å, and if the distance from its C_β to the nearest surface point was greater than 2.0 Å. The remaining residues were classified as surface positions if the sum of the distances from their C_α, along their C_α-C_β vector, to the solvent-accessible surface plus the distance from their C_β to the closest surface point was less than 2.7 Å. All remaining residues were classified as boundary. This classification was only necessary for the test cases chosen for sequence design calculations.

Table 6-2. Results of Side Chain Placement Calculations

	Number rotamers incorrect [a]				ΔE (kcal/mol) [a]			
	SCMF	MCQ	MC	GA	SCMF	MCQ	MC	GA
1aac	7	7	26	15	0.6	0.8	6.2	4.6
1cbn	0	0	4	0	0.0	0.0	0.5	0.0
256b	23	9	29	21	14.0	0.4	7.8	8.9
1isu	1	0	5	0	0.2	0.0	0.4	0.0
5rxn	3	0	12	3	0.2	0.0	1.9	0.5
1arb	11	7	34	8	8.4	1.8	10.0	32.0
2hbg	6	10	37	8	0.7	10.8	17.2	1.7
1bpi	10	3	5	5	6.8	0.4	1.0	4.1
1igd	8	3	12	9	1.3	0.5	2.4	2.1
1cex	6	12	38	5	2.2	2.0	10.9	1.4
1xnb	20	4	34	9	37.6	0.7	8.5	0.8
1plc	6	0	21	0	6.8	0.0	3.2	0.0
1ptx	4	0	6	0	0.2	0.0	0.8	0.0
2erl	8	2	6	2	2.3	0.1	0.8	2.8
2end	19	11	28	21	13.7	4.7	12.2	13.9
1amm	20	6	36	9	12.2	1.5	10.4	1.9
1whi	15	10	33	20	3.2	1.9	7.0	8.4
9rnt	13	2	16	10	1.7	0.0	2.1	1.0
2ihl	10	7	24	10	2.2	0.6	5.4	2.3
1ctj	12	0	17	5	3.6	0.0	3.2	0.4
ave[b]	0.12	0.05	0.23	0.09	5.9	1.3	5.6	4.3

[a] The solution as compared to the GMEC found by DEE.
[b] The average fraction of rotamers incorrect and the average energy above the GMEC in kcal/mol.

Table 6-3. Times for Side Chain Placement Calculations

	Time (min)			
	DEE	SCMF	MC/Q [a]	GA
1aac	0.2	0.7	53.9	42.0
1cbn	0.0[b]	0.0[b]	7.1	6.1
256b	5.1	0.5	62.8	44.1
1isu	0.0[b]	0.1	16.5	14.5
5rxn	0.1	0.1	12.9	11.1
1arb	26	10.5	402.7	638.2[c]
2hbg	1.0	3.0	120.4	83.3
1bpi	0.0[b]	0.2	12.4	10.6
1igd	0.2	0.3	17.7	14.5
1cex	2.4	6.9	172.9	107.1
1xnb	2.7	3.2	148.9	107.5
1plc	0.2	0.6	48.9	37.3
1ptx	0.0[b]	1.6	12.1	10.2
2erl	0.0[b]	0.1	5.0	4.2
2end	13.8	6.2	118.8	74.3
1amm	1.9	9.6	212.7	158.1
1whi	1.0	3.3	81.3	55.6
9rnt	0.1	0.8	50.6	39.6
2ihl	0.1	1.1	74.7	54.0
1ctj	0.1	0.3	33.2	28.8
average	2.7	2.5	83.3	77.1

[a] The MC quench step requires insignificant additional time.
[b] Less than 0.05 min was recorded.
[c] The number of cycles was increased to 15 (see Section 2.2).

Table 6-4. DEE Explosion Behavior for Protein Sequence Designs

	number of design positions before explosion is observed[a]		
	core	boundary	surface
1aac	24	25	34
9rnt	36[b]	30[b]	38
256b	17	24	8
2hbg	18	28	27
1plc	18	24	25[b]

[a] Defined as the number of design residues at which a time was observed greater than 500 minutes on a single processor.
[b] No time explosion was observed. The largest number of attempted design positions is reported.

Table 6-5. Core Results for Sequence Design Calculations

		# incorrect [a]			time (min)		
	design [b]	DEE	MCQ	SCMF	DEE	MCQ	SCMF
1aac	8	0	0	0	5	57	3
	24	0	0	4	5382	73	13
9rnt	12	0	0	0	2	55	3
	36	0	1	3	71	76	71
256b	10	0	0	1	232	59	4
	17	0	2	4	7271	63	7
2hbg	9	0	2	2	6	120	4
	18	0	2	2	9999	109	7
1plc	9	0	2	1	15	50	3
	18	0	3	3	1704	60	4
ave[c]					52	68	3
					4885	76	20

[a] The solution as compared to the GMEC determined by DEE
[b] The number of sequence design positions. A representative example is shown for a medium and hard calculation. The hard calculation corresponds to the point at which a time explosion was observed for DEE.
[c] The averages are taken for the medium and hard design problems.

Table 6-6. Boundary Results for Sequence Design Calculations

		# incorrect [a]			time (min)		
	design [b]	DEE	MCQ	SCMF	DEE	MCQ	SCMF
1aac	15	0	5	6	68	62	13
	25	0	7	12	917	79	40
9rnt	5	0	0	0	1	53	3
	30	0	11	9	157	84	71
256b	8	0	1	4	51	71	14
	24	0	11	11	1855	90	74
2hbg	12	0	3	0	25	130	21
	28	0	9	12	1153	168	87
1plc	9	0	3	1	12	50	6
	24	0	8	8	599	64	20
ave[c]					31	73	11
					936	97	58

[a] The solution as compared to the GMEC determined by DEE
[b] The number of sequence design positions. A representative example is shown for a medium and hard calculation. The hard calculation corresponds to the point at which a time explosion was observed for DEE.
[c] The averages are taken for the medium and hard design problems.

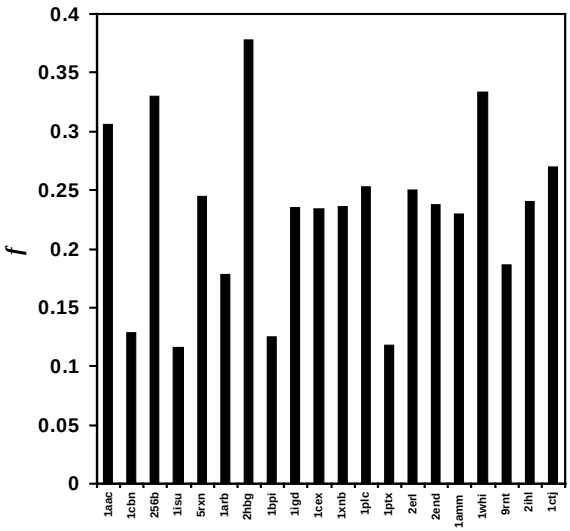
Table 6-7. Surface Results for Sequence Design Calculations

		# incorrect [a]			time (min)		
	design [b]	DEE	MCQ	SCMF	DEE	MCQ	SCMF
1aac	15	0	8	5	3	61	18
	34	0	14	14	918	75	58
9rnt	15	0	8	9	81	58	8
	38	0	21	18	870	79	63
256b	3	0	2	2	18	65	4
	8	0	5	5	2329	66	10
2hbg	9	0	2	4	47	108	8
	30	0	15	15	2006	141	41
1plc	10	0	2	2	13	53	4
	25	0	10	9	151	63	19
ave[c]					162	69	8
					1255	85	38

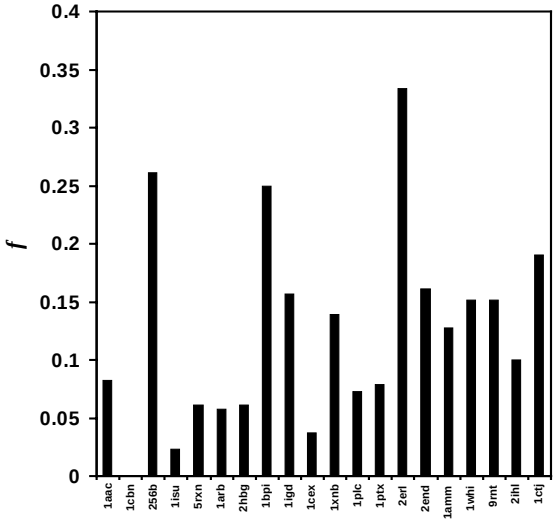
[a] The solution as compared to the GMEC determined by DEE
[b] The number of sequence design positions. A representative example is shown for a medium and hard calculation. The hard calculation corresponds to the point at which a time explosion was observed for DEE.
[c] The averages are taken for the medium and hard design problems.

Figure 6.1:

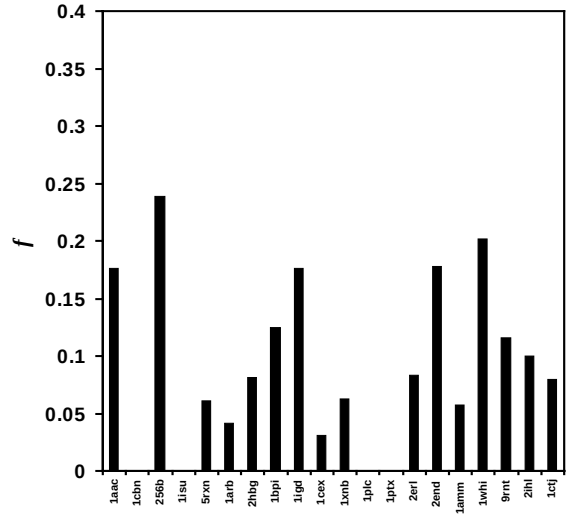
The GMEC was determined by DEE and compared to the result given by the other algorithms for the side chain placement calculations. The fraction of incorrect rotamers *f* predicted by **(A)** MC, **(B)** SCMF, **(C)** GA, and **(D)** MCQ is shown for each protein in the test set, as compared to the GMEC found by DEE.


f
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1aac
1cbn
256b
1isu
5rxn
1arb
2hbg
1bpi
1igd
1cex
1xnb
1plc
1ptx
2erl
2end
1amm
1whi
9rnt
2ihl
1ctj

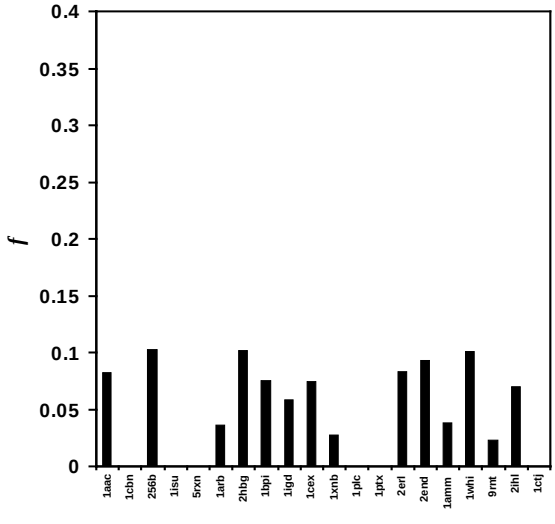
(A)


f
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1aac
1cbn
256b
1isu
5rxn
1arb
2hbg
1bpi
1igd
1cex
1xnb
1plc
1ptx
2erl
2end
1amm
1whi
9rnt
2ihl
1ctj

(B)


f
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1aac
1cbn
256b
1isu
5rxn
1arb
2hbg
1bpi
1igd
1cex
1xnb
1plc
1ptx
2erl
2end
1amm
1whi
9rnt
2ihl
1ctj

(C)


f
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1aac
1cbn
256b
1isu
5rxn
1arb
2hbg
1bpi
1igd
1cex
1xnb
1plc
1ptx
2erl
2end
1amm
1whi
9rnt
2ihl
1ctj

(D)

Figure 6-2:

The difference between the GMEC found by DEE and the energy determined by **(A)** MC, **(B)** SCMF, **(C)** GA, and **(D)** MCQ for the side chain placement calculations.

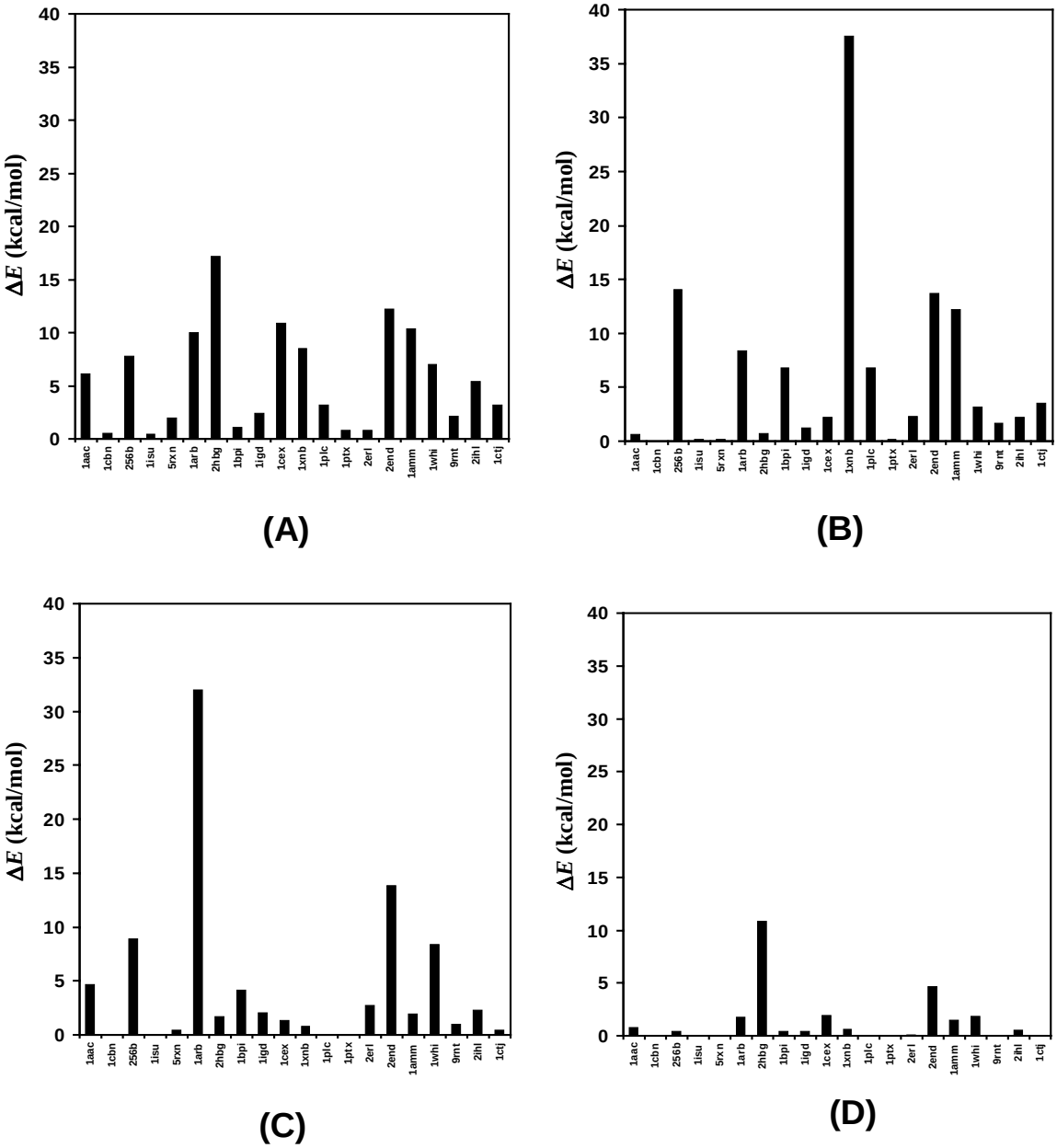

ΔE (kcal/mol)
0
5
10
15
20
25
30
35
40
1aac
1cbn
256b
1isu
5rxn
1arb
2hbg
1bpi
1igd
1cex
1xnb
1plc
1ptx
2erl
2end
1amm
1whi
9rnt
2ihl
1ctj
(A)
(B)
(C)
(D)

Figure 6-3:

The time for DEE to converge on the 2hbg test case plotted against the number of design positions for the core (thick solid), boundary (thin solid), and surface (dotted) regions. This graph is representative of the time explosion behavior of the remaining four sequence design test cases. Note that the y-axis is truncated at 2000; the time for the hardest core calculation continues to 9999 minutes (Table 6-5).

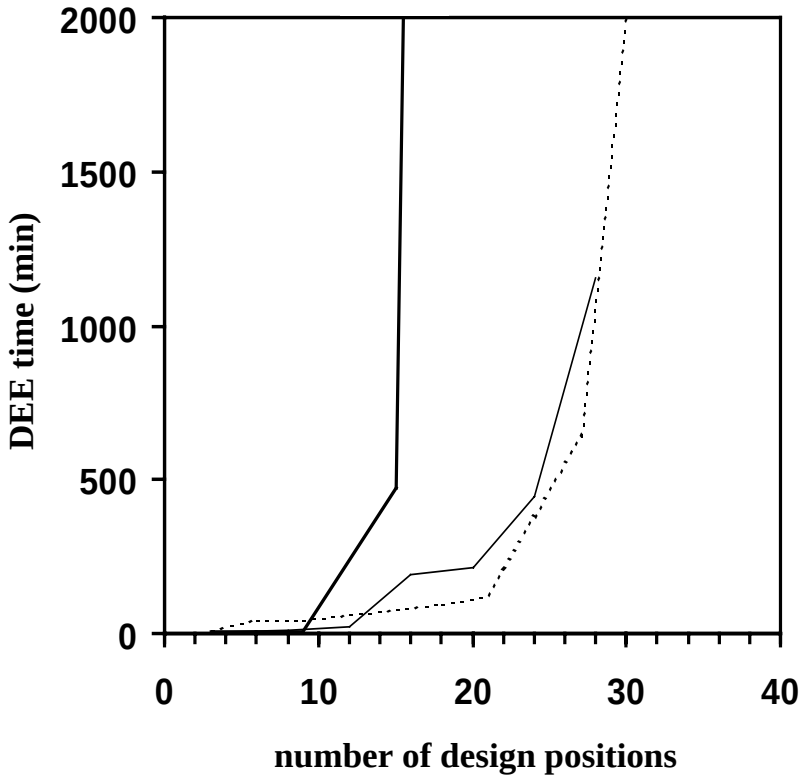
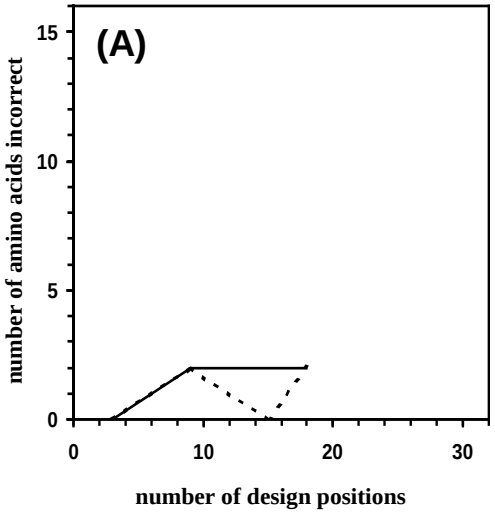
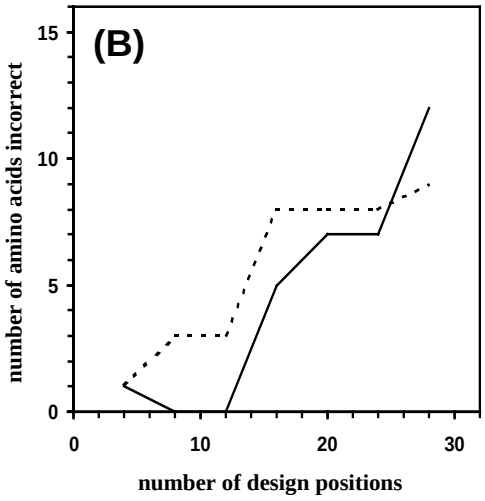
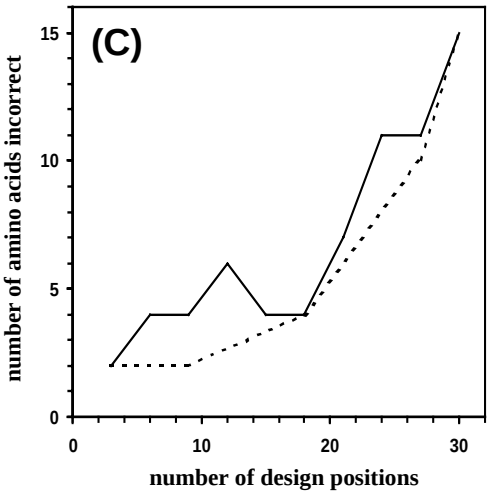

2000
1500
1000
500
0
DEE time (min)
0
10
20
30
40
number of design positions

Figure 6-4:

The results of the sequence design test in the **(A)** core, **(B)** boundary, and **(C)** surface regions for 2hbg. The number of amino acids incorrect with respect to the GMEC found by DEE is plotted against the number of design positions in the protein for the SCMF (solid line) and MCQ (dotted line) calculations. The end of each line either corresponds with the DEE explosion as shown in Figure 6-4 and Table 6-4 or the point at which all of the regional positions have been designed. Note that 2hbg is one of the two cases out of twenty where SCMF outperformed MCQ for the side chain placement tests.


(A)
number of amino acids incorrect
0
5
10
15
0
10
20
30
number of design positions


(B)
number of amino acids incorrect
0
5
10
15
0
10
20
30
number of design positions


(C)
number of amino acids incorrect
0
5
10
15
0
10
20
30
number of design positions

Chapter 7

Evolvable Systems in Biology

Portions of this chapter are reproduced from:
Voigt, C. A., Mayo, S. L., Wang, Z-G., and Arnold, F. H. (2002). Directing the evolvable: Utilizing robustness in in vitro evolution, In: Robustness, Ed: Erica Jen, Santa Fe Institute Press

1. Introduction

The use of directed evolution techniques has greatly accelerated the discovery of new and useful biological molecules and systems (Arnold, 2001a). Through iterative cycles of diversity creation (e.g., mutation or recombination) and selection, proteins, antibodies, pathways, viruses, and organisms have been evolved to perform tasks optimized for pharmaceutical and industrial applications. Before directed evolution became established, it was unclear how successful such an approach would be. It was not obvious that randomized mutagenesis and selection would find improvements, due to combinatorial explosion in the number of possible offspring and the observation that few of these are functional, much less have improvements in desired properties.

Directed evolution is successful, in part, because prior to being evolved *in vitro* these systems have a long history of evolution *in vivo*. As a result of this history, they have properties that make them amenable to both natural and laboratory evolution. This evolvability represents the ability of a system to produce fit offspring in a dynamic environment. This paper will review some of the features that make a system evolvable. A particular emphasis will be made on the contribution of robustness, or the ability for a system to survive perturbations of its internal parameters. Robustness enhances the ability

of a population to sample parameter space, thus enabling the discovery of novel phenotypes.

Understanding the basis for evolvability will aid the design of efficient evolutionary algorithms that accelerate the *in vitro* discovery process. Achieving this goal will require the combination of computational models with data from *in vitro* evolution experiments. In this review, we describe the initial steps of this effort. First, we provide a general definition of robustness and explain its relationship to evolvability. In the following sections, we apply these ideas: first to the evolution of proteins through mutagenesis and recombination and then to the evolution of genetic networks.

1.1. Robustness

The behavior of a system can be defined by a set of internal parameters. In the case of a metabolic network, the parameters are the kinetic constants and concentrations of the component enzymes, which determine the products and their rates of production. Similarly, the activity of an enzyme is defined by its amino acid sequence, solvent conditions, and temperature. A system is robust if it can absorb variations in these parameters without disrupting its behavior (Savageau, 1971; Savageau, 1972). The parameters can be perturbed by various insults, for example, kinetic constants can be altered by mutations, temperature variations, or exposure to different environments. Because we are describing robustness from the perspective of directed evolution, the insults for these systems are defined by the experimental technique, such as point mutagenesis or recombination.

It is important to contrast parameter robustness with variable stability, which describes resilience to perturbations in the inputs of a system. Stability implies that the state of the system before and after the perturbation remains unchanged. As an illustration, a bridge is stable regarding variables such as car weight. The bridge is considered stable if it reliably returns to the same state after cars have passed over it. In contrast, robustness describes the collection of bridge design parameters, such as the cable strength for a suspension bridge. If the Golden Gate Bridge in San Francisco can be reproduced in Alaska without disturbing its function, then the design is robust. In this example, the strength of the materials is the parameter, and this parameter is perturbed by a change in environment.

In determining the robustness of a system, the behavior and the parameters need to be defined. A system may be robust with regards to one behavior while being sensitive with regards to a different behavior (Barkai and Leibler, 1997). There are several convenient metrics for measuring robustness. For parameters that are continuous, the sensitivity S is defined as the change in a behavior b with respect to a parameter p,

$$S(b, p) = \frac{\mathrm{d}b}{\mathrm{d}p} \tag{7-1}$$

(Savageau, 1971). When the parameter is discontinuous (e.g., amino acid sequence), then it is useful to define an entropy which captures the number of states that are consistent with retaining the system behavior. The entropy of parameter i is defined as

$$s_i = \sum_{j=1}^{20} p_i(s) \ln p_i(s), \tag{7-2}$$

where $p_i(s)$ is the probability that parameter i in state s does not disrupt the behavior of the system. These probabilities can be derived from several sources. They can be

calculated explicitly if the energetic consequence of each state is known. They can also be obtained from a list of possible states, obtained either through a simulation or by experimental observation of variation in the system (e.g., examining amino acid variability through a sequence alignment). A useful application of Equation (7-2) is to quantify the variability of a protein residue *i* with respect to amino acid substitutions *s* (Saven and Wolynes, 1997; Voigt *et al.*, 2001b)

1.2. Evolvability

Evolvability is the capacity of a system to react at the genetic level to changing requirements for survival (Kirshner and Gerhart, 1998). Upon environmental change, an evolvable system will produce offspring whose perturbed parameters improve the new fitness. Those systems that require the least dramatic parameter changes (e.g., the fewest mutations) have the smallest entropic barrier to being discovered (Kirshner and Gerhart, 1998; van Nimwegen, 1999). Architectures that minimize the entropic barriers are the ones that are likely to find improvements first and therefore survive. In directed evolution, the evolutionary constraints *in vitro* differ from those properties that were selected for by nature. With this "environmental change," an evolvable architecture is more likely to result in a successful directed evolution experiment.

Robustness can reduce the entropic barrier by separating the parameters that define the various behaviors of a system. For example, an enzyme may improve the evolvability of its activity by separating those residues that maintain the stability from those that tune the activity. If a residue contributes to both properties, then it would be more difficult to make substitutions to improve one property without degrading the other.

To further reduce the entropic barrier, it is important that small changes in the evolvable parameter lead to large changes in the behavior of the system. Evolvability will be reduced if regions of parameter space that are devoid of any behavior have to be traversed (Figure 7-1). Robustness can also improve evolvability by facilitating the exploration of parameter space through neutral drift (Eigen and McCaskill, 1989; Huynen *et al*., 1996; Huynen, 1998; Kirshner and Gerhart, 1998; Barkai and Leibler, 1997). Neutral drift drastically increases the fraction of parameter space that can be sampled, thus increasing the likelihood of discovering novel behaviors.

Measuring robustness, either by calculating sensitivity or entropy, is fairly straightforward. In contrast, the various ways in which evolvability can manifest itself makes its quantification a more challenging task. For example, an evolvable parameter does not have to be robust. Some measures of evolvability include the number of behaviors that can be sampled and the nature of the transitions between these behaviors. A highly cooperative transition and small separations in parameter space are indicators of evolvability (Figure 7-1).

2. Utilizing Robustness to Optimize Mutant Libraries

In directed evolution, genetic diversity can be tolerated due to the intrinsic robustness of proteins. The ability to predict how and where a protein is robust has led to the design of evolutionary algorithms where point mutations or recombination is targeted (Voigt *et al*., 2001a, Arnold, 2001b, Voigt *et al*., 2002). In describing the robustness of a protein, the behavior describes the combination of properties that needs to be retained for function. This is mainly the stability of the three-dimensional structure, but can be a more

complex combination of properties. For example, in evolving an antibody for use as a pharmaceutical, besides maintaining the stability and affinity for the target, it may also be important to evade the human immune response. When diversity is generated using random point mutagenesis, the relevant parameter space is the amino acid state of each residue. A system that is robust can absorb variation in the parameters (amino acid substitutions) without altering a defined behavior (e.g., stability). The capacity to improve existing functions or discover new ones via amino acid substitutions is the protein's evolvability. In this section, we describe the realization of robustness and evolvability in protein structures. This provides a basis for the introduction of strategies that accelerate evolutionary searches.

2.1. Robust Protein Architectures

The total number and order of interactions between amino acids affect the average robustness of a protein (Figure 7-2). When there are many interactions, there are more constraints that need to be satisfied, thus increasing the probability that a mutation is deleterious. This effect worsens as the order of the interactions increase. For example, a system that is dominated by 2-body interactions is more robust on average than one that is dominated by 3-body interactions (Kauffman and Levin, 1987; Kauffman and Weinberger, 1989). Robustness is also affected by the distribution of interactions in the protein structure. A scale-free distribution of interactions has been demonstrated to be particularly robust (Jeong *et al*., 2000). A property of scale-free distributions is that there are a few, highly interacting residues and many weakly interacting residues. A protein can achieve a scale-free-like distribution of interaction by increasing the ratio of surface

area to volume. Reducing the number of interactions at many residues increases the average robustness of the protein at the cost of making a few residues highly sensitive to perturbation. If a mutation occurs at such a residue, this has a catastrophic effect on the stability. In other words, increasing the sensitivity of a few parameters makes the overall system robust to random, but not directed perturbation, a scenario described as “Highly Optimized Tolerance” (Carlson and Doyle, 2000).

Modularity is also important for structural robustness. This effect was demonstrated using a simple protein lattice model to enumerate the number of sequences that fold into various two-dimensional structures (Li *et al*., 1996). Many sequences folded into a few highly robust structures, whereas most structures were fragile with only a few or no sequences folding into them. The robust structures were found to be modular, with a small robust motif copied throughout the larger structure (Figure 5-7) (Wang *et al*., 2000). The repetition of modular peptide subunits is a common theme in protein structures (Orengo *et al*., 1994).

Another mechanism to improve robustness is to increase the thermal stability of a structure. Many theoretical models have demonstrated that mutational and thermal stability are strongly correlated (Broglia *et al*., 1999; Buchler and Goldstein, 1999; Bornberg-Bauer and Chan, 1999; Mélin *et al*., 1999; Ancel and Fontana, 2000). In understanding this correlation, it is important to note that the adjectives “mutational” and “thermal” describe insults rather than behaviors. The relevant behavior is existing in the state of a folded protein. This behavior can be rewritten as maintaining a large energy gap between the folded ground state and the ensemble of unfolded conformations. The energy

gap can be perturbed either by increasing temperature or by the disruption amino acid interactions via a mutation.

2.2. Measuring Robustness

One feature of robustness is that the fitnesses of sequences close in sequence space are highly correlated so information is retained upon mutation (Weinberger, 1990). Fitness correlation provides an experimentally attainable measure of robustness. This information is typically in the form of a plot of the mutation rate versus the percent of offspring that retain some function (Figure 7-3) (Suzuki *et al.*, 1996; Daugherty *et al.*, 2000). A small initial slope indicates that more mutations can be accumulated without degrading the function, indicating that this function is more robust (Wilke and Adami, 2001).

We have developed a computational algorithm that calculates the structural robustness of a protein (Voigt *et al.*, 2001b). This algorithm calculates the stabilization energy of all amino acid sequences folded onto a specified three-dimensional structure using the ORBIT protein design software to calculate the amino acid interactions (Dahiyat and Mayo, 1997) and mean-field theory to accelerate the calculation (Saven and Wolynes, 1997). The energetic information is condensed into a residue entropy (Equation 7-2), where a high entropy indicates that a residue is tolerant to amino acid substitution. Using this algorithm, those residues that can be mutated while preserving the structural stability can be identified.

2.3. Robustness Improves Functional Plasticity

Proteins are particularly plastic with regard to tuning function and exploring novel function space. This is evidenced by the observation that some common structural motifs are able to perform a wide variety of functions (Orengo *et al*., 1994; Bolon and Mayo, 2001). Directed evolution experiments have demonstrated that significant functional variability can be obtained with few mutations (Shao and Arnold, 1996). One way to achieve functional evolvability is to improve the robustness of the behaviors that are essential for function but are not being optimized. By improving the robustness of a structure, mutations are less likely to be destabilizing and sequence space can be more readily explored for new properties through neutral drift (Aronson *et al*., 1994; Govidarajan and Goldstein, 1997). Indeed, mutagenesis experiments have repeatedly demonstrated that protein structures are amazingly robust regarding mutations (Loeb *et al*., 1989; Rennell *et al*., 1991; Aronsson *et al*., 1994; Axe *et al*., 1996; Baase *et al*., 1996; Huang *et al*., 1996). Besides improving the overall robustness of the structure, there are several additional mechanisms for improving evolvability, including the separation of parameters and the presence of suppressor mutations.

Evolvability can be improved through the separation of parameters that control different behaviors (Kirshner and Gerhart, 1998). This allows one behavior to be optimized without negatively affecting the remaining behaviors. For example, a protein is more evolvable when those residues that maintain stability are isolated from those that control activity. This isolation can be observed when independent mutations that improve different behaviors have additive effects when combined. Additivity has been observed frequently in mutagenesis data and it has been proposed to take advantage of this

property in protein engineering (Wells, 1990; Sandberg and Terwilliger, 1993; Skinner and Terwilliger, 1996).

When several homologous parents have different properties and high sequence identity, functional additivity can be utilized to produce a library of offspring with combinations of properties from the parents. In one such study, 26 parental subtilisin genes were recombined to produce a library of offspring (Ness *et al.*, 1999). The offspring were screened for activity and stability in various conditions, such as acidic or basic environments or in organic solvents. The hybrid proteins in the library demonstrated a broad range of combined properties. Additivity can also be on the level of parameters that control activity, such as individual components of substrate specificity. By separating the effects of the parameters that confer specificity from the requirements imposed by the catalytic mechanism and structural maintenance, these parameters can be perturbed individually to produce offspring with diverse specificities. Additive parameters that confer specificity are apparent in the recombination of two triazine hydrolases (Raillard *et al*., 2001). These two enzymes only differ at 9 residues out of 475, but have very different activities. Recombined offspring were found to catalyze reactions on a variety of triazine compounds with chemically distinct R-groups. Sequence analysis revealed that different residues were important in controlling different physical components of specificity. For example, residues 84 and 92 determine the size of the R-group that could be bound. Mutations at these residues are free to alter aspects of specificity independently without disturbing the catalytic mechanism or stability.

Separating the residues that control activity and stability can be achieved by minimizing the number of stabilizing interactions at functionally important residues. One

way this is manifested is through the prevalence of loops and exposed regions near the active site. These loops often control substrate specificity and have been the frequent target of mutagenesis (Hedstrom *et al*., 1992; Palzkill *et al*., 1994; El Hawrani *et al*., 1994; Burks, *et al*., 1997; Matsumura and Ellington, 2001). Further, the complementarity-determining regions (CDRs) of antibodies are composed of loops and have been shown to be robust (Brown *et al*., 1996; Burks *et al*., 1997). Structural tolerance can be achieved without resorting to loop structures. Patel and Loeb demonstrated that the active site of DNA polymerase I, an antiparallel β-strand, is very tolerant to mutagenesis (2000).

Deleterious mutations can sometimes be overcome by additional, compensating mutations that act to suppress the negative effect (Baase *et al*., 1999; Jucovic and Poteete, 1999). The appearance of compensating mutations is apparent in diagrams that plot the percent of a library that is functional versus the mutation rate (Figure 7-3). Typically, these plots demonstrate an exponential decay proportional to the robustness. However, some of these curves recover at high mutation rates, implying the existence of compensating mutations (Suzuki *et al*., 1996; Daugherty *et al*., 2000). When a single mutation compensates for an extraordinary number of deleterious mutations, it is referred to as a "global suppressor." An example is the M182T mutation in beta-lactamase, which was found to generally compensate for locally destabilizing mutations in a loop near the active site (Huang and Palzkill, 1997; Sideraki *et al*., 2001; Orencia *et al*., 2001). When this mutation is present, it becomes possible to make additional mutations that rearrange the active site without degrading the stability. The trend of initially accumulating mutations that improve the evolvability has been observed in directed evolution

experiments. In the directed evolution of the substrate specificity of beta-glucoronidase, the intermediate mutants first broadened the substrate specificity (Matsumura and Ellington, 2001). After the activity was made more plastic, additional synergistic mutations tuned activity towards the new substrate, and the broad specificity was lost. In another study, the crystal structures of wild-type and evolved esterases were compared (Spiller *et al*., 1999). It was found that several loops that form the entrance to the active site cavity are ordered into a specific conformation by mutations distant from the active site. This initial fixation provided the basis for additional mutations in later generations.

2.4. Targeting Diversity

Algorithms that have been proposed to optimize directed evolution can be separated into two general categories (Voigt *et al*., 2001a). Several methods have been proposed that optimize the mutation rate as a function of the number of mutants that can be screened. In addition, the effectiveness of a screening algorithm, such as pooling, can be explored. Another approach has been to target specific residues for mutagenesis, either by comparing sequence alignments or using computational methods to identify those residues that are structurally tolerant. Each of these methods is fundamentally reliant on underlying assumptions regarding the robustness and evolvability of the enzyme.

Several theoretical models have been used to study the optimal mutation rate as a function of the size of the screening library and the ruggedness of the fitness landscape (Matsuura *et al*., 1998, Voigt *et al*., 2001a). As the number of interactions increases, the probability that a mutation is deleterious also increases. When multiple mutations are accumulated on a gene, a larger fraction of these mutations will decrease the fitness. This

effect quickly erodes the beneficial effect of any positive mutations. Therefore, to search rugged landscapes, a smaller mutation rate is optimal.

We used a computational model to demonstrate that the directed evolution algorithm preferentially discovers beneficial mutations at structurally tolerant residues (Voigt *et al*., 2001b). For a given structure, the energetic effect of each amino acid was calculated using mean-field theory and condensed into a residue entropy (Equation 7-2). Seventeen out of the twenty-two mutations found by directed evolution to improve the activity of subtilisin E and T4 lysozyme were found to occur at structurally robust residues (Figure 3-5). Targeting those residues that are structurally robust increases the fraction of the library that is folded and stable. This should also increase the probability of discovering functional improvements.

There is evidence that the immune system targets the generation of diversity to structurally tolerant residues during somatic mutagenesis. There are residue hotspots where mutations are concentrated by various cellular mechanisms during the affinity maturation process (Berek and Milstein, 1987; Sharon *et a*l., 1989; Betz *et al*., 1993; Neuberger and Milstein, 1995; Cowell *et al*., 1999). Through structural studies of germline and affinity-matured antibodies, it has been observed that somatic mutations generally preserve the structure of the binding site and antibody-antigen binding interactions (Spinelli and Alzari, 1994, Orencia, 2000). The somatic mutations could be targeted towards structurally robust residues to accelerate the discovery of higher affinity mutant antibodies. Antibodies are frequently the target of directed evolution, to improve antigen binding as well as to improve the activity of catalytic antibodies (Schultz and Lerner, 1995; Low *et al*., 1996; Xu *et al*., 1999; Boder *et al*., 2000). Targeting those

residues that have been identified as hotspots or those residues that are calculated to be structurally tolerant may improve the diversity of an *in vitro* library (Chapter 3).

Additivity is essential in the success of pooling algorithms and recombination strategies. A pooling algorithm involves the screening of multiple mutations simultaneously and then recombining the best mutants from each pool. If all of the mutations are additive, then a pooling strategy drastically reduces the screening requirements to discover the optimal combination of mutations. As the number of non-additive mutations increases, then pooling strategies become less reliable (Kauffman and Macready, 1995). Similarly, the success of recombining several mutations onto a single offspring is dependent on the strength of interactions between the mutants (Moore *et al*., 1997). If the mutations do not interact, then simply combining all of the mutations onto a single offspring is optimal. A theoretical method to identify the optimal strategy for combining the mutations has been proposed for the case when the mutants are interacting and the number of mutant that can be screened is limited (Aita and Husimi, 2000).

Consensus design has been proposed as a method to improve the thermostability of enzymes (Lehman *et al*., 2000; Jermutus, *et al*., 2001). A sequence alignment of naturally divergent sequences is used to create a consensus sequence that contains the most common amino acid at each location. This method has been used successfully to improve the thermostability (and the mutational robustness) of several enzymes (Lehman *et al*., 2000). It is unclear why the consensus sequence improves thermostability, rather than just accumulating neutral mutations. One possibility is that if natural evolution behaves like a random walk, then it is expected that the time spent in an amino acid state is proportional to the energy of that state and more stable amino acids will reside longer.

It is possible that the consensus amino acids reflect large residence times, and therefore low energies.

The success of each of these optimization strategies depends on the robustness of the enzyme. These strategies can be improved through the development of algorithms that can predict the effect of mutations on the structure (Dahiyat and Mayo, 1997; Voigt *et al*., 2000b). Those mutations that are additive are more likely to be combined successfully by pooling, consensus and recombination strategies. Further, the ability to predict the overall robustness of a system, either computationally or through the analysis of an experimentally generated knockout graph will be useful. Besides calculating the additivity of some properties, there are currently no computational methods to can predict the evolvability of specific residues. Understanding how to identify residues that contribute to various properties will lead to powerful design tools.

3. Robustness to Recombination

Recombination is a powerful tool in directed evolution as it can combine traits from multiple parents onto a single offspring (Stemmer, 1994; Crameri *et al*., 1998). Recombination plays a key role in the natural evolution of proteins, notably in the generation of diverse libraries of antibodies, synthases, and proteases (Gō, 1985). These proteins have well-defined domain boundaries and recombination shuffles domains into different configurations. The bead-like or loop topologies of these structures make them robust to recombination events (Campell and Barton, 1991). When there is no obvious domain topology, mechanisms, such as introns, can focus crossovers towards specific regions of the protein structure. In terms of *in vitro* recombination, the ability to focus the

diversity towards regions that are robust with regards to recombination will improve the quality of the library and reduce the number of hybrids that need to be screened. In this section, we first describe the observed correlation between intron locations and protein structures and then demonstrate how exon shuffling can achieve functional diversity. Finally, an algorithm based on identifying compact structural units will be used to demonstrate that successful recombination events occur in regions separating structural modules.

3.1. Evolution of Intron Locations

Many eukaryotic genes are composed of pieces of coding DNA (exons), separated in the genome by regions of non-coding DNA (introns). After transcription, introns are removed from the mRNA through a splicing mechanism. Of the many proposed functions of introns, one is that they facilitate the swapping of exons (Gilbert, 1978; Blake, 1978; Gō, 1985). When two genes are recombined, the crossovers in the mature gene will be biased towards the interface between exons. Longer introns will increase the crossover probability at that location under the assumption that crossovers can occur at each nucleotide with equal probability. If exons correspond to structural or functional subunits of protein structure, then the reconstructed gene would have a higher probability of being stable and functional. Indeed, this correlation has been demonstrated for a large number of genes (Gō, 1981; Gō, 1983; Gō, 1985; de Souza *et al*., 1996; Panchenko *et al*., 1996).

There are several possible routes by which introns could have emerged in eukaryotic genes (Gō, 1985; Gilbert *et al*., 1997; de Souza *et al*., 1998). The “introns-early” theory states that exons correspond to structural motifs that were discovered early

in evolutionary history. These exons were pieced together by recombination and gene duplication to build the genes that are now observed. This view asserts that prokaryotes lost their introns due to the strong selection on genome size. In contrast, the “introns-late” theory states that introns were inserted in genes late in evolutionary history, thus explaining their existence in eukaryotes. The early versus late debate is ongoing and it is likely that some introns emerged early and were lost and others emerged late.

If an intron emerged due to the early mechanism, then it is clearly going to correspond to a structural subunit. Arriving late, it could appear anywhere throughout the structure equally, without any structural preference. This idea has led to the argument that the observed correlation between introns and structural units is evidence of an early mechanism (Gilbert *et al*., 1997). However, if introns were to appear at random locations in a population of genes, then selection could drive the introns towards regions separating structural modules, if the existence of an intron increases an organism’s fitness by promoting successful recombination events on a reasonably fast time scale. In other words, selection drives the creation of a robust gene structure.

Theory that has been developed to optimize genetic algorithms provides insight into the relationship between recombination and protein structure. In this literature, the concept of a schema, or a cluster of interacting bits, is useful in predicting the success, or failure, of recombination (Holland, 1975). When crossovers frequently divide a schema, then these interactions are disrupted and the offspring are more likely to have inferior fitnesses. When schema disruption is not controlled, genetic algorithms will often fail to converge on an optimal solution (Mitchell, 1994; Mitchell, 1998). In a particularly interesting study, the success of a genetic algorithm was improved by recording where

past crossovers resulted in fit offspring (Schaffer and Morishima, 1987). This information was used to bias crossovers in future generations. In this way, selection automatically biased the recombination markers towards regions that separated schemas. Extending these results to biology, this simulation demonstrates the advantage of shifting introns towards the regions that separate structural schemas.

3.2. Exons as Functional Switches

Exon swapping can occur on evolutionary timescales or on the timescale of gene splicing in the cell (Gō, 1985). The ability to swap exons without disrupting the structure improves the evolvability of the gene by promoting functional switches between different molecular properties (Gilbert, 1978; Blake, 1979). These switches have been found to alter the substrate specificity, the tissue distribution, and the association properties of the translated proteins. It has been suggested that performing exon swapping *in vitro* will produce functionally diverse libraries (Fisch *et al*., 1996; Kolkman and Stemmer, 2001).

There have been several examples of achieving functional diversity through the *in vitro* swapping of exons that correspond to structural modules of enzymes. Gō and co-workers altered the coenzyme specificity of isocitrate dehydrogenase by calculating the structural module corresponding to the NADP-binding site (Yaoi *et al*., 1996). When this module was swapped with a NAD-binding site, the reaction was shown to proceed with the new coenzyme. In another experiment, a module of the β-subunit of hemoglobin was swapped with the corresponding module of the α-subunit (Wakasugi *et al*., 1994; Inaba *et al*., 1997). This hybrid protein folded into the correct tertiary structure, but the association of different subunits was altered, suggesting that the function of the fourth

module is to regulate subunit association. This substitution did not affect other properties of hemoglobin, including oxygen binding. In a particularly dramatic experiment, the catalytic activities of α-lactalbumin and lysozyme were swapped by shuffling the exon corresponding to the amino acids that surround their active sites (Kumagai *et al*., 1992). The success of this experiment hinged on the observation that the two enzymes share the same structure and distribution of exons. Swapping exons can also alter specificity. For example, the swapping of alternate exons in a human cytochrome P450 changed its substrate specificity and tissue distribution (Christmas *et al*., 2001). This implies that the gene structure of P450 promotes the swapping of functional modules such that this enzyme can participate in different biological functions. In the dehydrogenase, hemoglobin and lysozyme experiments, subportions of the structural module were swapped as a control. In each case, swapping a portion rather than the whole module resulted in an unstable or non-functional enzyme.

The immune system effectively uses exon shuffling to create antibody variants that can bind a broad range of antigens. Mimicking *in vivo* antibody selection, Borrebaeck and co-workers used recombination techniques to shuffle the naturally occurring human exons that encode the CDR regions to generate a large binding repertoire (Soderlind et *al*., 2000). The library containing ~10^9 antibodies was screened against a wide array of hapten and protein targets and antibodies with nanomolar binding affinities were reliably found. This stunning work represents the ability to create a full antibody repertoire in the test tube. When combined with directed-evolution-like somatic mutagenesis, a nearly complete artificial immune system will be created.

3.3. In Vitro Recombination Preserves Structural Schema

The success of *in vitro* recombination is based on the assumption that the parents share similar structures. For a hybrid protein to demonstrate new or improved properties, a prerequisite is that it folds into a well-defined (and presumably similar) structure. Therefore, crossovers are more likely to be successful when they occur in regions that lie between schemas (Voigt *et al*., 2002). In this context, schemas are defined by the pattern of stabilizing interactions between amino acids and recombination is most successful when the crossovers break the fewest interactions. The hybrids with the minimum schema disruption are the most likely to retain the structure of the parents.

Using a computational algorithm that predicts the location of schemas, data were analyzed from five independent directed evolution experiments where several parents were shuffled to create random libraries of recombinant offspring (Voigt *et al*., 2002). Crossovers in the offspring that survived selection were strongly biased towards regions that minimize the schema disruption. To further demonstrate the requirement that schema be preserved, two β-lactamases were recombined that have similar structures, but share little sequence identity. The three-dimensional structure was divided into schemas and the interaction strengths between the different schemas were calculated. Experimentally, hybrid proteins were constructed where the schemas were exchanged between structures and each hybrid was tested for activity. A sharp transition was found in the activity as the disruption of the hybrid increased (Figure 5-9). Recombination events that cause disruption above this threshold resulted in non-functional hybrids. These experiments demonstrate in real-time how selection for folded, function offspring can bias crossover-focusing mechanisms towards regions separating structural schemas. In addition, this

algorithm will improve directed evolution as it enables the design of libraries with an enriched fraction of properly folded hybrid proteins.

4. Evolution of Genetic Circuits and Metabolic Pathways

Metabolic pathways and genetic circuits have recently become targets for *in vitro* evolution (McAdams and Arkin, 2000; Schmidt-Dannert *et al*., 2000). The robustness of a network describes the resilience of a behavior to perturbations in parameters. In the case of a metabolic pathway, the behavior may be the chemicals generated and the rate at which they are produced. For a genetic circuit, the behavior is the integrity of the computation, for example, the ability to behave like an oscillator, toggle switch, logic gate, or memory (Bhalla and Iyengar, 1999; Yuh *et al*., 2000; Gardner *et al*., 2000; Elowitz and Leibler, 2000). A network is evolvable if it can change behaviors through the perturbation of its internal parameters.

There are several means by which a diverse library of networks can be created. One method involves the randomization of the component genes through mutagenesis or recombination. Mutations can change the behavior of a network by altering the kinetic constants for an activity, the substrate specificity, and the products produced. If the mutated DNA encodes a repressor or activator protein or a related DNA binding site, then mutations will vary the strength of repression or activation (Becskei and Serrano, 2000). Further, mutations can stabilize or destabilize a protein, which affects the dynamics of the network by changing the protein's residence time. The library of offspring is more likely to contain the desired properties if the diversity is applied to those parameters that are

evolvable or those parameters that are robust regarding properties required for the desired behavior, but not being explicitly optimized.

A combinatorial library of networks can also be created by randomly combining modules of pre-constructed combinations of transcription units. For example, the combination of different genes with different repressor sites can produce a library of many possible dynamic circuits. In this way, different interaction topologies can be created and tested for various behaviors. For this combinatorial approach to be successful, it is necessary that the modules are robust regarding the form of the inputs and the network and cellular environments in which they are inserted (von Dassow *et al.*, 2000; Savageau, 2001).

4.1. Robust Network Topologies

The topology of a network describes the architecture of interactions between the network components. For a metabolic pathway, an interaction may represent the enzymatic conversion of a substrate into a product or the effect of a species on the control of a reaction. An interaction may also describe the effect of a repressor or activator on the expression of another gene. Network topologies can be visualized as a set of nodes and edges, where a node represents a component and an edge represents an interaction (Figure 7-1). The topologies of large networks can be very complex, and it is difficult to predict those features that are robust from those that are fragile (Figure 7-4). Some topological motifs that confer robustness have been identified, including feedback loops, a modular architecture, and a scale-free distribution of nodes and edges.

Feedback control buffers the variables of the system towards external perturbations; in other words, it improves the stability of the system. Beyond improving stability, feedback control also improves robustness with respect to internal parameters (Savageau, 1972). Feedback control can occur with various topologies. The simplest, autoregulation, where the immediate products inhibit the reaction, has been shown to improve robustness (Savageau, 1974; Becskei and Serrano, 1998). Larger feedback loops, when downstream products control the first steps of metabolism, proved to be more robust (Savageau, 1972). Alternative topologies that contained multiple, small feedback loops in sequence or nested, proved to be less robust than the systems with larger feedback loops.

Barkai and Leibler proposed that robustness is manifested in signal transduction networks through the topology of feedback control (1997). They constructed a model of bacterial chemotaxis and found a set of parameters that reproduced the desired network behavior, in this case, to turn the flagella on and off in such a way so the bacteria swims up an attractant gradient (adaptation). This network turned out to be remarkably robust as nearly 80% of networks generated by randomly varying all of the parameters 2-fold from the starting point were found to demonstrate adaptation. When varied individually, each parameter could be varied by several orders of magnitude. This robustness was later verified experimentally (Alon *et al*., 1999) and proved to result from an integral feedback control loop (Yi *et al*., 2000). This result supports the concept that robustness is inherent to the topology of the network and does not emerge through a specific combination of parameter values.

The robustness of a network topology towards the removal of nodes or edges, as the result of knockout or specificity-altering mutations, has been studied using a number of theoretical models (Albert *et al*., 2000; Cohen *et al*., 2000; Jeong *et al*., 2000; Cohen *et al*., 2001). A general result from these studies is that a scale-free topology of interactions tends to be robust. In a scale-free network, the distribution of nodes and edges follows a power law, $P(k) \sim k^{-\alpha}$, where $P(k)$ is the probability of a node having connectivity k, and α is a power-log exponent. Considering robustness, there is an optimal value of $\alpha \approx 2.5$, which is shared by the structure of the internet (Cohen *et al*., 2000) and metabolic networks (Wagner and Fell, 2001). In a comparison across metabolic networks of 43 species, it was found that the most highly connected nodes were the most conserved, whereas the least connected varied considerably (Jeong *et al*., 2000). In other words, the least connected nodes are highly tolerant to changes, similar to what was found for protein structures (Voigt *et al*., 2001b).

4.2. Evolvability of Networks

An evolvable network has the ability to sample many behaviors by varying the internal parameters. To be robust, a network has to retain some behavior under parameter variation. In contrast, evolvability requires that other behaviors are attainable without traversing regions of parameter space that are devoid of behavior. Very little is understood as to how the topology of a network affects evolvability. There is some evidence that the evolvability of a network can be improved through modularity, switch points, and broad substrate specificities. Understanding these properties will facilitate the creation of *in vitro* strategies to create libraries of networks with diverse properties.

When shuffling portions of gene networks, it is essential that the subcomponents are modular. A portion of a genetic network forms a module if it can be substituted in other environments or other networks and performs the same qualitative behavior. If a network is modular, evolution can reuse motifs by rewiring the network inputs and outputs (Hartwell *et al*., 1999). In terms of the distribution of nodes and edges, a modular topology can be divided into subsets of nodes, where the nodes are highly interacting within the subset and not interacting between subsets. In addition, the network should generate the same behavior for a variety of input stimuli (von Dassow *et al*., 2000). This reduces the demands on the form of the inputs, making it easier to combine with other subnetworks. Using kinetic models, Bhalla and Iyengar found that a wide range of complex dynamic behaviors was attainable by coupling multiple, independent signaling pathways (1999). The behaviors that could be obtained included memory, timers, switching regulatory wiring, and time and concentration thresholds for response.

Stochastic mechanisms can alter the behavior of a network (McAdams and Arkin, 1997; McAdams and Arkin, 1999). The random component can be introduced by low concentrations of reacting species, slow reaction rates, or limited availability of catalytic centers. Biological systems exploit this randomness as a switch that can determine the behavior of an individual cell or create diversity in a population in cells. For a network to be able to utilize stochastic effects, it is necessary to stabilize the effect of fluctuations on the remainder of the network through redundancy and feedback loops. In other words, maintaining the overall robustness of the network promotes evolvability through the variation of other parameters. Stochastic switches that select between two alternate pathways can occur (Arkin *et al*., 1998; McAdams and Arkin, 1999). Identifying a switch

point and then targeting it using directed evolution could be used to create a diverse library of network behaviors.

Gene duplication creates functional redundancy. The robustness conferred by a duplication event is likely to deteriorate rapidly as there is no disadvantage for the function of one of the copied genes to be destroyed (Nowak *et al*., 1997; Wagner, 2000; Wagner, 2001). However, the evolutionary stability of the duplication can be ensured initially by increasing regulatory reliability or by changing the function of the duplicated gene. This creates the opportunity for the duplicated gene to add or optimize functions, possibly in unrelated pathways (McAdams and Arkin, 1999). An ancestral gene with broad substrate specificity can potentially participate in more pathways and is therefore more evolvable. This mechanism has been duplicated *in vitro* by demonstrating that a metabolic enzyme with broad substrate specificity could be evolved to specifically participate in two different metabolic pathways (Jürgens *et al*., 2000).

Secondary metabolites are produced by various organisms and are typically not essential for survival. These chemicals may confer properties, such as the color or fragrance, or they may have therapeutic properties, such as antibiotic or tumor-suppressing activities, or serve in biological warfare or defense. While many of the chemicals produced by an organism may not confer a selective advantage, the existence of a secondary metabolism is advantageous because it gives the organism the potential to discover a few potent chemicals (Firn and Jones, 2000). To optimize the discovery process, secondary metabolisms appear to have evolved to maximize the diversity of chemicals that are attainable when the components of the pathway are perturbed. The evolvability of secondary metabolic enzymes is enhanced by low substrate specificities,

thus increasing the number of potential downstream products. These specificities can be tuned and fixed by evolution as beneficial metabolites are discovered. The components of secondary metabolism have been observed to be robust to environmental changes. This facilitates the transfer of a network from one organism to another.

The robustness and evolvability of secondary pathways make them particularly well-suited targets for *in vitro* evolution to tune the production of desired chemicals. New biosynthetic pathways can be created by combining genes from different sources and subjecting these genes to mutagenesis and recombination (Figure 7-5). Selection is then applied for the production of specific compounds or for the generation of diverse chemicals. Directed evolution has been applied to the carotenoid biosynthetic pathway to alter the diversity of the chemicals produced (Schmidt-Dannert *et al*., 2000). Four genes that encode different component enzymes were combined. The basic C_{40} carotenoid building block is constructed by two synthases. The inclusion of a phytoene desaturase (*crtI*) and lycopene cyclase (*crtY*) allowed the C_{40} backbone to be modified to produce distinct carotenoids. These two modifying enzymes control the number of double bonds and the formation of cyclic rings, respectively. Libraries of *crtI* and *crtY* genes were created through the recombination of homologous genes. The production of different carotenoids could be visualized by the change in color caused by these modifications. Mutants were found that produced carotenoids with different degrees of desaturation (fewer double bonds results in a yellowish color, more double bonds results in a pinkish color) and different degrees of cyclization (resulting in a orange-red to purple-red color). One clone produced torulene, a carotenoid that had not been previously observed in organisms that contain these parental biosynthetic genes. Organisms that produce

torulene in nature do so by a different metabolic route. These experiments demonstrate that applying directed evolution to a secondary metabolic pathway can alter the range of chemicals produced by the network and extend the product diversity beyond what is produced by the parents.

5. Conclusions

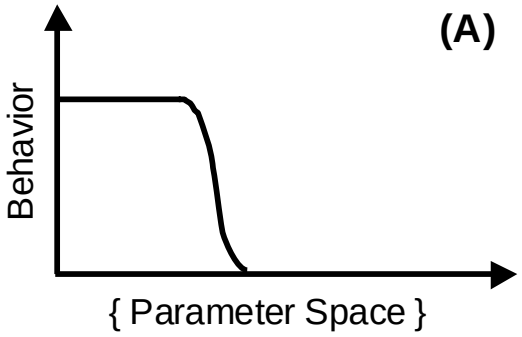
Understanding the robustness and evolvability of biological systems is essential for constructing algorithms that guide evolutionary design strategies. To quantify the potential for evolutionary improvement, stability theory is useless as it is concerned with the resilience of a single state of a system when variables are perturbed. In contrast, evolution samples many states by perturbing the system's internal parameters, such as kinetic constants, interaction topologies, and component stabilities. A robust system has the ability to sample many of these states while preserving the fundamental behaviors of the system. During this process of drifting through parameter space, new behaviors can be sampled. Some systems have a higher capacity for such change and are therefore evolvable.

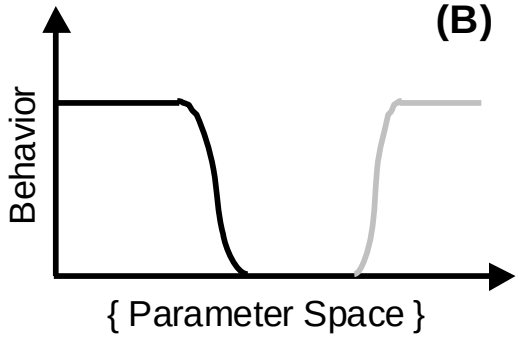
The robust and evolvable properties of three hierarchies in biology have been reviewed: the mutation of individual genes, the recombination of clusters of amino acids from different genes, and the evolution of biological networks. While these systems have some fundamental differences, there are many common motifs that confer robustness and evolvability. In all of these systems, there are similar motifs that improve robustness, such as the distribution of interactions, modularity, the separation of parameters (additivity), and redundancy. These themes are common to biological systems not

reviewed here, such as RNA structures, viruses, genomes, and whole ecologies. An understanding of the basis for robust and evolvable systems in biology will facilitate the design of a new generation of techniques in directed evolution.

Figure 7-1:

The behavior of a biological system is plotted as the function of some internal parameter. When the behavior is robust, it is insensitive to large variation in parameter space **(A)**. It is possible that variation in a parameter can sample new behaviors **(B)**. If this behavior is attainable without having to go through dead parameter space, the behavior is evolvable **(C)**.


(A)
Behavior
{ Parameter Space }


(B)
Behavior
{ Parameter Space }

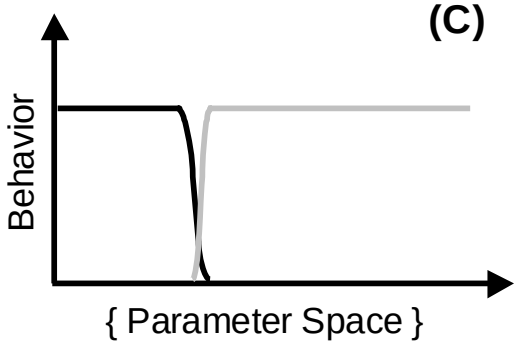
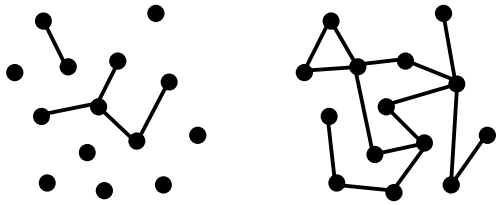
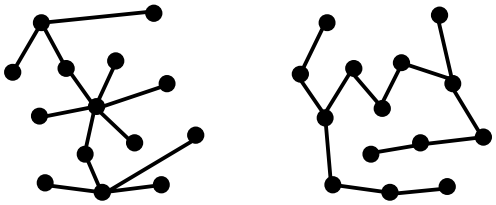

(C)
Behavior
{ Parameter Space }

Figure 7-2:

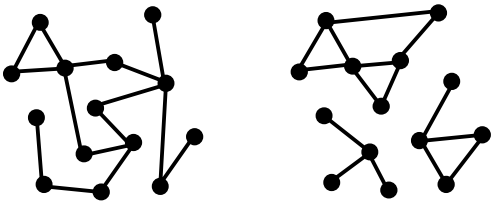
Some examples of robust and fragile architectures for a system of nodes and edges. If the system is a protein structure, the nodes represent residues and the edges are amino acid interactions. If the system is a protein network, then the nodes can be component proteins and the edges represent protein-protein interactions. More robust structures have fewer interactions (**A** – left), a scale-free-like distribution of interactions (**B**-left), and a modular structure (**C** – left). Modularity does not necessarily have to be on the level of interactions. The most robust two-dimensional protein structure, as determined by Wingreen and co-workers, is shown on the left, where gray lines mark the progression of the carbon backbone (Li *et al*., 1996). This structure represents the repetition of a smaller, robust motif that is not ascertainable from the interaction topology (black lines, right).



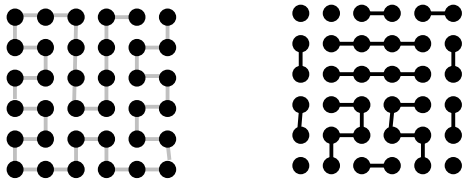
(A)



(B)



(C)



(D)

Figure 7-3:

The knockout graphs of HIV RT (□), T4 Polymerase (Δ), and an antibody (○) are shown (Suzuki *et al*., 1996; Daugherty *et al*., 2000). A steeper initial slope indicates the behavior is more sensitive to mutations. Note that the same protein could have different slopes if different behaviors (or different stringencies) were measured. At high mutation rates, there is a transition in the slope, implying the emergence of compensating mutations.

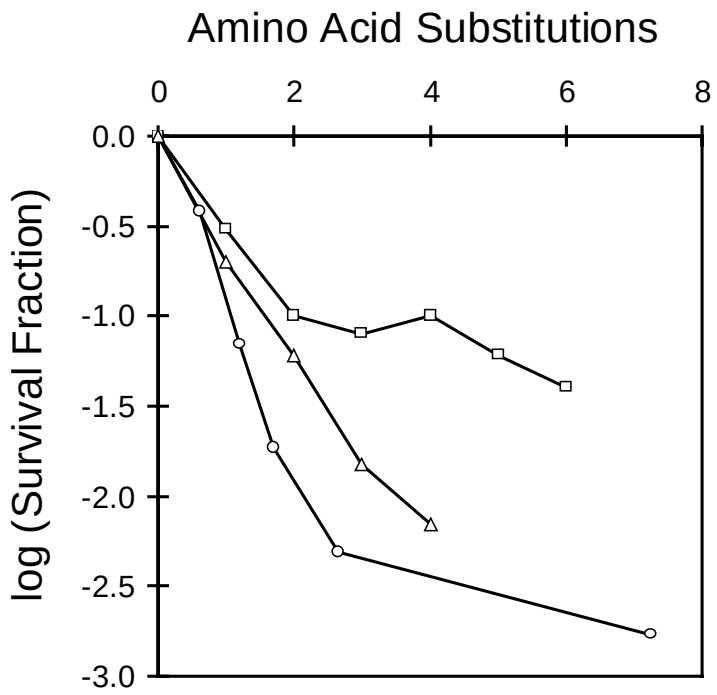

Amino Acid Substitutions
0
2
4
6
8
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
log (Survival Fraction)

Figure 7-4:

The segment polarization network of *Drosophila*, as modeled by Odell and co-workers (von Dassow *et al*., 2000). **(A)** The topology of the network showing the interactions between proteins, including intercellular interactions. The behavior of this network is to produce a specific pattern of expression for a group of cells **(B)**. The wild-type behavior was found to be remarkably robust. Often, the parameters controlling the network could be varied by several orders of magnitude **(C)**. In addition to the robustness, it was found that varying some parameters could change the behavior of the network (different expression patterns), indicating that the network is evolvable.

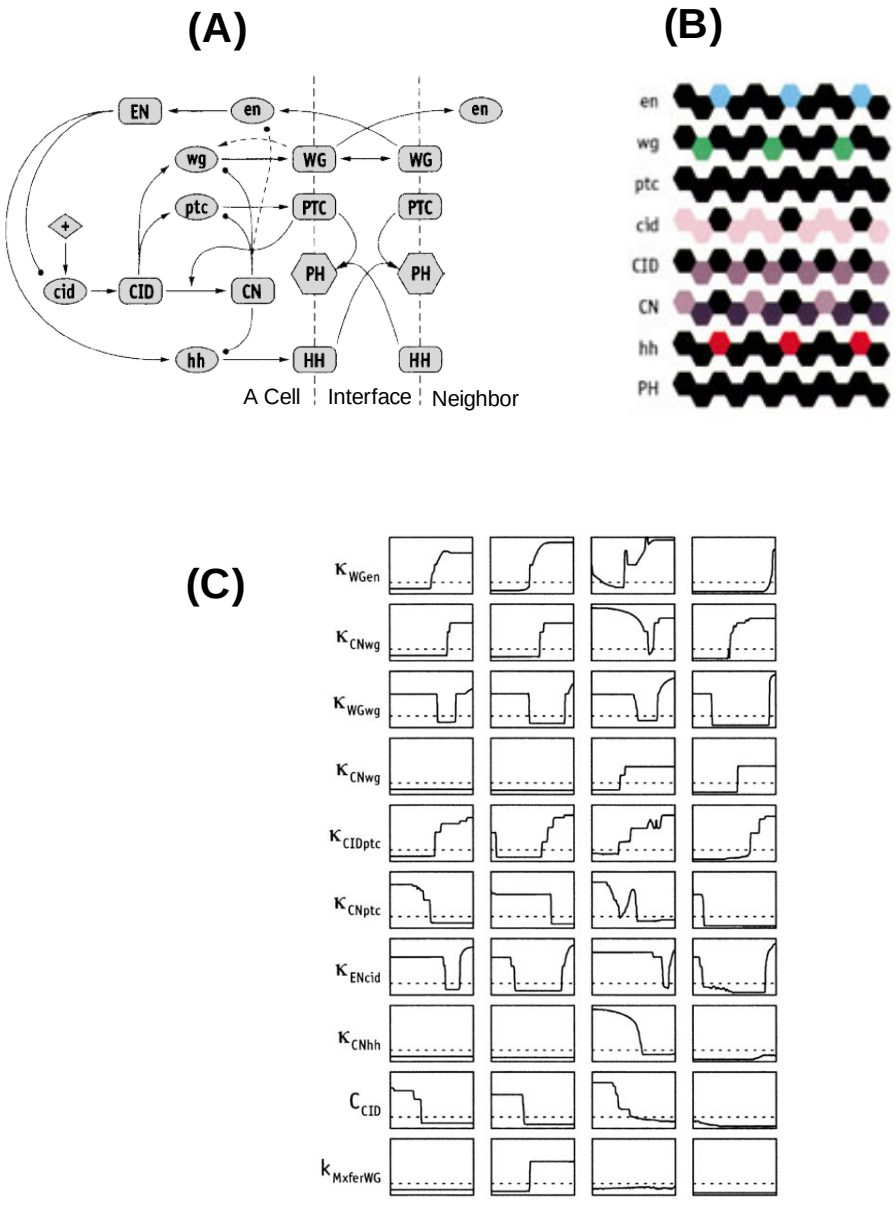
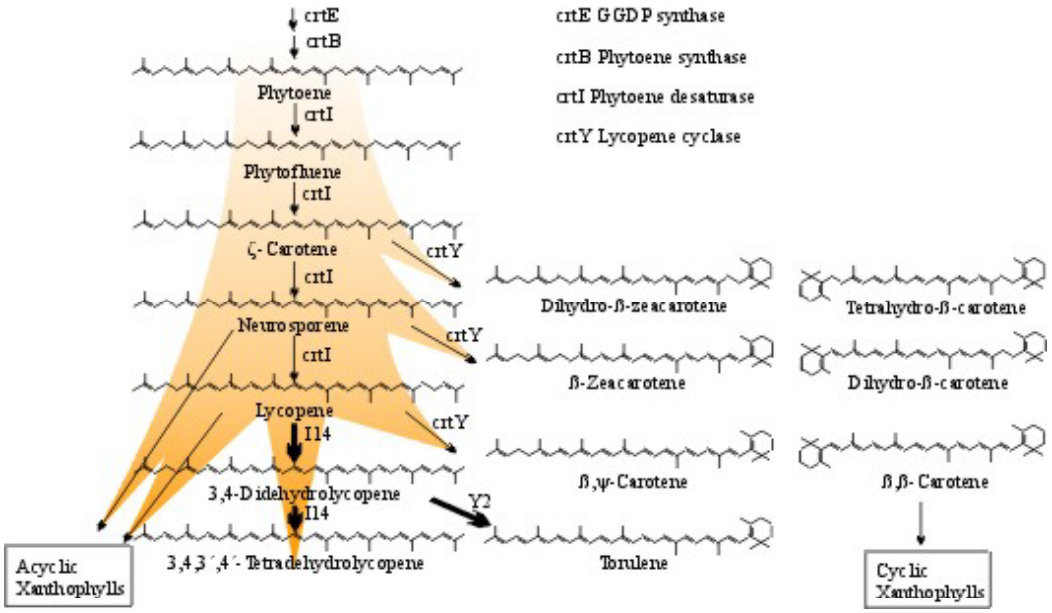
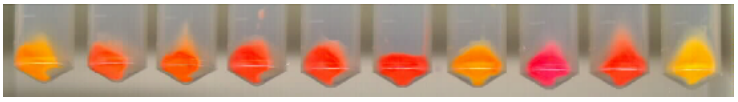

(A)
EN
en
en
wg
WG
WG
ptc
PTC
PTC
+
cid
CID
CN
PH
PH
hh
HH
HH
A Cell
Interface
Neighbor
(B)
en
wg
ptc
cid
CID
CN
hh
PH
(C)
κWGen
κCNwg
κWGwg
κCNwg
κCIDptc
κCNptc
κENcid
κCNhh
CCID
kMxferWG

Figure 7-5:

The range of carotenoids **(A)** and their colors **(B)** that can be created through directed evolution (Schmidt-Dannert *et al*., 2000). The *crtE* and *crtB* enzymes create the initial C_{40} backbone (top, left). This can then be transformed into different carotenoids by desaturating bonds with *crtI* and cyclizing the ends with *crtY*. By evolving the properties or *crtI* and *crtY*, the production of a variety of carotenoids was possible.


crtE
crtB
Phytoene
crtI
Phytofluene
crtI
ζ- Carotene
crtI
Neurosporene
crtI
Lycopene
I14
3,4-Didehydrolycopene
I14
3,4,3´,4´- Tetradehydrolycopene
crtY
crtY
crtY
Y2
Acyclic Xanthophylls
Dihydro-ß-zeacarotene
ß-Zeacarotene
ß,ψ-Carotene
Torulene
Tetrahydro-ß-carotene
Dihydro-ß-carotene
ß,ß- Carotene
Cyclic Xanthophylls
crtE GGDP synthase
crtB Phytoene synthase
crtI Phytoene desaturase
crtY Lycopene cyclase

(A)



(B)

Appendix A

Higher-Order Moments of the Mutant Distribution

The n^{th}-moment of the mutant fitness distribution is calculated from

$$\mu_n = \left\langle \left(w - < w > \right)^n \right\rangle \quad , \tag{A-1}$$

giving the familiar equation, $\mu_2 = <w^2> - <w>^2$, for the second moment. The more familiar standard deviation is simply $\sigma = \sqrt{\mu_2}$. Note that there is a difference in the naming convention between Chapter 2 and this appendix. Here, amino acid *a* at residue is indicated by a_i, not i_a. First, we square Equation (2-24),

$$\frac{\langle w \rangle^2}{C^2} = \left\{ \sum_i^N \sum_{a_i, b_i} \Delta\gamma_i P_i + \frac{b}{2} \sum_i^N \sum_{j \neq i}^N \sum_{a_i, b_i} \sum_{a_j, b_j} \Delta\gamma_{ij} P_i P_j \right\}^2 \tag{A-2}$$

and expand the sums,

$$\begin{aligned}
\frac{\langle w \rangle^2}{C^2} &= \sum_i^N \sum_a \sum_b \Delta\gamma_i{}^2 P_i^2 + b \sum_i^N \sum_j^N \sum_{a_i, a_j} \sum_{b_i, b_j} \Delta\gamma_i \Delta\gamma_{ij} P_i^2 P_j + \frac{b^2}{4} \sum_i^N \sum_j^N \sum_{a_i, a_j} \sum_{b_i, b_j} \Delta\gamma_{ij}{}^2 P_i^2 P_j{}^2 \\
&+ \sum_i^N \sum_{i' \neq i}^N \sum_{a_i, a_{i'}} \sum_{b_i, b_{i'}} \Delta\gamma_i \Delta\gamma_{i'} P_i P_{i'} + b \sum_i^N \sum_{i' \neq i}^N \sum_j^N \sum_{a_i, a_{i'}, a_j} \sum_{b_i, b_{i'}, b_j} \Delta\gamma_i \Delta\gamma_{i'j} P_i P_{i'} P_j \\
&+ \frac{b^2}{4} \sum_i^N \sum_{i' \neq i}^N \sum_j^N \sum_{a_i, a_{i'}, a_j} \sum_{b_i, b_{i'}, b_j} \Delta\gamma_{ij} \Delta\gamma_{i'j'} P_i P_{i'} P_j{}^2 + \frac{b^2}{4} \sum_i^N \sum_j^N \sum_{j' \neq j}^N \sum_{a_i, a_j, a_{j'}} \sum_{b_i, b_j, b_{j'}} \Delta\gamma_{ij} \Delta\gamma_{i'j'} P_i^2 P_j P_{j'} \\
&+ \frac{b^2}{4} \sum_i^N \sum_{i' \neq i}^N \sum_j^N \sum_{j' \neq j}^N \sum_{a_i, a_{i'}, a_j, a_{j'}} \sum_{b_i, b_{i'}, b_j, b_{j'}} \Delta\gamma_{ij} \Delta\gamma_{i'j'} P_i P_{i'} P_j P_{j'}
\end{aligned} \tag{A-3}$$

Then, *w* is squared such that

$$w^2 = \sum_i^N \sum_{i'}^N \Delta\gamma_i \Delta\gamma_{i'} + b \sum_i^N \sum_{i'}^N \sum_j^N \Delta\gamma_i \Delta\gamma_{i'j} + \frac{b^2}{4} \sum_i^N \sum_{i'}^N \sum_j^N \sum_{j'}^N \Delta\gamma_{ij} \Delta\gamma_{i'j'} \ , \tag{A-4}$$

and expanded to

$$w^2 = \sum_i^N \Delta\gamma_i{}^2 + b\sum_i^N\sum_j^N \Delta\gamma_i\Delta\gamma_{ij} + \frac{b^2}{4}\sum_i^N\sum_j^N \Delta\gamma_{ij}{}^2 + \sum_{i\neq i'}^N\sum^N \Delta\gamma_i\Delta\gamma_{i'} + b\sum_i^N\sum_{i'\neq i}^N\sum_j^N \Delta\gamma_i\Delta\gamma_{i'j}$$
$$+ \frac{b^2}{4}\sum_i^N\sum_{i'\neq i}^N\sum_j^N \Delta\gamma_{ij}\Delta\gamma_{i'j'} + \frac{b^2}{4}\sum_i^N\sum_j^N\sum_{j'\neq j}^N \Delta\gamma_{ij}\Delta\gamma_{i'j'} + \frac{b^2}{4}\sum_i^N\sum_{i'\neq i}^N\sum_j^N\sum_{j'\neq j}^N \Delta\gamma_{ij}\Delta\gamma_{i'j'} \qquad \text{(A-5)}$$

and averaged over all amino acid sequences

$$\frac{\langle w^2\rangle}{C} = \sum_i^N\sum_a\sum_b \Delta\gamma_i{}^2 P_i + b\sum_i^N\sum_j^N\sum_{a_i,a_j}\sum_{b_i,b_j} \Delta\gamma_i\Delta\gamma_{ij}P_iP_j + \frac{b^2}{4}\sum_i^N\sum_j^N\sum_{a_i,a_j}\sum_{b_i,b_j} \Delta\gamma_{ij}{}^2 P_iP_j$$
$$+ \sum_i^N\sum_{i'\neq i}^N\sum_{a_i,a_{i'}}\sum_{b_i,b_{i'}} \Delta\gamma_i\Delta\gamma_{i'}P_iP_{i'} + b\sum_i^N\sum_{i'\neq i}^N\sum_j^N\sum_{a_i,a_{i'},a_j}\sum_{b_i,b_{i'},b_j} \Delta\gamma_i\Delta\gamma_{i'j}P_iP_{i'}P_j$$
$$+ \frac{b^2}{4}\sum_i^N\sum_{i'\neq i}^N\sum_j^N\sum_{a_i,a_{i'},a_j}\sum_{b_i,b_{i'},b_j} \Delta\gamma_{ij}\Delta\gamma_{i'j'}P_iP_{i'}P_j + \frac{b^2}{4}\sum_i^N\sum_j^N\sum_{j'\neq j}^N\sum_{a_i,a_j,a_{j'}}\sum_{b_i,b_j,b_{j'}} \Delta\gamma_{ij}\Delta\gamma_{i'j'}P_iP_jP_{j'} \qquad \text{(A-6)}$$
$$+ \frac{b^2}{4}\sum_i^N\sum_{i'\neq i}^N\sum_j^N\sum_{j'\neq j}^N\sum_{a_i,a_{i'},a_j,a_{j'}}\sum_{b_i,b_{i'},b_j,b_{j'}} \Delta\gamma_{ij}\Delta\gamma_{i'j'}P_iP_{i'}P_jP_{j'}$$

Finally, subtracting A.3 from A.6 gives the second moment,

$$\mu_2 = \sum_i^N\sum_a\sum_b \Delta\gamma_i{}^2\frac{P_i}{C}\left(1-\frac{P_i}{C}\right) + b\sum_i^N\sum_j^N\sum_{a_i,a_j}\sum_{b_i,b_j} \Delta\gamma_i\Delta\gamma_{ij}\frac{P_iP_j}{C}\left(1-\frac{P_i}{C}\right)$$
$$+ \frac{b^2}{4}\sum_i^N\sum_{i'\neq i}^N\sum_j^N\sum_{a_i,a_{i'},a_j}\sum_{b_i,b_{i'},b_j} \Delta\gamma_{ij}\Delta\gamma_{i'j'}\frac{P_iP_{i'}P_j}{C}\left(1-\frac{P_j}{C}\right)$$
$$+ \frac{b^2}{4}\sum_i^N\sum_j^N\sum_{j'\neq j}^N\sum_{a_i,a_j,a_{j'}}\sum_{b_i,b_j,b_{j'}} \Delta\gamma_{ij}\Delta\gamma_{i'j'}\frac{P_iP_jP_{j'}}{C}\left(1-\frac{P_i}{C}\right) \qquad \text{(A-7)}$$
$$+ \frac{b^2}{4}\sum_i^N\sum_j^N\sum_{a_i,a_j}\sum_{b_i,b_j} \Delta\gamma_{ij}{}^2\frac{P_iP_j}{C}\left(1-\frac{P_iP_j}{C}\right)$$

Appendix B

Dead-end Elimination and Monte Carlo Entropy Calculations

Mean-field theory is an approximate method and is expected to worsen as the coupling in the system increases. To overcome this problem, we have developed two algorithms that calculate the entropy based on a series of minimizations performed by the dead-end elimination (DEE) algorithm or a Monte Carlo (MC) simulation.

The DEE-entropy algorithm calculates the substitution energy of all amino acids at all positions in the wild-type amino acid background (Figure B-1). First, a residue is chosen (residue *i*) and the remaining residues in the structure are held in their wild-type amino acid identity. Then, residue *i* is assigned an amino acid identity *a*. The flexibility of all the amino acid side chains are discretized into rotamers and the global minimum energy conformation is found using the DEE algorithm. This minimum energy is then assigned to that amino acid at that residue. This procedure is used to find the energy of all twenty amino acid substitutions at residue *i*. This process is repeated for all residues in the protein so the outcome of the algorithm is the energy for all single-mutant amino acid substitutions at every position.

The probability of each amino acid at each residue is calculated from the energies using a Boltzmann weighting,

$$p(i_a) = \frac{e^{-\beta E(i_a)}}{\sum_{b}^{A} e^{-\beta E(i_b)}}, \qquad \text{(B-1)}$$

where $A = 20$ is the total number of amino acids, $p(i_a)$ is the probability of amino acid a existing at residue i, and $E(i_a)$ is the energy of amino acid a at residue i. The entropy s_i is then calculated using Equation (3-2). The temperature in (B-1) is similar in interpretation

to the mean-field temperature in that it represents a threshold energy, above which sequences are unstable (Figure 3-1). When the entropies calculated by the mean-field algorithm and DEE-algorithm are compared, both algorithms agree on the assignment of the high entropy positions (Figure B-2). The disagreement increases for the low entropy residues. This is, in part, due to the mean-field assumption. As the coupling between residues increases, this assumption becomes less valid (see Chapter 6). Also, the restriction that the DEE algorithm must calculate the substitution energies based on the wild-type amino acid background leads to disagreement between the methods.

Monte Carlo methods can also be used to calculate a sequence entropy (Figure B-3). First, the rotamer conformation is minimized for a structure. From the global minimum energy conformation (GMEC), a Monte Carlo algorithm is run where rotamer substitutions are made at random and accepted with a weighted probability if the mutant sequence is higher in energy than the previous rotamer sequence. During the Monte Carlo run, the rotamer sequences that are within a cutoff energy from the GMEC sequence are recorded (Figure B-3). At the end of a defined number of Monte Carlo steps, the set of low-energy sequences is used to calculate the sequence entropy. The probabilities that each amino acid exists at a residue are calculated by counting by the number of times that amino acid appears in the low-energy list and dividing by the total number of low-energy sequences. The sequence entropies can then be calculated using Equation (3-2) (Figure B-4). While the high entropy residues remain approximately the same, many mean-field low-entropy positions are predicted by Monte Carlo to have zero entropy. This could be due to the limited sampling of the Monte Carlo algorithm.

Figure B-1:

A schematic is shown for the DEE-entropy algorithm. First, a residue is chosen and all twenty amino acids are substituted at that residue. For each substitution, the minimum energy conformation is obtained using dead-end elimination. After all substitutions are made at all residues, the resulting list of energies is used to produce the entropy profile. This is a more exact method than the mean-field algorithm, but is limited to making point substitutions in the wild-type amino acid background.

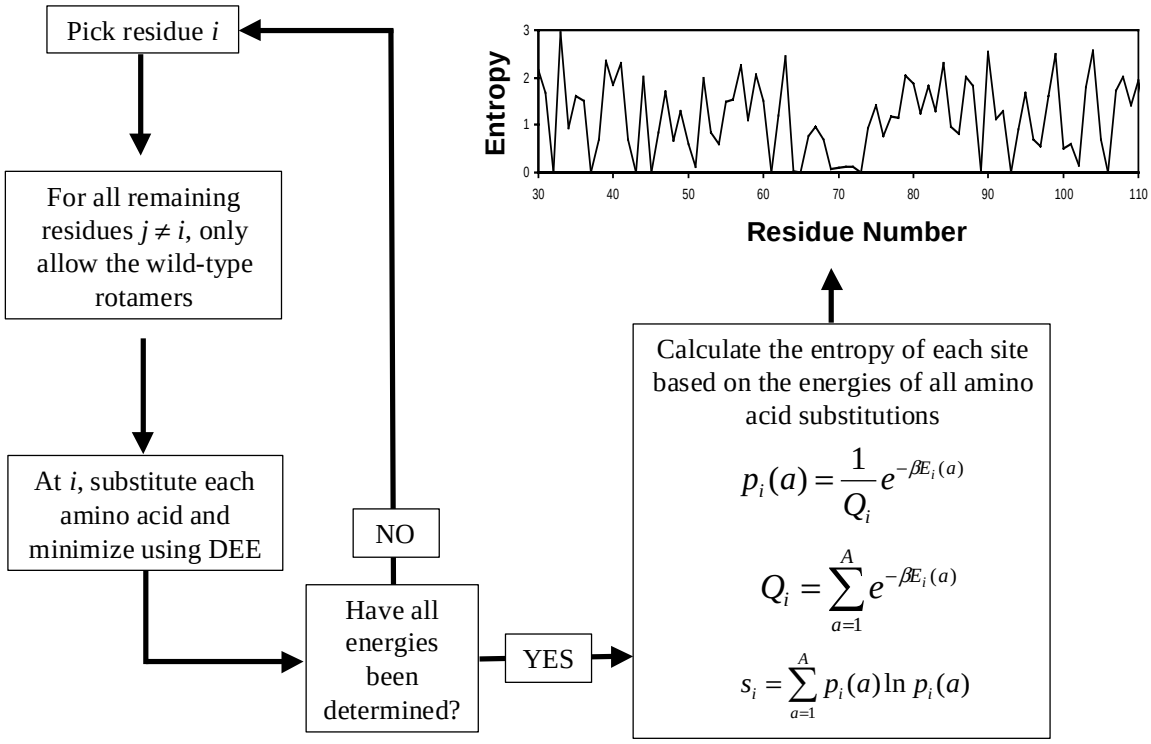

Pick residue *i*
For all remaining residues $j \neq i$, only allow the wild-type rotamers
At *i*, substitute each amino acid and minimize using DEE
Have all energies been determined?
NO
YES
Calculate the entropy of each site based on the energies of all amino acid substitutions
$p_i(a) = \frac{1}{Q_i} e^{-\beta E_i(a)}$
$Q_i = \sum_{a=1}^{A} e^{-\beta E_i(a)}$
$s_i = \sum_{a=1}^{A} p_i(a) \ln p_i(a)$
Entropy
0
1
2
3
30
40
50
60
70
80
90
100
110
Residue Number

Figure B-2:

A comparison of the entropy calculated by the mean-field algorithm and the DEE algorithm for the T4 lysozyme structure. Both algorithms identify find the same high-entropy positions, but differ in their rank ordering of low-entropy positions. This is a demonstration of the fact that the mean-field approximation is less accurate at highly coupled positions.

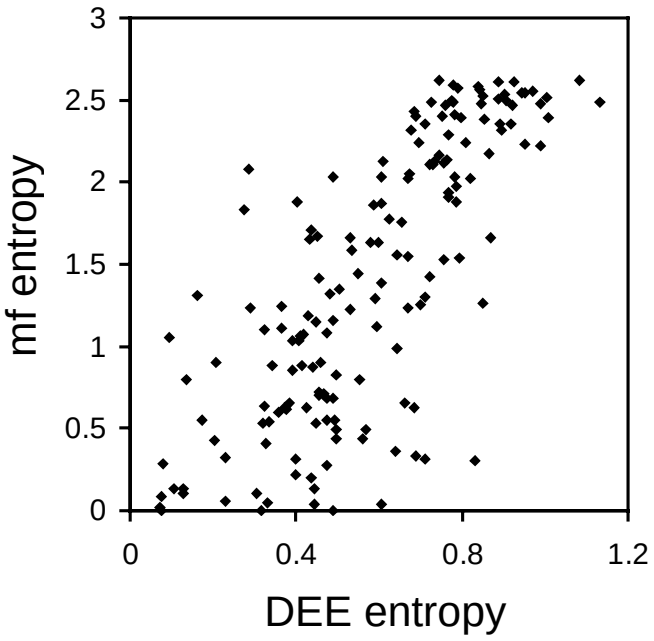

3
2.5
2
1.5
1
0.5
0
0
0.4
0.8
1.2
mf entropy
DEE entropy

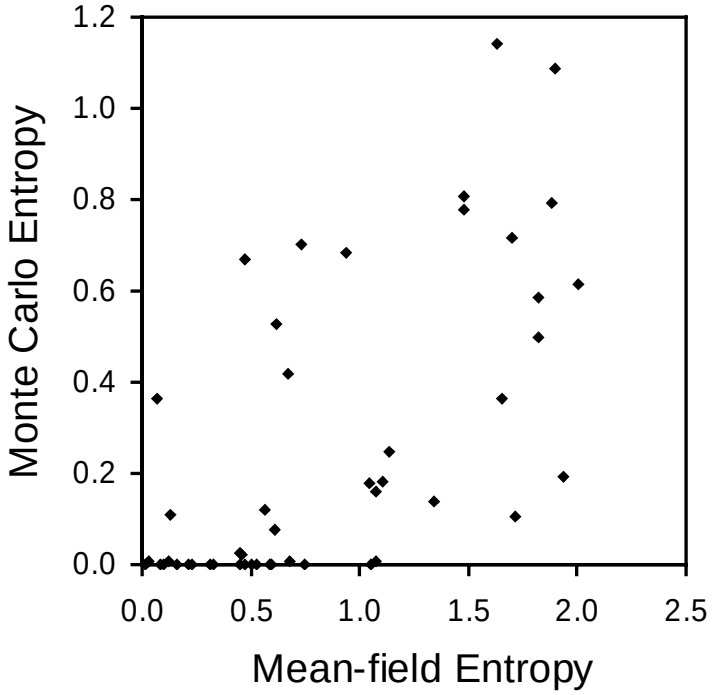
Figure B-3:

An example of the sequence list obtained by a Monte Carlo sampling from the GMEC. The structure used is antibody D1.3 (from Chapter 4) and only residues 24–34 are shown. The GMEC is the first sequence in the list and the other sequences are ranked according to energy. If the amino acid of a mutant sequence is the same as the GMEC, this is marked with a '|'. Only forty sequences are shown here; typically, over 10000 are recorded in order to calculate the sequence entropies of each residue.

	24A	25A	26A	28A	29A	30A	31A	32A	33A	34A
-485.4186	GLU	ALA	THR	ASP	VAL	TYR	PHE	PHE	ILE	ALA
-484.9586	|	|	|	|	|	|	|	|	|	|
-484.7058	|	|	|	|	|	|	|	|	|	|
-484.6677	|	|	|	|	|	|	|	|	|	|
-484.6606	|	|	|	|	|	|	|	|	|	|
-484.6336	|	|	|	|	|	|	|	|	|	|
-484.627	|	|	|	|	|	|	|	|	|	|
-484.572	|	|	|	|	|	|	|	|	|	|
-484.5691	|	|	|	|	|	|	|	|	|	|
-484.4584	|	|	|	|	|	|	|	|	|	|
-484.4427	|	|	|	|	|	|	LEU	|	|	|
-484.4276	|	|	|	|	|	|	|	|	|	|
-484.4043	|	|	|	|	|	|	LEU	|	|	|
-484.3664	|	|	|	|	|	|	LEU	|	|	|
-484.3235	|	|	|	|	|	|	|	|	LEU	|
-484.2793	|	|	|	|	|	|	|	|	|	|
-484.2761	|	|	SER	|	|	|	|	|	|	|
-484.2675	|	|	|	|	|	|	|	|	|	|
-484.2586	|	|	|	|	|	|	|	|	|	|
-484.2077	|	|	|	|	|	|	|	|	|	|
-484.2007	|	|	|	|	|	|	|	|	|	|
-484.1314	|	|	|	|	|	|	|	|	|	|
-484.1204	|	|	|	|	|	|	|	|	|	|
-484.1074	|	|	|	|	|	|	|	|	|	|
-484.0949	|	|	|	|	|	|	|	|	|	|
-484.0793	|	|	|	|	|	|	LEU	|	|	|
-484.0732	|	|	GLN	|	|	|	|	|	|
-484.048	|	|	|	|	|	|	|	|	|	|
-484.0299	|	|	|	|	|	|	|	|	|	|
-483.9904	|	|	|	|	|	|	LEU	|	|	|
-483.9867	|	|	|	|	|	|	|	|	|	|
-483.9828	|	|	|	|	|	|	LEU	|	|	|
-483.9718	|	|	|	|	|	|	|	|	|	|
-483.971	|	|	|	|	|	|	|	|	|	|
-483.9572	|	|	|	|	|	|	LEU	|	|	|
-483.9508	|	|	|	|	|	VAL	|	|	|	|
-483.95	|	|	|	|	|	|	|	|	|	|
-483.9265	|	|	|	|	|	|	|	|	|	|
-483.9246	|	|	|	|	|	|	|	|	|	|
-483.924	|	|	|	|	|	|	LYS	|	|	|
-483.894	ASP	|	|	|	|	|	|	|	|	|	|

Figure B-4:

The entropies that are calculated by the mean-field and Monte Carlo algorithms are compared. The data was obtained for the CDR residues of antibody D1.3. Many residues are predicted by Monte Carlo to have zero entropy. This is due to the sampling limitations of that algorithm.


Monte Carlo Entropy
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
Mean-field Entropy

Appendix C

Adding Ambient Temperature to the Sequence Entropy

In the protein design model, the mean-field calculation is done on the rotamer level. However, it is necessary to translate the rotamer probabilities into the amino acid probabilities required to calculate the site entropies (Equation 3-2). In doing this translation, it is necessary to include an additional temperature-the physical temperature of the system-as well as the simulation temperature that arises from the variational treatment. The physical temperature alters the importance that multiple rotamers of an amino acid have acceptable energies at a given residue. A high temperature biases the amino acid probabilities to be high when multiple rotamers of that amino acid have low energies. Conversely, a low temperature indicates that is more important for an amino acid to have a single, acceptable rotamer. Conceptually, this is similar to the flexible rotamer model where rotamers and sub-rotamers were considered (Mendes *et al.*, 1999). In the flexible rotamer model, the rotamer temperature and sub-rotamer temperature are equal.

The two temperatures can be included into the mean-field derivation by introducing a new free energy to be minimized is

$$E = \left\langle F \right\rangle_A - T_S S_S , \tag{C-1}$$

where E is the variational free energy, $\langle F \rangle_A$ indicates the free energy averaged over all sequences, T_S is the sequence temperature, and the sequence entropy is

$$S_S = -k_S \sum_i^N \sum_a p_i(a) \ln p_i(a) . \tag{C-2}$$

The probability that amino acid *a* exists at position *i* is the sum of the rotamer probabilities

$$p_i(a) = \sum_{r \in a} p_i(r) \tag{C-3}$$

where $r \in a$ indicates the set of all the rotamers for amino acid *a*,

The free energy is

$$\langle F \rangle_A = \langle U \rangle_A - T_R \langle S_R \rangle_A, \tag{C-4}$$

where T_R is the rotamer (physical) temperature. The internal energy averaged over all sequences is

$$\langle U \rangle_A = \sum_i^N \sum_a \sum_{r \in a} \gamma_i(r) p_i(r) + \sum_i^N \sum_j^N \sum_a \sum_{r \in a} \sum_b \sum_{r \in b} \gamma_{ij}(r,s) p_i(r) p_j(s) \lambda_{ij}\,, \tag{C-5}$$

and the rotamer entropy is

$$S_R = -k_R \sum_i^N \sum_a p_i(a) \sum_{r \in a} \frac{p_i(r)}{\sum_{r' \in a} p_i(r')} \ln \frac{p_i(r)}{\sum_{r' \in a} p_i(r')}. \tag{C-6}$$

The total free energy can then be written as

$$\begin{aligned} E = \langle U \rangle_A + \frac{1}{\beta_R} \sum_i^N \sum_a \sum_{r \in a} p_i(r) \ln p_i(r) + \left(\frac{1}{\beta_S} - \frac{1}{\beta_R} \right) \sum_i^N \sum_a \sum_{r \in a} p_i(r) \ln \sum_{r' \in a} p_i(r') \\ + \sum_i^N \mu_i \left(\sum_r p_i(r) - 1 \right) \end{aligned} \tag{C-7}$$

where β_R and β_S are the inverse sequence and rotamer temperatures, and μ are the lagrange multipliers such that the rotamer probabilities sum to unity at each position. The derivative of the free energy for all $p_i(r)$ is set to zero so

$$\gamma_i(r) + \sum_j^N \sum_b \sum_{s \in b} \gamma_{ij}(r,s) p_j(s) \lambda_{ij} + \frac{1}{\beta_R} \ln p_i(r) + \frac{1}{\beta_R} + \mu_i + \left(\frac{1}{\beta_S} - \frac{1}{\beta_R} \right) \left(\ln p_i(a) + \frac{p_i(r)}{p_i(a)} \right) = 0 \quad . \qquad \text{(C-8)}$$

Solving for μ_i, and rearranging gives

$$p_i(r) = \frac{1}{Z_i} \exp\left[-\beta_R \varepsilon_i(r) + \left(1 - \frac{\beta_R}{\beta_S} \right) \left(\ln p_i(a) + \frac{p_i(r)}{p_i(a)} \right) \right], \qquad \text{(C-9)}$$

where

$$\varepsilon_i(r) = \gamma_i(r) + \sum_j^N \sum_b \sum_{s \in b} \gamma_{ij}(r,s) p_j(s) \lambda_{ij} \qquad \text{(C-10)}$$

and

$$Z_i = \sum_r \exp\left[-\beta_R \varepsilon_i(r) + \left(1 - \frac{\beta_R}{\beta_S} \right) \left(\ln p_i(a) + \frac{p_i(r)}{p_i(a)} \right) \right]. \qquad \text{(C-11)}$$

Equations (C-9) and (C-11) demonstrate that the entropy contribution can be neglected when $\beta_R \cong \beta_S$.

Appendix D

Calculating a Joint Entropy for Two Structures

This work was done in collaboration with Deepshika Datta in the Mayo group, who has run the DEE calculations and is making and testing the protein G and engrailed homeodomain variants.

Protein evolution can be described as a random walk in sequence space. This space consists of all combinations of amino acids, connected through single mutational moves. Through rounds of mutagenesis and selections, a sequence or population of sequences drifts can drift through this space (Eigen and McCaskill, 1989). During this process, not all of sequence space is accessible. Most amino acid combinations are either unfolded or non-functional. The fraction of sequence space that retains structure or function defines the topology of the space that can be reached via evolution (Taverna and Goldstein, 2000).

It is particularly interesting to understand the relationship between sequence space and the space of all possible three-dimensional structures (Figure D-1). There are far fewer structures than sequences, so the mapping of structure space onto sequence space is highly degenerate (Chothia and Lesk, 1986; Aronson *et al*., 1994; Cordes *et al*., 1996). In addition, diffusion through sequence space is much faster than diffusion through structure space (Govindarajan and Goldstein, 1997). This is determined by the connectivity of the space. If more than a few mutations separate the portions of sequence space corresponding to different structures, then it is unlikely that evolution will be able to discover the new structure (Blanco *et al*., 1999). However, if the spaces corresponding to different structures are interwoven, as is the case with RNA secondary structures, then

evolution can rapidly discover new structural topologies through diffusion (Fontana and Schuster, 1998).

It is desirable to characterize sequence-structure map from two perspectives. First, the problem of understanding the minimal number of amino acid substitutions to convert one structure to another has been defined as the "Paracelsus Challenge" (Rose and Creamer, 1996). Using sequence comparison and rational substitutions, the closest two sequences have converged is 50% identity, which is a huge separation in sequence space (Blanco *et al*., 1999). Second, it is useful to understand the energetics of converting one structure into another (Glykos *et al*., 1999; Cordes *et al*., 1999). A switch could be achieved through some external perturbation, such as through the addition of a metal ligand that stabilizes one of the structures. The computational tools used to study inverse folding have the potential of making progress towards understanding both of these goals (Babajide *et al*., 2001; Koehl and Levitt, 2002). There is theoretical evidence that designing a sequence to be close in sequence space for two structures is equivalent to designing a sequence to be close in energy to two structures (Bornberg-Bauer and Chan, 1999; Wang *et al*., 2000)

Measuring the sequence entropy effectively provides a set of amino acid sequences that are consistent with a structure (Chapter 3, Appendix B). By comparing the entropy profiles for two structures, those residues can be identified that are important contributions to each structure (Figure B-2). For example, if a residue has a high entropy in protein G, but a low entropy in engrailed homeodomain, this indicates that the protein G sequence can traverse towards engrailed in sequence space, but not vise versa. Similarly, when a residue has a high entropy in for both structures, this means the amino

acid identity at this position is not important in either structure and this dimension in sequence space can be traversed by both.

There is a limitation to simply comparing the sequence entropy for two structures. If at residue *i*, amino acids 1-10 are allowed with equal probability in protein G and amino acids 11-20 are allowed in engrailed with equal probability, then the entropy at this position would be high. However, there is no single amino acid that can adopt both structures. This is a particular problem when considering binary patterning as one structure may allow all hydrophobic amino acids and the other all hydrophilic amino acids. This motivates the calculation of a joint probability that amino acid *a* exists at residue *i* of both structure α and structure β,

$$p_{\alpha,\beta}(i_a) = p_\alpha(i_a)\, p_\beta(i_a), \tag{D-1}$$

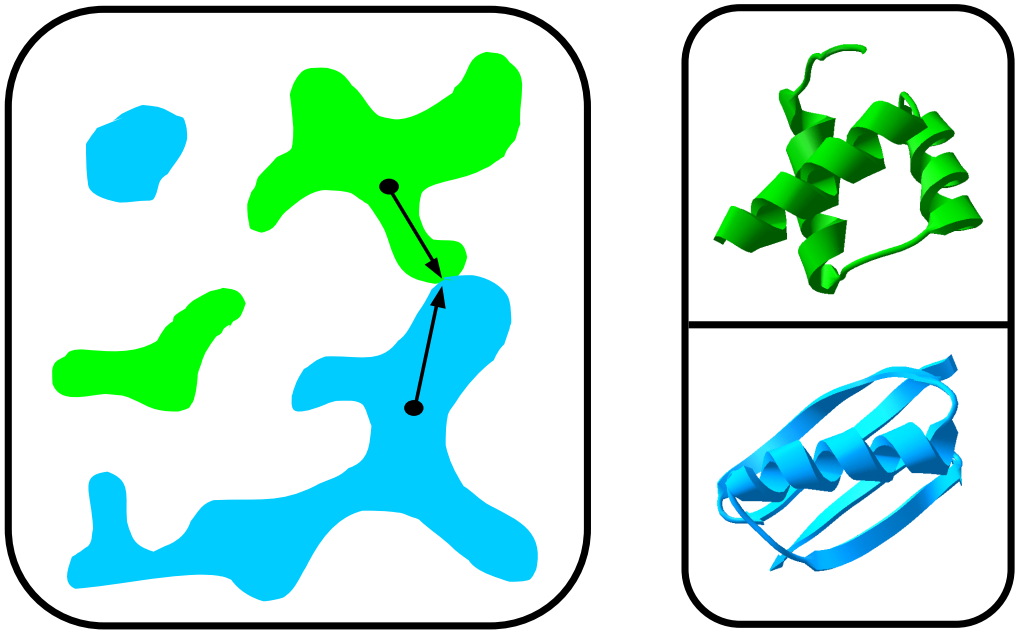
where the individual probabilities represent the ability for amino acid *a* to exist in each structure. If a probability is high for one structure, but low for the other, then this will result in a low joint probability. If the two probabilities are high, then this means that the amino acid can be substituted at that position for both. These probabilities can be calculated by mean-field theory and then used to calculate the corresponding entropies from Equation (3-2). The joint entropy for the protein G / engrailed comparison is shown in Figure D-2. The specific amino acids that are allowed in both structures are shown in Figure D-3.

The joint entropy profile provides guidance when trying to design a sequence that is close to two structures in sequence space (or energy). Deepshika Datta is testing this hypothesis by using the entropy profile to determine those residues that should be mutated to produce a sequence that has very low energies in both the protein G and

engrailed topologies. DEE was used to design these positions from two starting sequences to test how close the two structures can come together in sequence space. Further, several positions have been mutated to histidine that will collectively coordinate a Ni ligand in the engrailed structure, but not the protein G structure. If a sequence that spontaneously folds into protein G is close in energy to the engrailed structure, then the addition of this metal could facilitate a structural switch. These ideas are currently being tested in the Mayo lab.

Figure D-1:

A cartoon is shown demonstrating the mapping between sequence and structure space. Sequence space is the set of all possible amino acid sequences for a given protein length and structure space is the set of all possible structures. These structures are somehow distributed in sequence space, although the properties of this distribution remain unclear. The goal of the joint entropy algorithm is to identify those directions in sequence space that will lead to the conversion of one structure into another (black arrows).

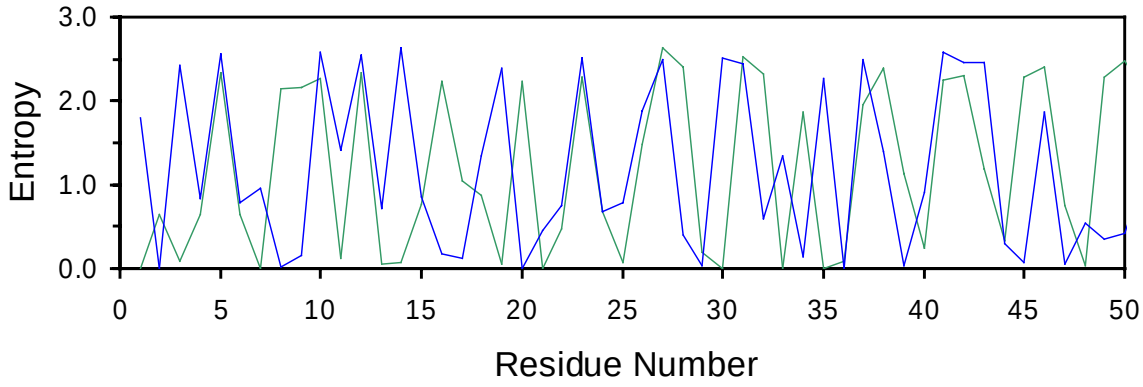


Sequence Space

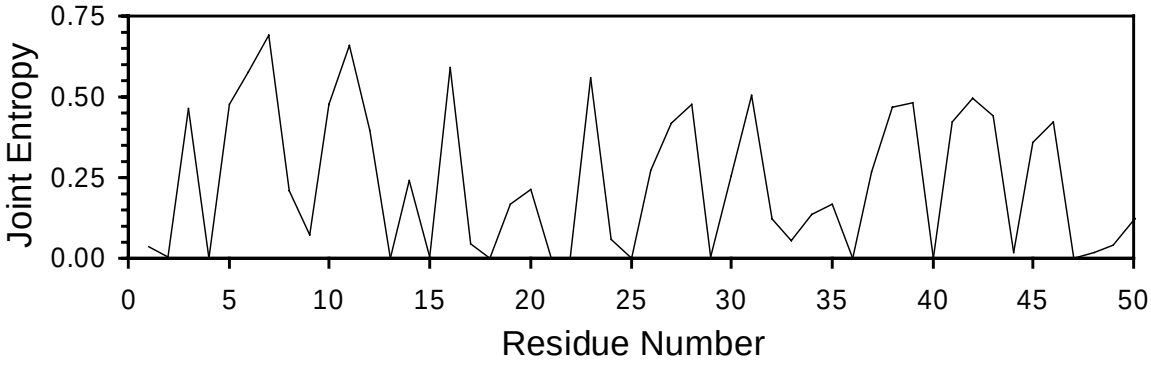
Structure Space

Figure D-2:

(A) The sequence entropies are compared for protein G (blue) and engrailed homeodomain (green) for T = 400 K. The numbering convention is for engrailed, so the first residue is residue six of protein G. Regions of high entropy in both profiles indicate that the amino acid identity is unconstrained for both structures, whereas a dual low entropy means the wild-type amino acid is fixed in both structures. A residue that has a high entropy for structure A and a low entropy for structure B indicates that a mutation can be made to convert the sequence from A to B, but not vise versa. **(B)** The joint entropy of the two structures as calculated from Equations (D-2) and (3-2).


3.0
2.0
1.0
0.0
Entropy
0
5
10
15
20
25
30
35
40
45
50
Residue Number

(A)


0.75
0.50
0.25
0.00
Joint Entropy
0
5
10
15
20
25
30
35
40
45
50
Residue Number

(B)

Figure D-3:

The amino acids are listed that are predicted to be acceptable in both the protein G (blue sequence) and engrailed homeodomain (green sequence) structural contexts. The amino acids for which $p(i_a) > 0.1$ for both structures are shown in red and $p(i_a) > 0.01$ are shown in black. The numbering convention is for engrailed, so the first residue is residue six of protein G. Residues for which the amino acid identity is identical in the parents are indicated.

Position	enh	pg	
1	T	I	
2	K	L	
3	F	N	I
4	D	G	
5	E	K	E Q R M Y S C D W T K N
6	Q	T	R Q S K
7	L	-	
8	K	K	K
9	R	G	M C K
10	R	E	R E Q L I M V C T K N
11	L	T	L I
12	E	T	M R E Q L Y F S C W T K N
13	E	T	E Q
14	E	-	
15	F	A	
16	K	V	R
17	R	D	N
18	D	A	
19	R	A	N
20	R	T	E
21	L	A	
22	T	E	
23	N	K	R Q E L I M Y S C D W T K N
24	Q	V	
25	R	F	
26	R	K	S C D T N
27	H	Q	Q R E L M Y S C D W T K N
28	D	Y	Q E N
29	L	A	
30	S	N	S
31	Q	D	E Q R L M Y S C D T K N
32	K	N	N
33	L	G	
34	G	V	C
35	I	D	L
36	N	G	
37	E	-	
38	E	W	N R E Q S C D
39	L	T	Q
40	I	Y	
41	E	D	R Q E L I M S C D T K N
42	D	-	
43	W	A	E Q M T K
44	F	T	
45	R	K	R
46	R	T	R E Q M S C D T K N
47	K	F	

Appendix E

Combinatorial Libraries Based on Schema Disruption

These calculations were inspired by experiments being performed by Jonathan Silberg and Michelle Meyer in the Arnold group. They performed all of the experiments described in this appendix.

The schema profile described in Chapter 5 provides a method to visualize where crossovers are likely to disrupt the three-dimensional structure. The profile condenses higher-dimensional information in a way that can be rapidly calculated and visualized. The difficulty with the profile is that nearly every position in the protein structure could be an acceptable crossover if a proper set of additional compensating crossovers is made. A minimum in the profile predicts that this is unlikely when the average fragments size is large (the number of crossovers is small). This appendix aims to test this assumption and to introduce new computational tools for visualizing disruption for constructing and analyzing libraries for which the number and location of crossovers are fixed. A motivation for this work is to develop computational methods to complement experimental strategies for analyzing the quality of shuffled libraries (Joern *et al*., 2002).

A library that is constructed with a fixed set of defined crossovers can be computationally analyzed by creating all possible hybrids and testing the disruption of each. It is desirable to find a set of crossovers that enriches the fraction of the library that is predicted to be functional. If two parents are being shuffled, a library of c defined crossovers produces 2^{c+1} hybrids, half of which have to be computationally analyzed. This is because the schema theory predicts that the disruption caused by a hybrid sequence and the inverse parental heredity is identical.

Three libraries of hybrid TEM-1/PSE-4 β-lactamases are studied (Figure E-1). The MIN library corresponds to a crossover at every minimum in the schema disruption profile. The MAX library corresponds to a crossover at every maximum in the library. Both of these libraries have seven crossovers, corresponding to 256 possible hybrids. In addition, a library corresponding to crossovers at both the minima and maxima (MIN-MAX) is analyzed. This library has thirteen crossovers, corresponding to 16384 hybrids.

For the TEM-1/PSE-4 system, it was experimentally found that there is a transition in disruption at ~ 28 before function is lost (Figure 5-9). After calculating the disruption of each library, the fraction that is below this transition is determined (Figure E-2). The MIN library is the most enriched, with nine (7%) low disruption hybrids and the MAX library has four (3%) low disruption hybrids (Figure E-3). When the locations for acceptable crossovers are compared with the schema disruption profile, the minima in the profile tend to have more crossovers. However, it is not impossible for crossovers to occur near maxima, only less likely. An exception to the inverse correlation is the region around residue 105, which is a minimum in the profile, but few crossovers occur in the designed libraries. The MIN-MAX library has the most low-disruption hybrids (54), but this is a small fraction of the entire library (0.5%) (Figure E-4). When compared to the schema disruption profile, an inverse correlation exists, but is not as strong as was observed for the MIN and MAX libraries. This is due to the smaller fragment size (14 residues) in the MIN-MAX library as compared to the MIN and MAX libraries (33 residues). The assumptions behind the construction of the schema disruption profile are not valid for small fragment sizes.

The MIN-MAX library was constructed experimentally by synthesizing oligonucleotides to correspond to each of the fourteen fragments. Each fragment had sufficient DNA overlap so that the hybrid library could be combinatorially constructed via PCR techniques. Members of the naive library were sequenced to confirm that the fragments were incorporated without bias. The library was then screened for activity towards the degradation of ampicillin. The hybrids that survive this selection are strongly biased towards having a low disruption (Figure E-5). It is found that the low-disruption hybrids are over 200-fold enriched in the selected library. Based on this library analysis, it is found that the transition is slightly higher than observed in the smaller two-cut library presented in Chapter 5. Based on the MIN-MAX library, the transition is around 35. These results demonstrate the importance of considering the disruption in designing targeted libraries. Further, by analyzing the small number of high-disruption hybrids that survive selection, the parameters used to calculate the disruption can be more finely tuned.

The optimal set of seven crossovers to produce a library should satisfy two constrains: a large fraction of the library should be folded, and the folded hybrids should be diverse. Diversity can be measured as the number of effective mutations that occur for a hybrid. The three libraries that have been studied were constructed using the schema profile as a guide, but there are many possible sets of crossovers that satisfy the profile. It unclear as to whether the MIN library, where a single crossover is placed at each minimum, is truly the optimal seven-crossover library. Libraries were constructed from sets of seven randomly assigned crossovers to examine the limits of optimizing the crossover locations. A minimum fragment size of ten residues is enforced. For each

library, the schema disruption was computationally determined for all possible combinations of fragments. Then, the fraction of the library that has a disruption below 28 and the average diversity in the low-disruption portion of the library is recorded. This process is repeated for 100,000 libraries of randomly chosen crossovers and some of these libraries are highly enriched in diverse, low-disruption hybrids (Figure E-6). When the crossovers that occur on the low-disruption hybrids are compared with the schema disruption profile, an inverse correlation is observed (Figure E-7). Optimal sets of crossovers can also be identified that produce 40% low disruption hybrids with an average of 40 mutations per sequence. The best set of crossovers found in the 100,000 random libraries is shown in Figure E-7B.

Figure E-1:

The three TEM-1/PSE-4 β-lactamase libraries referred to in this Appendix are shown. The MIN library has crossovers at each minimum of the schema profile. The MAX library has crossovers at each maximum of the schema profile. Both the MAX and MIN libraries have seven crossovers, corresponding to eight fragments. The MIN-MAX library has thirteen crossovers (fourteen fragments), where each fragment corresponds to a maximum and minimum of the schema profile. The average fragment in the MIN and MAX libraries is 33 and the MIN-MAX library is 14.

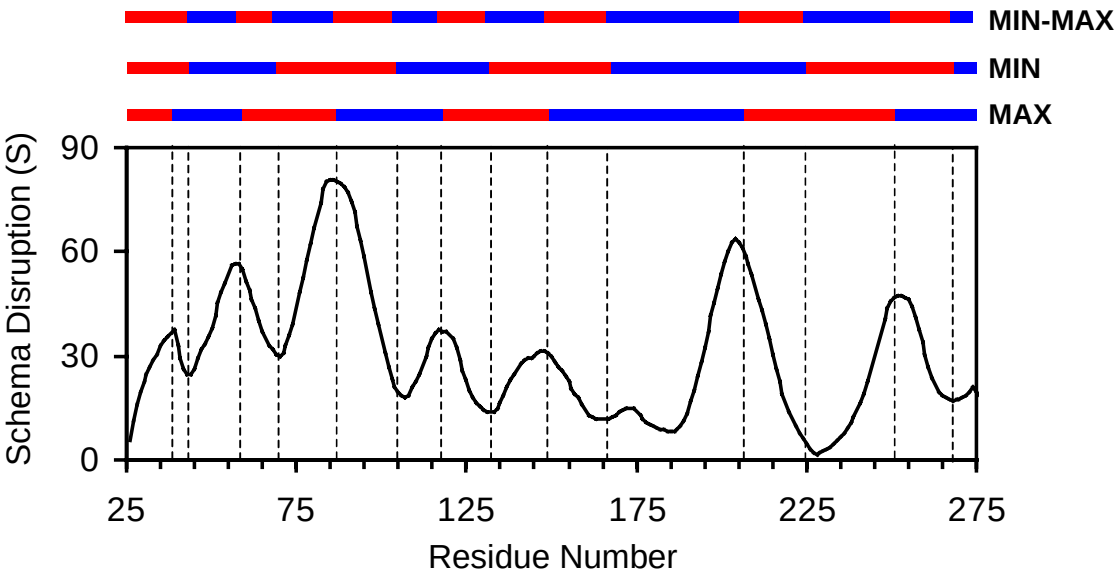

MIN-MAX
MIN
MAX
Schema Disruption (S)
90
60
30
0
25
75
125
175
225
275
Residue Number

Figure E-2:

The fraction of each targeted library that has a crossover disruption less than the threshold beyond which function is lost. The MIN library is enriched with 7% functional hybrids whereas the MAX library has 3% functional hybrids. The MIN-MAX library has the most functional hybrids, but the fraction is on 0.5% because the size of the library is much larger than the MIN or MAX libraries.

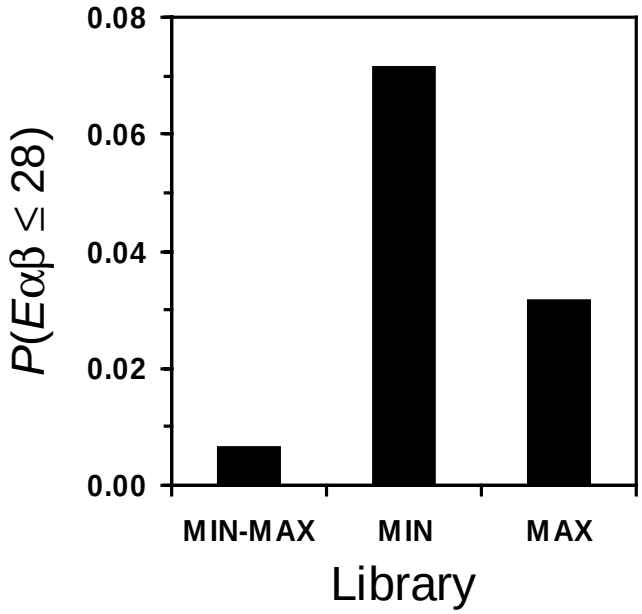
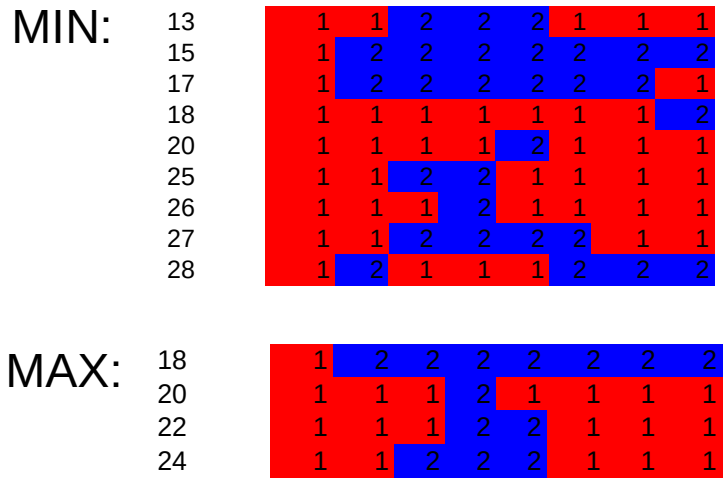
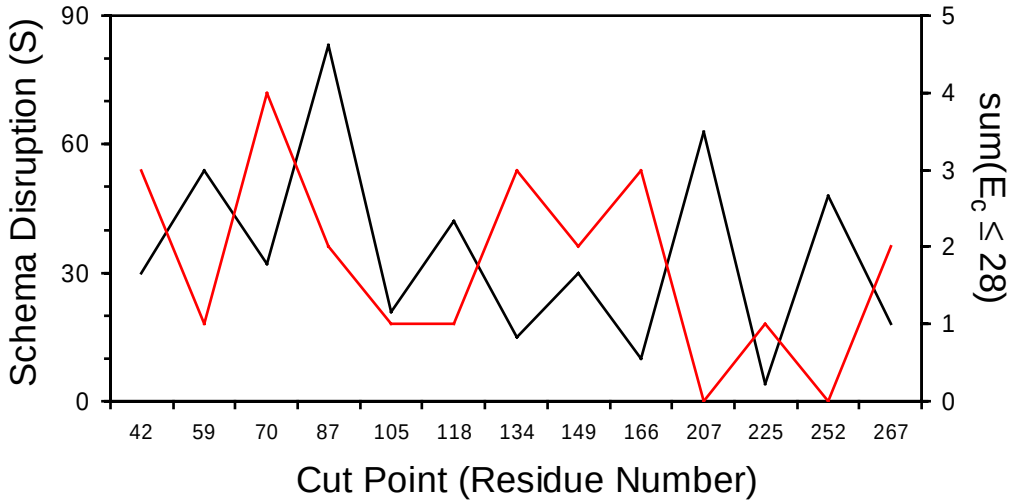

0.08
0.06
0.04
0.02
0.00
P(Eαβ ≤ 28)
MIN-MAX
MIN
MAX
Library

Figure E-3:

(A) The hybrids that are predicted to be functional in the MIN and MAX libraries are shown with the predicted disruption. The average number of fragments that are shuffled in the MIN library is 2.11 and the MAX library is 1.75. **(B)** The number of crossovers in the MIN and MAX libraries is compared with the schema profile. There is an inverse correlation: more crossovers occur at minima in the profile. The exception is at residue 105, which is a minimum, but very few crossovers occur at this residue.


MIN:
13 1 1 2 2 2 1 1 1
15 1 2 2 2 2 2 2 2
17 1 2 2 2 2 2 2 1
18 1 1 1 1 1 1 1 2
20 1 1 1 1 2 1 1 1
25 1 1 2 2 1 1 1 1
26 1 1 1 2 1 1 1 1
27 1 1 2 2 2 2 1 1
28 1 2 1 1 1 2 2 2
MAX:
18 1 2 2 2 2 2 2 2
20 1 1 1 2 1 1 1 1
22 1 1 1 2 2 1 1 1
24 1 1 2 2 2 1 1 1

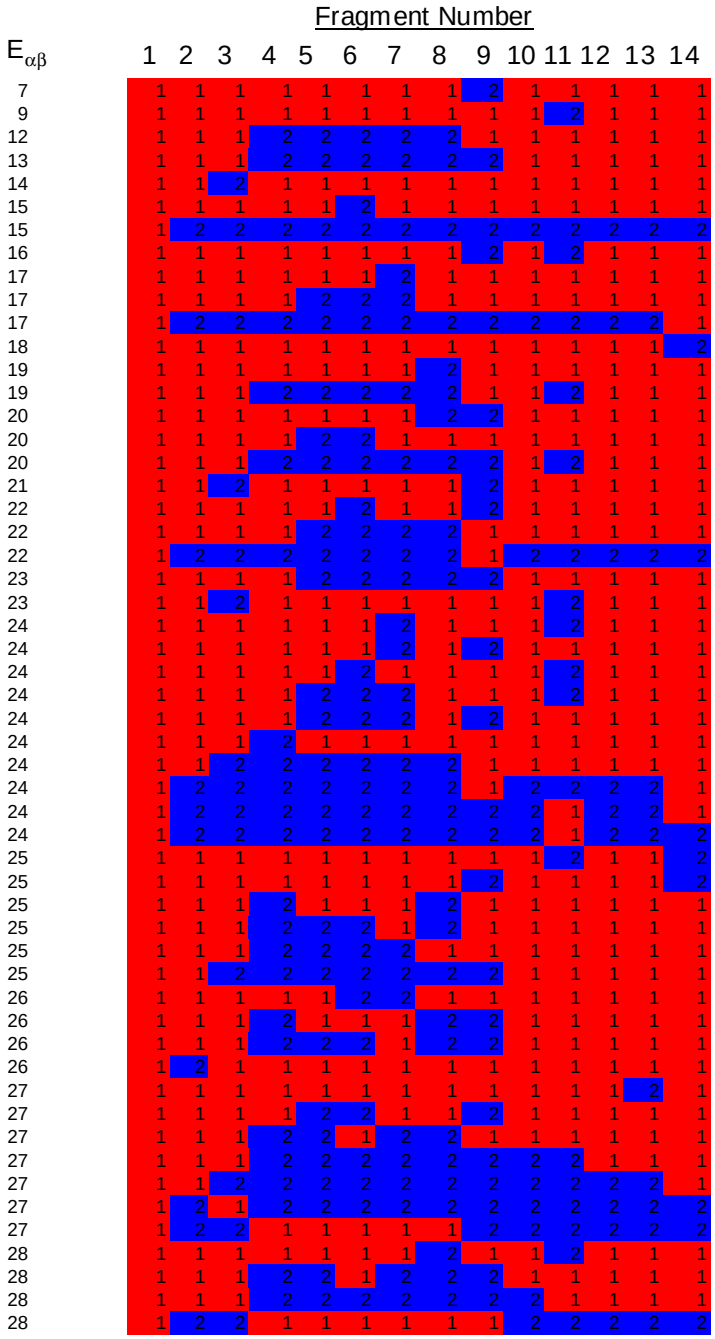
(A)


Schema Disruption (S)
90
60
30
0
5
4
3
2
1
0
sum(E_c ≤ 28)
42 59 70 87 105 118 134 149 166 207 225 252 267
Cut Point (Residue Number)

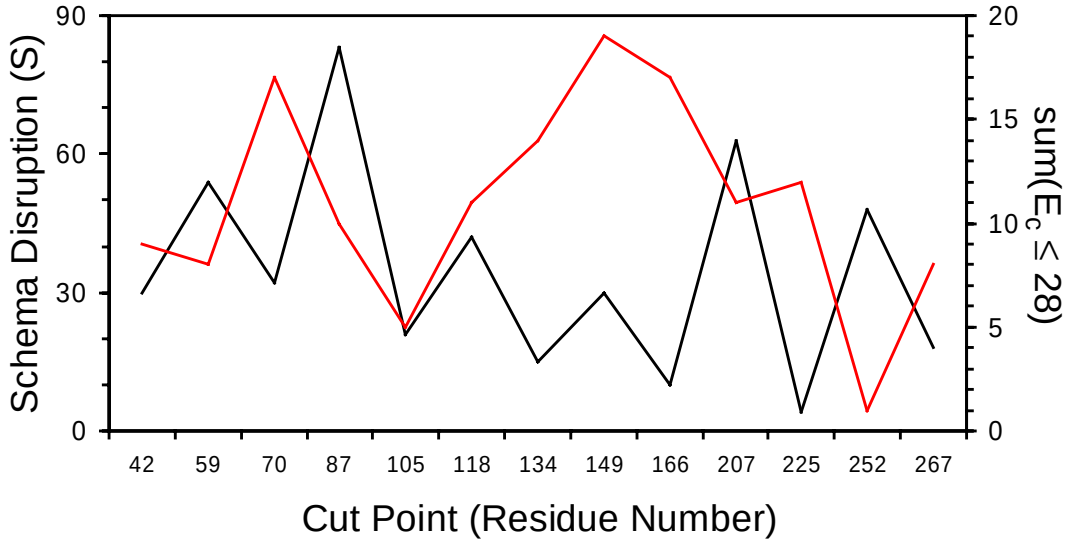
(B)

Figure E-4:

(A) The hybrids that have a calculated disruption below 28 are shown for the MIN-MAX library. There are more viable hybrids than the MIN or MAX libraries, but the fraction of low-disruption hybrids is less because the MIN-MAX library has many more possible hybrids. **(B)** The results in (A) are compared with the schema disruption profile. There is still an anti-correlation between the number of crossovers and the schema disruption, but the relationship is less strong than in the seven crossover libraries. This is due to the smaller average fragment size (14 versus 33 for the MIN and MAX libraries).


Fragment Number
$E_{\alpha\beta}$
1 2 3 4 5 6 7 8 9 10 11 12 13 14
7 1 1 1 1 1 1 1 1 2 1 1 1 1 1
9 1 1 1 1 1 1 1 1 1 1 2 1 1 1
12 1 1 1 2 2 2 2 2 1 1 1 1 1 1
13 1 1 1 2 2 2 2 2 2 1 1 1 1 1
14 1 1 2 1 1 1 1 1 1 1 1 1 1 1
15 1 1 1 1 1 2 1 1 1 1 1 1 1 1
15 1 2 2 2 2 2 2 2 2 2 2 2 2 2
16 1 1 1 1 1 1 1 1 2 1 2 1 1 1
17 1 1 1 1 1 1 2 1 1 1 1 1 1 1
17 1 1 1 1 2 2 2 1 1 1 1 1 1 1
17 1 2 2 2 2 2 2 2 2 2 2 2 2 1
18 1 1 1 1 1 1 1 1 1 1 1 1 1 2
19 1 1 1 1 1 1 1 2 1 1 1 1 1 1
19 1 1 1 2 2 2 2 2 1 1 2 1 1 1
20 1 1 1 1 1 1 1 2 2 1 1 1 1 1
20 1 1 1 1 2 2 1 1 1 1 1 1 1 1
20 1 1 1 2 2 2 2 2 2 1 2 1 1 1
21 1 1 2 1 1 1 1 1 2 1 1 1 1 1
22 1 1 1 1 1 2 1 1 2 1 1 1 1 1
22 1 1 1 1 2 2 2 2 1 1 1 1 1 1
22 1 2 2 2 2 2 2 2 1 2 2 2 2 2
23 1 1 1 1 2 2 2 2 2 1 1 1 1 1
23 1 1 2 1 1 1 1 1 1 1 2 1 1 1
24 1 1 1 1 1 1 2 1 1 1 2 1 1 1
24 1 1 1 1 1 1 2 1 2 1 1 1 1 1
24 1 1 1 1 1 2 1 1 1 1 2 1 1 1
24 1 1 1 1 2 2 2 1 1 1 2 1 1 1
24 1 1 1 1 2 2 2 1 2 1 1 1 1 1
24 1 1 1 2 1 1 1 1 1 1 1 1 1 1
24 1 1 2 2 2 2 2 2 1 1 1 1 1 1
24 1 2 2 2 2 2 2 2 1 2 2 2 2 1
24 1 2 2 2 2 2 2 2 2 2 1 2 2 1
24 1 2 2 2 2 2 2 2 2 2 1 2 2 2
25 1 1 1 1 1 1 1 1 1 1 2 1 1 2
25 1 1 1 1 1 1 1 1 2 1 1 1 1 2
25 1 1 1 2 1 1 1 2 1 1 1 1 1 1
25 1 1 1 2 2 2 1 2 1 1 1 1 1 1
25 1 1 1 2 2 2 2 1 1 1 1 1 1 1
25 1 1 2 2 2 2 2 2 2 1 1 1 1 1
26 1 1 1 1 1 2 2 1 1 1 1 1 1 1
26 1 1 1 2 1 1 1 2 2 1 1 1 1 1
26 1 1 1 2 2 2 1 2 2 1 1 1 1 1
26 1 2 1 1 1 1 1 1 1 1 1 1 1 1
27 1 1 1 1 1 1 1 1 1 1 1 1 2 1
27 1 1 1 1 2 2 1 1 2 1 1 1 1 1
27 1 1 1 2 2 1 2 2 1 1 1 1 1 1
27 1 1 1 2 2 2 2 2 2 2 2 1 1 1
27 1 1 2 2 2 2 2 2 2 2 2 2 2 1
27 1 2 1 2 2 2 2 2 2 2 2 2 2 2
27 1 2 2 1 1 1 1 1 2 2 2 2 2 2
28 1 1 1 1 1 1 1 2 1 1 2 1 1 1
28 1 1 1 2 2 1 2 2 2 1 1 1 1 1
28 1 1 1 2 2 2 2 2 2 2 1 1 1 1
28 1 2 2 1 1 1 1 1 1 2 2 2 2 2

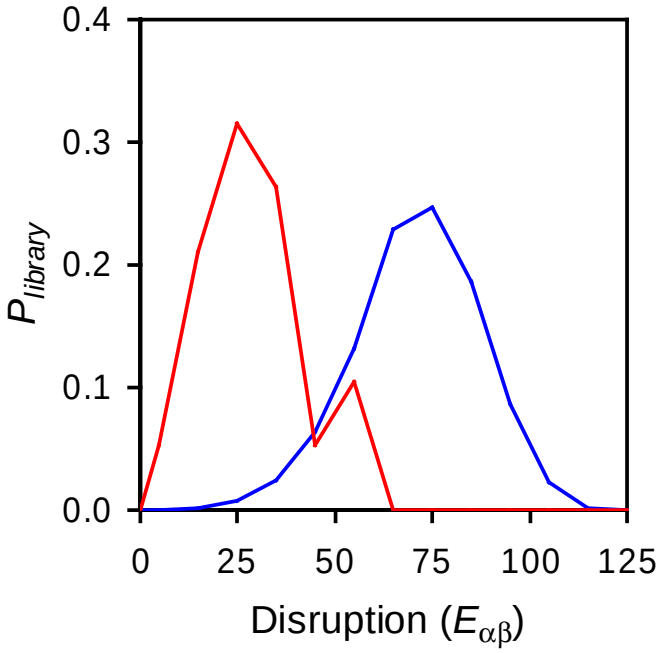
(A)


Schema Disruption (S)
90
60
30
0
20
15
10
5
0
sum(Ec ≤ 28)
42
59
70
87
105
118
134
149
166
207
225
252
267
Cut Point (Residue Number)

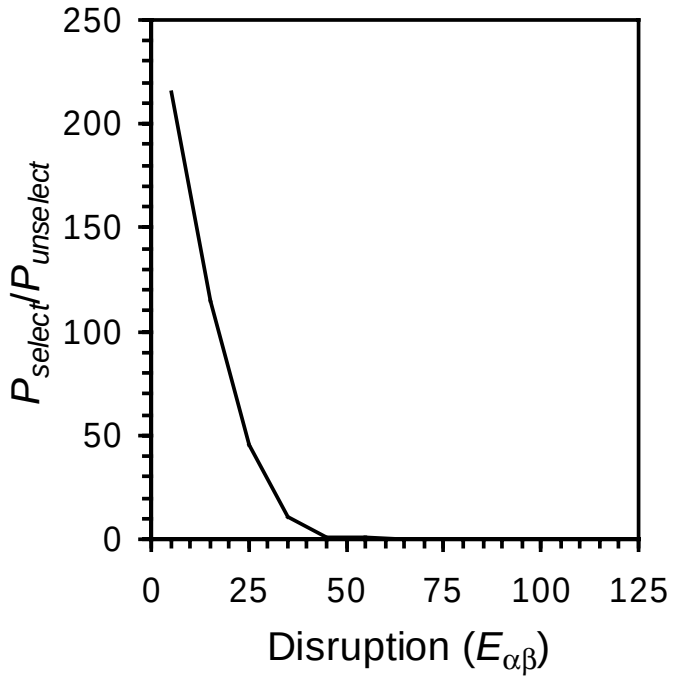
(B)

Figure E-5:

The experimental results from the selection experiment with the MIN-MAX library are shown. **(A)** The blue line represents the distribution of disruption that is contained in the entire unselected library. The red line shows the distribution of the hybrids that survive selection on ampicillin with a MIC > 20 μg/ml. The selected distribution is significantly shifted towards low-disruption hybrids. **(B)** The enrichment of the selected library is plotted as a function of the disruption. Here, the enrichment is defined as the ratio of the probability of a hybrid being found in the selected library to the probability of being found in the unselected library. The low-disruption hybrids have over 200-fold enrichment in the selected library.


0.4
0.3
0.2
0.1
0.0
$P_{library}$
0
25
50
75
100
125
Disruption ($E_{\alpha\beta}$)

(A)


250
200
150
100
50
0
$P_{select}/P_{unselect}$
0
25
50
75
100
125
Disruption ($E_{\alpha\beta}$)

(B)

Figure E-6:

The properties of 10,000 libraries with random sets of seven crossovers are shown. The minimum fragment size was restricted to be 10 residues. For each library, the fraction of the library with a disruption below 28 ($E_{\alpha\beta} \leq 28$) was calculated as well as the average diversity in the low-disruption hybrids. The diversity was calculated by counting the average number of effective mutations that occur for each hybrid with a schema disruption below 28. The optimal libraries are in the upper right portion of the graph. These sets of crossovers correspond to the largest and most diverse libraries. While the MIN library (red point) is better than the MAX library (blue point), there are more optimal libraries.

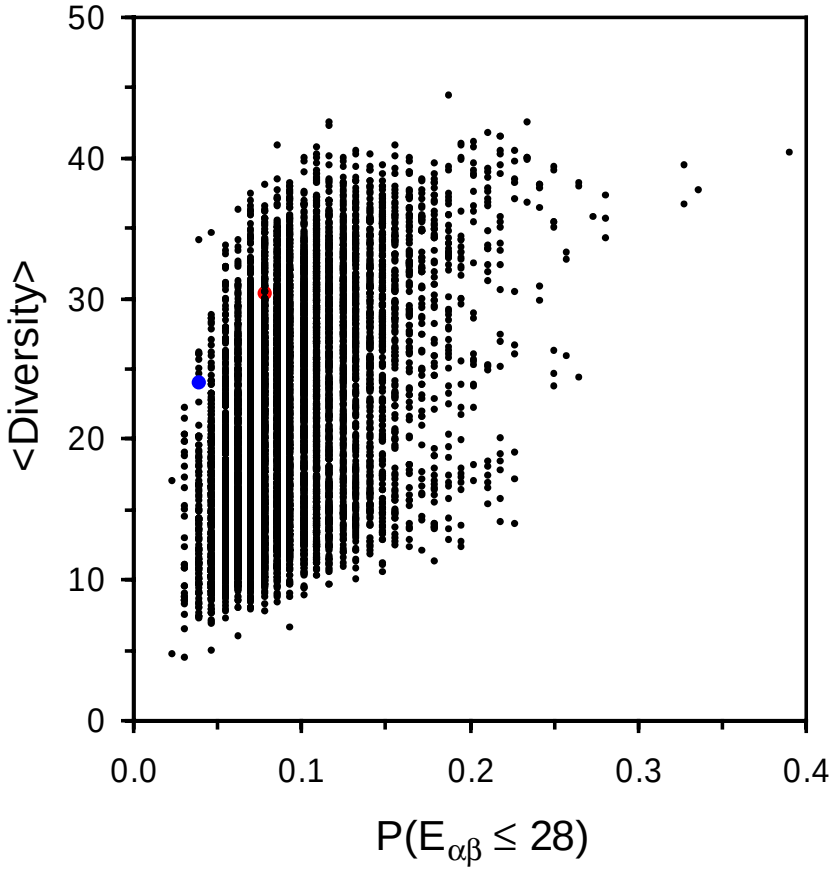
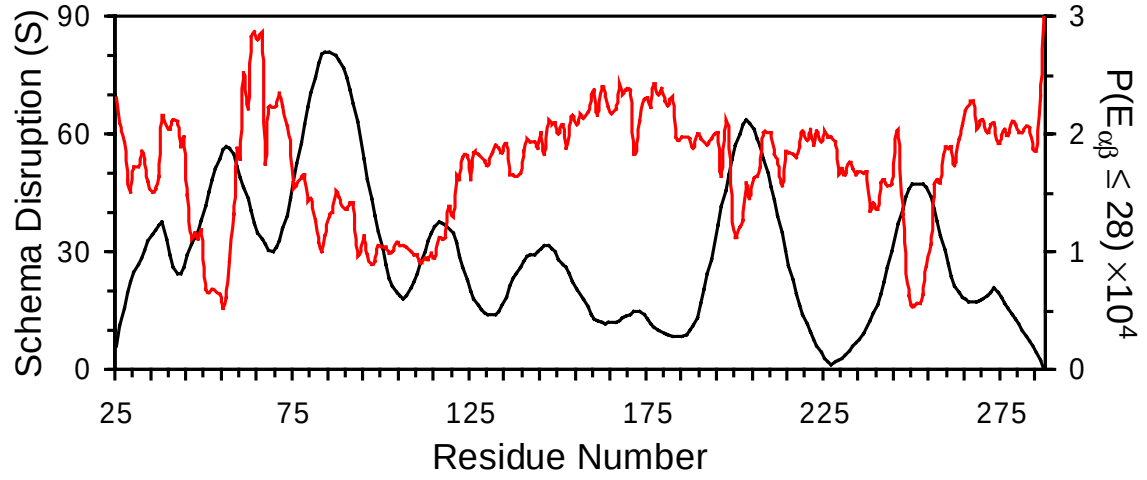
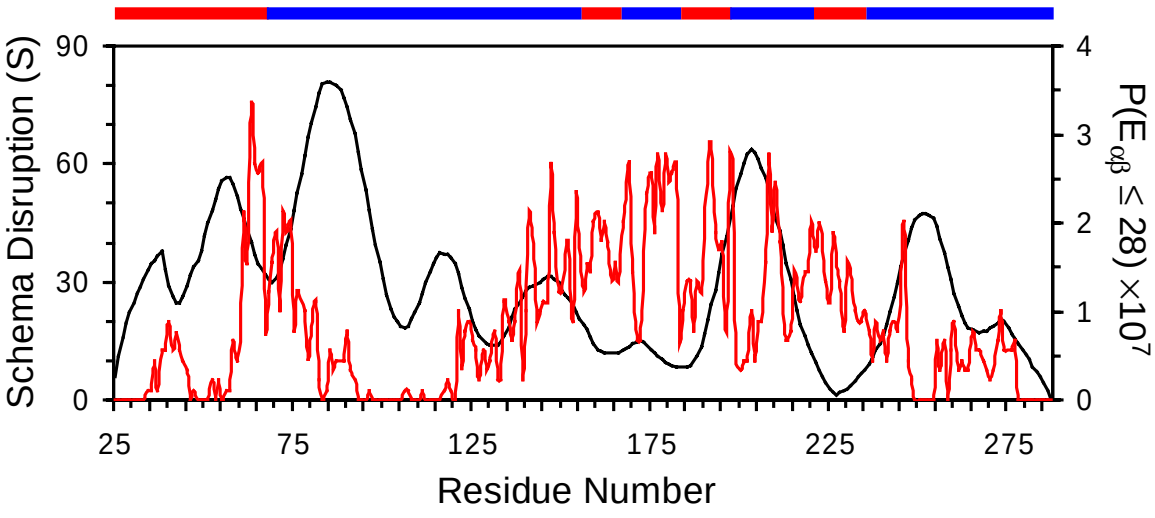

50
40
30
20
10
0
<Diversity>
0.0
0.1
0.2
0.3
0.4
P(E$_{\alpha\beta}$ ≤ 28)

Figure E-7:

The properties of 100,000 randomly generated libraries of seven crossovers are shown. **(A)** All of the crossovers that lead to hybrids with disruption below 28 are shown. For these data, there are no fragment size or diversity restrictions. With the exception of the region around residue 105, crossovers are less likely to occur at maxima in the schema disruption profile. **(B)** The properties of 100,000 libraries of seven randomly generated crossovers with the additional restriction that crossovers cannot occur within 10 residues of each other. The subset of libraries that have greater than 25% of hybrids with disruption less than 28 and an average diversity greater than 35 (Figure E-6) are shown. The best set of crossovers found is marked above the graph (40% of the library is low-disruption and the average diversity is 40.4 effective mutations per low-disruption hybrid).


Schema Disruption (S)
90
60
30
0
3
2
1
0
$P(E_{\alpha\beta} \leq 28) \times 10^4$
25
75
125
175
225
275
Residue Number

(A)


Schema Disruption (S)
90
60
30
0
4
3
2
1
0
$P(E_{\alpha\beta} \leq 28) \times 10^7$
25
75
125
175
225
275
Residue Number

(B)

Appendix F

Non-homologous Recombination

It has been observed that certain structural motifs are ubiquitous in proteins (Orengo *et al.*, 1994; Orengo *et al.*, 1997). The underlying cause of this observation is unclear. It is possible that motifs exist because they represent topologies that fold rapidly, they are robust to mutagenesis, they are functionally important, they represent biases in structure-determining experiments, or they are the result of evolutionary accident. It is intriguing to think that motifs could have emerged early in evolution and then were combined by recombination to create more complicated topologies. In Chapter 5, it was demonstrated that recombination tends to preserve clusters of interacting amino acids when parents sharing homologous structures are recombined. When the structures of these clusters, or schema, are compared between non-homologous proteins, common subunits are detected. In this appendix, we study the prevalence of β-lactamase schema in the protein structure database (PDB) and experimentally swap a schema from MADS box into β-lactamase.

The sequence identity between schemas from two non-homologous parents is expected to be low, so the schema profile is calculated from the TEM-1 structure with no probability matrix (Figure F-1). This divides the protein structure into schema that are similar to those presented in Chapter 5. The PDB is then searched using the combinatorial extension of the optimal path (CE) algorithm (Shindyalov, and Bourne, 1998). The CE program is particularly useful because it can compare smaller fragments than competing search algorithms. Still, several of the schema are too small to record matches in the database. To overcome this problem, schema 1 and 2 and schema 8 and 9

are combined to create a larger subunit. Even though schemas 4 and 6 are also small, combining them with adjacent schema does not increase the total number of structures discovered.

The results of the PDB search are shown in Figure F-2. Of the matches that are found, the homologous proteins (such as β-lactamase variants) are removed. In addition, multiple matches for a single protein (such as a set of HIV RT variants) are counted only once. Of the positive matches, most correspond with schema 1+2, 3, and 8+9. These schema participate in a larger, non-contiguous domain that has been identified previously to commonly occur in proteins (Orengo *et al*., 1994; Orengo *et al*., 1997). Despite this, there are examples where each of these schema occur independently (Figure F-3). It is interesting that while these matches correspond with a schema in β-lactamase, they are not necessarily complete schema in the non-homologous structure.

Schema 1+2 is found in many non-homologous structures. In addition, it was successfully swapped between the PSE-4 and TEM-1 sequences (Chapter 5 – hybrid 4A). For these reasons, it is chosen as a good candidate to test recombination between non-homologous structures. When the best matches to schema 1+2 are compared, their rms is nearly identical (Figure F-4). However, there are several problems in predicting whether they can be swapped between structures. One problem is the variability in the connection point between the schemas. It may impossible to twist a non-homologous schema 1+2 into the background of the β-lactamase structure. This may be overcome by introducing a variable peptide sequence to connect the schema when they are recombined (O'Maille *et al*., 2002). Another problem is the disparity between the properties of the amino acids when compared in a sequence alignment. The most striking problem is the lack of a

common binary pattern. Some of the non-homologous matches to schema 1+2 are completely buried in their wild-type structural context and are therefore entirely composed of hydrophobic amino acids. In contrast schema 1+2 is relatively exposed in β-lactamase and has a distinct binary pattern. The non-homologous schema 1+2 that is the closest in rms and amino acid properties is from a transcription factor, MADS box protein (Figure F-5). The binary pattern is fairly conserved despite sharing only 11% amino acid identity with TEM-1 and 6% identity with PSE-4. However, there are some notable amino acid substitutions at residues that are conserved.

The DNA fragment corresponding to the MADS box protein was constructed via recursive PCR from oligos that contain the codons optimized for expression in *E. coli*. The DNA fragments corresponding to the remaining portions of TEM-1 and PSE-4 were then generated via PCR as described in Chapter 5. The entire gene was then constructed using the SOEing procedure and was then inserted into the PMON vector. The reconstructed gene was sequenced to confirm that no mutations had been made. The vector was then transformed into XL1-Blue cells and plated on increasing concentrations of ampicillin. No cells grew on plates with 10 μg/ml ampicillin, indicating a complete loss of β-lactamase activity.

While this experiment represents a single unsuccessful recombination attempt, it does expose some of the difficulties underlying the swapping of non-homologous schema. Most importantly, the structural context may dominate the ability for one subunit to be inserted into a different structure. Interesting future experiments include the attempt to evolve the TEM/MADS hybrid to regain stability or activity, the development of

selections for stability, and the testing of adding connecting peptides between schema when recombining very divergent sequences.

Figure F-1:

The schema profile is shown for β-lactamase TEM-1 without any parental sequence identity. The minima of the profile are marked by the dotted lines and the structures of the schema are shown above the graph. The first two and last two schema are combined because the CE search algorithm poorly scores small fragments. The active site residues are S70, E73 (schema 3), and K166 (schema 5).

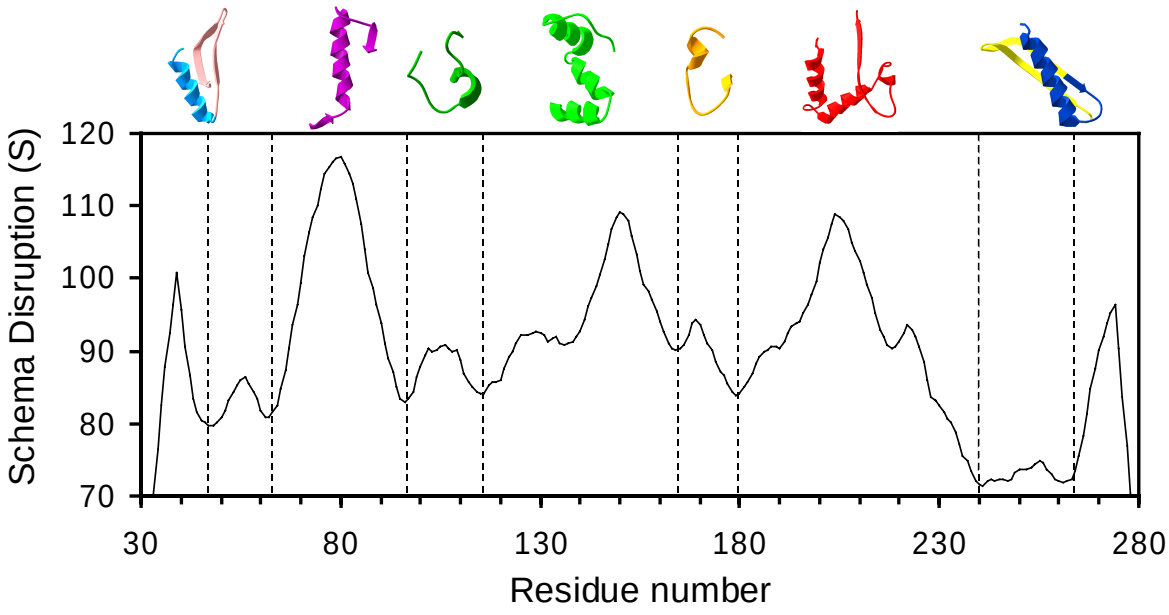

Schema Disruption (S)
120
110
100
90
80
70
30
80
130
180
230
280
Residue number

Figure F-2:

The number of unique matches found in the protein databank, as discovered by the CE algorithm (Shindyalov and Bourne, 1998). All of the β-lactamase variants and redundant structures were removed from this list. Schemas 1+2 and 8+9 are participating in a non-contiguous domain that has been identified previously as a motif in protein structures (Orengo *et al*., 1994; Orengo *et al*., 1997).

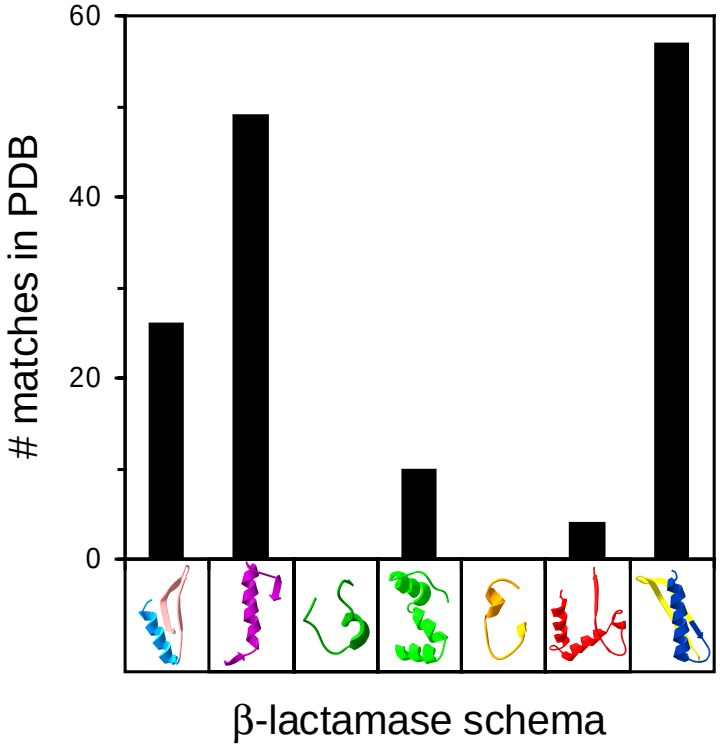
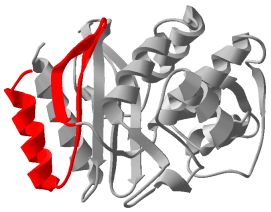
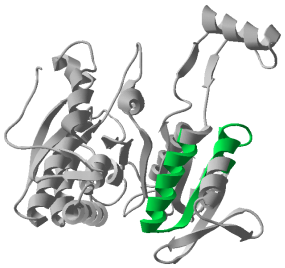

60
40
20
0
matches in PDB
β-lactamase schema

Figure F-3:

Six representative structures found in the protein databank that contain schema 1+2. The structural context of the schema differs greatly between the structures. For example, in β-lactamase, this schema is largely exposed whereas in myosin it is entirely buried. The transcription factor shown is MADS box.

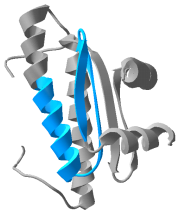


Schema 1+2
β-lactamase TEM

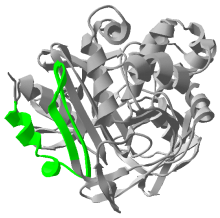


Isocitrate dehydrogenase

Phosphofructokinase

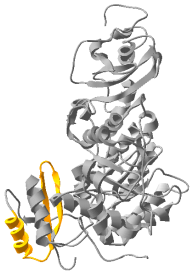


Transcription factor



Peptidase

Myosin motor domain



Rubisco

Figure F-4:

The twelve structures of schema 1+2 with the closest rms to β-lactamase are overlaid. While the backbones of the schema are very similar, the amino acid content of the sequences varies significantly. A concern in shuffling non-homologous schema is whether the end points of the divergent schema will properly connect (dotted circle). Allowing for insertions at the connection points may overcome this problem (O'Maille *et al*., 2002).

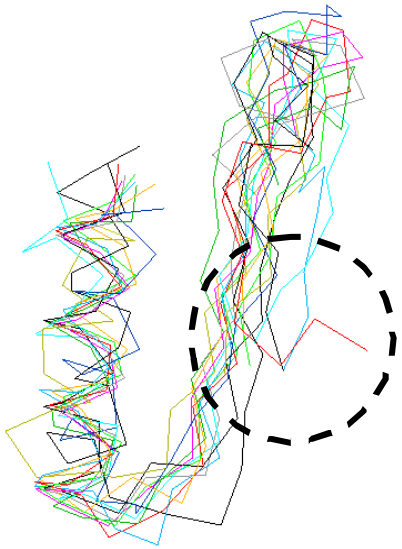
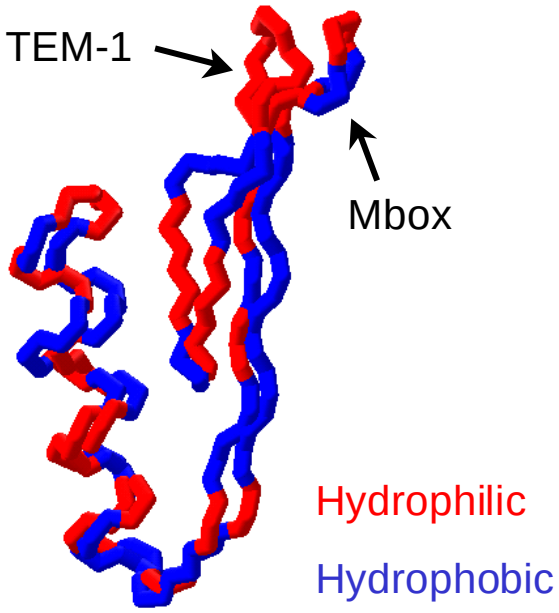


Figure F-5:

(A) The sequence for schema 1+2 of TEM-1 and PSE-4 is compared with the MADS box protein sequence. There is very little amino acid sequence identity shared between MADS box and the β-lactamases (11% with TEM-1 and 6% with PSE-4). To illustrate differences in the binary pattern, the hydrophobic amino acids are shown in blue and the polar residues are shown in red. Each disagreement in the binary pattern is highlighted in yellow. A conserved glycine residue is also shown. **(B)** The binary pattern of TEM-1 and MADS box are compared in the context of the three-dimensional structure.


tem: HPETLVKVKDAEDQLGARVGYIELDLNSGKILESF
Mbox: KFGLMKKAYELSVLCDCEIALIIFNS-SNKLFQYA
pse: FQQVEQDVKAIEVSLSARIGVSVLDTQNGEYW-DY

(A)


TEM-1
Mbox
Hydrophilic
Hydrophobic

(B)

References

Abramson, N. (1963). "Information Theory and Coding," McGraw-Hill, New York.

Adami, C., Ofria, C., and Collier, T. C. (2000). Evolution of biological complexity, *Proc. Natl. Acad. Sci. USA* **97**, 4463-4468.

Aita, T. and Husimi, Y. (1996). Fitness spectrum among random mutants on Mt. Fuji-type fitness landscape. *J. theor. Biol.* **182**, 469-485.

Aita, T. and Husimi, Y. (1998). Adaptive walks by the fittest among finite random mutants on a Mt. Fuji type fitness landscape. *J. theor. Biol.* **193**, 383-405.

Aita, T., and Husimi, Y. (2000). Theory of evolutionary molecular engineering through simultaneous accumulation of advantageous mutations, *J. theor. Biol*. **207**, 543-556.

Albert, R., Jeong, H., and Barabasi, A.-L. (2000). Error and attack tolerance of complex networks, *Nature* **406**, 378-382.

Almassy, R. J., Janson, C. A., Kan, C. C. and Hostomska, Z. (1992). Structures of apo and complexed *Escherichia coli* glycinamide ribonucleotide transformylase. *Proc. Natl. Acad. Sci. USA* **89**, 6114-6118.

Alon, U., Surette, M. G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis, *Nature* **397**, 168-171.

Alves, R., and Savageau, M. A. (2000). Comparing systemic properties of ensembles of biological networks by graphical and statistical methods, *Bioinformatics* **16**, 527-533.

Ancel, L. W., and Fontana, W. (2000). Plasticity, evolvability, and modularity in RNA, *J. Exp. Zoology* **288**, 242-283.

Anderson, P. W. (1983). Suggested model for prebiotic evolution: the use of chaos. *Proc. Natl. Acad. Sci. USA* **80**, 3386-3390.

Arkin, A., Ross, J., and McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage l-infected Escherichia coli cells, *Genetics* **149**, 1633-1648.

Arnold, F. H., editor (2001a). "Evolutionary Protein Design," *Advances in Protein Chemistry* **55**, Academic Press, San Diego, CA.

Arnold, F. H. (2001b). Combinatorial and computational challenges for biocatalyst design, *Nature* **409**, 253-257.

Arnold, F. H. and Wintrode, P. L. (1999). Enzymes, Directed Evolution. In *Encyclopedia of bioprocess technology: fermentation, biocatalysis, and bioseparation* (Flickinger, M. C. and Drew, S. W., eds.), Vol. 2, pp. 971-987. John Wiley & Sons, Inc., New York.

Aronson, H-E. G., Royer, W. E. Jr., Hendrickson, W. A. (1994). Quantification of tertiary structure conservation despite primary sequence drift in the globin fold. *Protein Science* **3**, 1706-1711.

Axe, D. D., Foster, N. W., and Fersht, A. R. (1996). Active barnase variants with completely random hydrophobic cores, *Proc. Natl. Acad. Sci. USA* **93**, 5590-5594.

Baase, W. A., *et al*. (1999). How much sequence variation can the functions of biological molecules tolerate?, In: *Simplicity and Complexity in Proteins and Nucleic Acids*, Edited by: H. Frauenfelder, J. Deisenhofer, and P. G. Wolynes, Dahlem University Press, pp. 297-311.

Babajide, A., Farber, R., Hofacker, I. L., Inman, J., Lapedes, A. S., and Stadler, P. F. (2001). Exploring protein sequence space using knowledge-based potentials. *J. theor. Biol*. **212**, 35-46.

Barkai, N., and Leibler, S. (1997). Robustness in simple biochemical networks, *Nature* **387**, 913-917.

Becskei, A., and Serrano, L. (2000). Engineering stability in gene networks by autoregulation, *Nature* **405**, 590-593.

Berek, C. and Milstein, C. (1987). Mutation drift and repertoire shift in the maturation of the immune response. *Immunol. Rev*. **96**, 23-41.

Berstein, F. C., *et al*. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Betz, A. G., Rada, C., Pannell, R., Milstein, R., and Neuberger, M. S. (1993). Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: Clustering, polarity, and specific hot spots, *Proc. Natl. Acad. Sci. USA* **90**, 2385-2388.

Betzel, C. *et al.* (1992). Crystal-structure of the alkaline proteinase savinase from Bacillus-lentus at 1.4-A resolution. *J. Mol. Biol*. **223**, 427-445.

Bhalla, U. S., and Iyengar, R. (1999). Emergent properties of networks of biological signal pathways, *Science* **283**, 381-387.

Bhat, T. N., Bentley, G. A., Fischmann, T. O., Boulot, G. and Poljak, R. J. (1990). Small rearrangements in structures of Fv and Fab fragments of antibody D1.3 on antigen binding. *Nature* **347**, 483-485.

Bhat, T. N., *et al*. (1994). Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc. Natl. Acad. Sci. USA* **91**, 1089-1093.

Blake, C. C. F. (1979). Exons encode protein functional units, *Nature* **277**, 598-598.

Blanco, F. J., Angrand, I., and Serrano, L. (1999). Exploring the conformational properties of the sequence space between two proteins with different folds: an experimental study. *J. Mol. Biol*. **285**, 741-753.

Boder, E. T., Midelfort, K. S., and Wittrup, K. D. (2000). Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc. Natl. Acad. Sci. USA* **97**, 10701-10705.

Bogarad, L. D. and Deem, M. W. (1999). A hierarchal approach to protein molecular evolution. *Proc. Natl. Acad. Sci. USA* **96**, 2591-2595 (1999).

Bolon, D. N., and Mayo, S. L. (2001). Enzyme-like proteins by computational design, *Proc. Natl. Acad. Sci. USA* **98**, 14274-14279.

Bolon, D. N., Voigt, C. A., and Mayo, S. L., De novo design of biocatalysts, *Curr. Opin. Chem. Biol*. **6**, 125-129 (2002).

Bornberg-Bauer, E., and Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space, *Proc. Natl. Acad. Sci. USA* **96**, 10689-10694.

Bornholdt, S., and Rohlf, T. (2000). Topological evolution of dynamical networks: Global criticality from local dynamics, *Phys. Rev. Lett*. **84**, 6114-6117.

Braden, B. C., *et al*. (1996). Crystal structure of the complex of the variable domain of antibody D1.3 and turkey egg white lysozyme: a novel conformational change in antibody CDR-L3 selects for antigen. *J. Mol. Biol.* **257**, 889-894.

Braden, B. C., Goldman, E. R., Mariuzza, R. A. and Poljak, R. J. (1998). Anatomy of an antibody molecule: structure, kinetics, thermodynamics and mutational studies of the antilysozyme antibody D1.3. *Immunol. Rev.* **163**, 45-57.

Brock, B. J. and Waterman, M. R. (2000). The use of random chimeragenesis to study structure/function properties of rat and human P450c17. *Arch. Biochem. Biophys*. **373**, 401-408.

Broglia, R. A., Tiana, G., Roman, H. E., Vigezzi, E., and Shakhnovich, E. (1999). *Phys. Rev. Lett*. **82**, 4727-4730.

Brown, B. M. and Sauer, R. T. (1999). Tolerance of Arc repressor to multiple-alanine substitutions. *Proc. Natl. Acad. Sci. USA* **96**, 1983-1988.

Brown, M., Rittenberg, M. B., Chen, C., and Roberts, V. A. (1996). Tolerance to single, but not multiple, amino acid replacements in antibody VH CDR2. *J. Immunol*. **156**, 3285-3291.

Bryngelson, J. D. and Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **84**, 7524-7528.

Buchler, N. E. G., and Goldstein, R. A. (1999). Universal correlation between energy gap and foldability for the random energy model and lattice proteins, *J. Chem. Phys*. **111**, 6599-6609.

Burks, E. A., Chen, G., Georgiou, G., and Iverson, B. L. (1997). In vitro scanning saturation mutagenesis of an antibody binding pocket, *Proc. Natl. Acad. Sci. USA* **94**, 412-417.

Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs, *Phys. Rev. Lett*. **85**, 5468-5471.

Campbell, I. D., and Baron, M. (1991). The structure and function of protein modules, *Phil. Trans. R. Soc. Lond. B*. **332**, 165-170.

Carlson, J. M., and Doyle, J. (2000). Highly optimized tolerance: Robustness and design in complex systems, *Phys. Rev. Lett*. **84**, 2529-2532.

Chen, K. and Arnold, F. H. (1993). Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. USA* **90,** 5618-5622.

Chen, G., Dubrawsky, I., Mendez, P., Georgiou, G. and Iverson, B. L. (1999). In vitro scanning saturation mutagenesis of all the specificity determining residues in an antibody binding site. *Protein Engineering* **12**, 349-356.

Chien, N. C., Roberts, V. A., Giusti, A. M., Scharff, M. D. and Getzoff, E. D. (1989). Significant structural and functional change of an antigen-binding site by a distant amino acid substitution: proposal of a structural mechanism. *Proc. Natl. Acad. Sci. USA* **86**, 5532-5536.

Chothia, C., and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in

proteins. *EMBO J*. **5**, 823-826.
Chowdhury, P. S., and Pastan, I. (1999) Improving antibody affinity by mimicking somatic hypermutation in vitro. *Nature Biotech*. **17**, 568-572.
Christmas, P., *et al*. (2001). Alternative splicing determines the function of CYP4F3 by switching substrate specificity, *J. Biol. Chem*. **41**, 38166-38172.
Cohen, R., Erez, K., ben-Avraham, D., and Havlin, S. (2000). Resilience of the internet to random breakdowns, *Phys. Rev. Lett*. **85**, 4626-4628.
Cohen, R., Erez, K., ben-Avraham, D., and Havlin, S. (2001). Breakdown of the internet under intentional attack, *Phys. Rev. Lett*. **86**, 3682-3685.
Cordes, M. H. J., Davidson, A. R. and Sauer, R. T. (1996). Sequence space, folding and protein design. *Current Opinion in Structural Biology* **6**, 3-10.
Cordes, M H. J., Walsh, N. P., McKnight, C. J., and Sauer, R. T. (1999). Evolution of a protein fold in vitro. *Science* **284**, 325-327.
Cowell, L. G., Kim, H-J., Humaljoki, T., Berek, C., and Kepler, T. B. (1999). Enhanced evolvability in Immunoglobin V genes under somatic hypermutation, *J. Mol. Evol*. **49**, 23-26.
Crameri, A., Dawes, G., Rodriguez, E. Jr., Silver, S., and Stemmer, W. P. C. (1997). Molecular evolution of an arsenate detoxification pathway by DNA shuffling, *Nature Biotech*. **15**, 436-438.
Crameri, A., Raillard, S.-A., Bermudez, E. and Stemmer, W. P. C. (1998). DNA shuffling of genes from diverse species accelerates directed evolution. *Nature* **391,** 288-291.
Crippen, G. M. (1978). Tree structural organization of proteins. *J. Mol. Biol.* **126**, 315-332.
Dahiyat, B. I., Gordon, D. B. and Mayo, S. L. (1997a). Automated design of the surface positions of protein helices. *Protein Science* **6**, 1333-1337.
Dahiyat, B. I. and Mayo, S. L. (1996). Protein design automation. *Protein Science* **5**, 895-903.
Dahiyat, B. I. and Mayo, S. L. (1997a). Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **94**, 10172-10177.
Dahiyat, B. I., and Mayo, S. L. (1997b). De novo protein design: Fully automated sequence selection, *Science* **278**, 82-87.
Dahiyat, B. I., Sarisky, C. A. and Mayo, S. L. (1997b). De Novo protein design: towards fully automated sequence selection. *J. Mol. Biol.* **273**, 789-796.
Dall'Acqua, W., Goldman, E. R., Eisenstein, E. and Mariuzza, R. A. (1996). A mutational analysis of the binding of two different proteins to the same antibody. *Biochemistry* **35**, 9667-9676.
Dall'Acqua, W., *et al*. (1998). A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry* **37**, 7981-7991.
Derrida, B. (1981). Random-energy model: An exactly solvable model of disordered systems, *Phys. Rev. B* **24**, 2613-2626.
Daugherty, P. S., Chen, G., Iverson, B. I., and Georgiou, G. (2000). Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies, *Proc. Natl. Acad. Sci. USA* **97**, 2029-2034.
Desjarlais, J. R. and Clarke, N. D. (1998). Computer search algorithms in protein modification and design. *Current Opinion in Structural Biology* **8**, 471-475.
Desjarlais, J. R. and Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Science* **4**, 2006-2018.
De Maeyer, M., Desmet, J. and Lasters, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Folding & Design* **2**, 53-66.
de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W., and Gilbert, W. (1996). Intron positions correlate with module boundaries in ancient proteins, *Proc. Natl. Acad. Sci. USA* **93**, 14632-14636.
de Souza, S. J., Long, M., Klein, R. J., Roy, S., Lin, S., and Gilbert, W. (1998). Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins, *Proc. Natl. Acad. Sci. USA* **95**, 5094-5099.
Desmet, J., De Maeyer, M., Hazes, B., and Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542.
Desmet, J., De Maeyer, M. and Lasters, I. (1994). In *The Protein Folding Problem and Tertiary Structure Prediction* (Jr., K. M. and Grand, S. L., eds.), pp. 307-337. Birkhauser, Boston.
Dunbrack, R. L. and Karplus, M. (1993). Backbone-dependent rotamer library for proteins - application to side-chain prediction. *J. Mol. Biol.* **230,** 543-574.

Dunbrack, R. L. and Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nature Structural Biology* **1,** 334-340.

Edwards, R., and Glass, L. (2000). Combinatorial explosion in model gene networks, *Chaos*, **10**, 691-704.

Eigen, M., and McCaskill, J. (1989). The molecular quasi-species, *Adv. Chem. Phys*. **75**, 149-263.

El Hawrani, A. S., Moreton, K. M., Sessions, R. B., Clarke, A. R., and Holbrook, J. J. (1994). Engineering surface loops of proteins - a preferred strategy for obtaining new enzyme function, *TIBTECH* **12**, 207-211.

Elowitz, M. B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators, *Nature* **403**, 335-338.

England, P., Bregegere, F. and Bedoulle, H. (1997). Energetic and kinetic contributions of contact residues of antibody D1.3 in the interaction with lysozyme. *Biochemistry* **36**, 164-172.

England, P., Nageotte, R., Renard, M., Page, A.-L. and Bedouelle, H. (1999). Functional characterization of the somatic hypermutation process leading to antibody D1.3, a high affinity antibody directed against lysozyme. *J. Immunol.* **162**, 2129-2136.

Fields, B. A., *et al*. (1996). Hydrogen bonding and solvent structure in an antigen-antibody interface. Crystal structures and thermodynamic characterization of three Fv mutants complexed with lysozyme. *Biochemistry* **35**, 15494-15503.

Firn, R. D., and Jones, C. G. (2000). The evaluation of secondary metabolism - a unifying model, *Molecular Microbiology*, **37**, 989-994.

Fisch, I., *et al*. (1996). A strategy of exon shuffling for making large peptide repertoires displayed on filamentous bacteriophage, *Proc. Natl. Acad. Sci. USA* **93**, 7761-7766.

Fischer, K. H. and Hertz, J. A. (1991). "Spin Glasses," Cambridge University Press, Cambridge.

Freire, E. (1999). The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proc. Natl. Acad. Sci. USA* **96**, 10118-10122.

Fontana, W. and Shuster, P. (1987). A computer model of evolutionary optimization. *Biophysical Chemistry* **26**, 123-147.

Fontana, W. and Shuster, P. (1998). Continuity in evolution: On the nature of transitions. *Science* **280,** 1451-1455.

Forrest, S. and Mitchell, M. (1993). *Foundations of Genetic Algorithms 2*, L. D. Whitley (Morgan Kaufmann, San Mateo), p. 109.

Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in *Escherichia coli*, *Nature* **403**, 339-342.

Germain, R. N. (2001). The art of the probable: System control in the adaptive immune system, *Science* **293**, 240-245.

Gilbert, W. (1978). Why genes in pieces?, *Nature* **271**, 501-501.

Gilbert, W., de Souza, S. J. and Long, M. Y. (1997). Origin of Genes. *Proc. Natl. Acad. Sci. USA* **94**, 7698-7703.

Giver, L., Gershenson, A., Freskgard, P.-O. and Arnold, F. H. (1998). Directed evolution of a thermostable esterase. *Proc. Natl. Acad. Sci. USA* **95**, 12809-12813.

Giver, L., Gershenson, A., Freskgard, P.-O. and Arnold, F. H. (1998). Directed evolution of a thermostable esterase. *Proc. Natl. Acad. Sci. USA* **95,** 12809-12813.

Glykos, N. M., Cesareni, G., and Kokkinidis, M. (1999). Protein plasticity to the extreme: changing the topology of a 4-a-helical bundle with a single amino acid substitution. *Structure* **7**, 597-603.

Gō, M. (1981). Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**, 90-92.

Gō, M. (1983). Modular structural units, exons, and function in chicken lysozyme. *Proc. Natl. Acad. Sci. USA* **80**, 1964-1968.

Gō, M. (1985). Protein structures and split genes, *Adv. Biophys.* **19**, 91-131.

Godzik, A. (1995). In search of the ideal protein sequence. *Protein Engineering* **8**, 409-416.

Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses, *Biophys. J*. **66**, 1335-1340.

Gordon, D. B., Marshall, S. A., and Mayo, S. L. (1999) Energy functions for protein design. *Curr. Opin. Struct. Biol*. **9**, 509-513.

Gordon, D. B., Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end-elimination theorem. *J. Comp. Chem*. **19**, 1505-1514.

Gordon, D. B. and Mayo, S. L. (1999). Branch-and-Terminate: a combinatorial optimization algorithm for protein design. *Structure* **7**, 1089-1098.

Govindarajan, S., and Goldstein, R. A. (1997). Evolution of model proteins on a foldability landscape, *Proteins* **29**, 461-466.

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

Hartman, J. L. IV, Garvik, B., and Hartwell, L. (2001). Principles for the buffering of genetic variation, *Science* **291**, 1001-1004.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular biology, *Nature* **402**, C47-C52.

Hawkins, R. E., Russell, S. J., Baier, M. and Winter, G. (1993). The contribution of contact and non-contact residues of antibody in the affinity of binding to antigen. *J. Mol. Biol.* **234**, 958-964.

Hedstrom, L., Szilagyi, L., and Rutter, W. J. (1992). Converting trypsin to chymotrypsin: the role of surface loops, *Science* **255**, 1249-1253.

Hellinga, H. W. and Richards, F. M. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. USA* **91**, 5803-5087.

Hill, T. (1960). "An Introduction to Statistical Thermodynamics," Dover Books, New York.

Horton, R. M. (1995). PCR-mediated recombination and mutagenesis. *Mol. Biotech.* **3**, 93-99.

Holland, J. (1975). "Adaptation in Natural and Artificial Systems," The University of Michigan Press, Ann Arbor, MI.

Holm, L. and Sander, C. (1992). Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins: Struct, Funct, and Gen.* **14**, 213-223.

Holm, L. and Sander, C. (1994). Parser for protein folding units. *Proteins* 19, 256-268.

Hordijk, W. and Manderick, B. (1995). The usefulness of recombination, *Advances in Artificial Life* **929**, 908-919.

Huang, W., and Palzkill, T. (1997). A natural polymorphism in β-lactamase is a global suppressor, *Proc. Natl. Acad. Sci. USA* **94**, 8801-8806.

Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S., and Palzkill, T. (1996). Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.* **258**, 688-703.

Huynen, M. A. (1996). Exploring phenotype space through neutral evolution, *J. Mol. Evol.* **43**, 165-169.

Huynen, M. A., Stadler, P. F., and Fontana, W. (1996). Smoothness within ruggedness: The role of neutrality in adaptation, *Proc. Natl. Acad. Sci. USA* **93**, 397-401.

Inaba, K., Wakasugi, K., Ishimori, K., Konno, T., Kataoka, M., and Morishima, I. (1997). Structural and functional roles of modules in hemoglobin, *J. Biol. Chem.* **272**, 30054-30060.

Ito, W., Sakato, N., Fujio, H., Yutani, K., Arata, Y. and Kurosawa, Y. (1992). The His-probe method: effects of histine residues introduced into the complementary-determining regions of antibodies on antigen-antibody interactions at different pH values. *FEBS* **309**, 11483-11486.

Jain, S. C., Shinde, U., Li, Y., Inouye, M. and Berman, H. M. (1998). The crystal structure of an autoprocessed Ser221Cys-subtilisin E-propeptide complex at 2.0 angstrom resolution. *J. Mol. Biol.* **284,** 137-144.

Jelsch, C., Mourey, L., Masson, J. M., and Samama, J. P. (1993). Crystal structure of *Escherichia Coli* TEM-1 beta-lactamase at 1.8 angstroms resolution. *Proteins* **16**, 364-383.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A-L. (2000). The large-scale organization of metabolic networks, *Nature* **407**, 651-654.

Jermutus, L.,Tessier, M., Pasamontes, L., van Loon, A. P. G. M., and Lehmann, M. (2001). Structure-based chimeric enzymes as an alternative to directed evolution: phytase as a test case, *J. Biotechnology* **85**, 15-24.

Joern, J. M., Meinhold, P., and Arnold, F. H. (2002). Analysis of shuffled gene libraries. *J. Mol. Biol.* **316**, 643-656.

Jones, D. T. (1994). De novo protein design using pairwise potentials and a genetic algorithm. *Protein Science* **3**, 567-574.

Jucovic, M., and Poteete, A. R. (1999). Protein salvage by directed evolution: Functional restoration of a defective lysozyme mutant, *Ann. NY Acad. Sci*., **870**, 404-407.

Jürgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M., and Sterner, R. (2000). Directed

evolution of a $(\beta\alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways, *Proc. Natl. Acad. Sci. USA* **97**, 9925-9930.

Kaneko, S., *et al*. (2000). Module shuffling of a family F/10 xylanase: replacement of modules M4 and M5 of the FXYN of Streptomyces olivaceoviridis E-86 with those of the Cex of Cellomonas fimi, *Protein Engineering* **13**, 873-879.

Kauffman, S. (1993). "The Origins of Order," Oxford University Press, Oxford.

Kauffman, S. and Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. theor. Biol.* **128**, 11-45.

Kauffman, S. A., and Macready, W. G. (1995). Search strategies for applied molecular evolution, *J. theor. Biol*. **173**, 427-440.

Kauffman, S. A. and Weinberger, E. D. (1989). The NK model of rugged fitness landscapes and its application to the maturation of the immune response. *J. theor. Biol*. **141**, 211-245.

Kimura, M. (1983). "The neutral theory of molecular evolution," Cambridge University Press, Cambridge.

Kirschner, M., and Gerhart, J. (1998). Evolvability, *Proc. Natl. Acad. Sci. USA* **95**, 8420-8427.

Kobayashi, H., *et al*. (1999). Probing the interaction between a high-affinity single-chain Fv and a pyrimidine (6-4) pyrimidone photodimer by site-directed mutagenesis. *Biochemistry* **38**, 532-539.

Koehl, P. and Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239,** 249-275.

Koehl, P. and Delarue, M. (1995). A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modeling. *Nature Struct. Biol*. **2**, 163-170.

Koehl, P. and Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Cur. Opin. Struct. Biol.* **6,** 222-226.

Koehl, P., and Levitt, M. (2002). Protein topology and stability define the space of allowed sequences. *Proc. Natl. Acad. Sci. USA* **99**, 1280-1285.

Kolkman, J. A., and Stemmer, W. P. C. (2001). Directed evolution of proteins by exon shuffling, *Nature Biotechnology* **19**, 423-428.

Kono, H., and Saven, J. G. (2001). Statistical theory for protein conformational libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure, *J. Mol. Biol*. **306**, 607-628.

Koradi, R., Billeter, M. and Wuthrich, K. (1996). MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14,** 51-&.

Kumagai, I., Takeda, S., and Miura, K-I. (1992). Functional conversion of the homologous proteins α-lactalbumin and lysozyme by exon exchange, *Proc. Natl. Acad. Sci. USA* **89**, 5887-5891.

Lasters, I., De Maeyer, M. and Desmet, J. (1995). Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Engineering* **8**, 815-822.

Lauffenberger, D. A. (2000). Cell signaling pathways as control modules: Complexity for simplicity? *Proc. Natl. Acad. Sci. USA* **97**, 5031-5033.

Laughton, C. A. (1994). Prediction of protein side-chain conformations from local three-dimensional homology relationships. *J. Mol. Biol*. **235**, 1088-1097.

Lazar, G. A., Desjarlais, J. R. and Handel, T. M. (1997). De novo design of the hydrophobic core of ubiquitin. *Protein Science* **6**, 1167-1178.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility *J. Mol. Biol*. **55**, 379-400.

Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* **236,** 918-939.

Lee, C. (1996). Testing homology modeling on mutant proteins: predicting structural and thermodynamic effects in the Ala98->Val mutants of T4 lysozyme. *Folding & Design* **1**, 1-12.

Lee, C. and Levitt, M. (1991). Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* **352**, 448-451.

Lee, C. and Subbiah S. (1991). Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol*. **217**, 373-388.

Lehman, M., Pasamontes, L., Lassan, S. F., and Wyss, M. (2000). The consensus concept for thermostability engineering of proteins, *Biochemica et Biophysica Acta*, **1543**, 408-415.

Li, H., Helling, R., Tang, C. and Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666-669.

Li, Y., Li, H., Smith-Gill, S. J. and Mariuzza, R. A. (2000). Three-dimensional structures of free and antigen-bound Fab from monoclonal antilyzozyme antibody HyHEL-63. *Biochemistry* **39**, 6296-6309.

Linden, R. H. J. v. d., Geus, B. d., Frenken, L. G. J., Peters, H. and Verrips, C. T. (2000). Improved production and function of llama heavy chain antibody fragments by molecular evolution. *J. Biotech.* **80**, 261-270.

Lim, D. *et al*. (2001). Insights into the molecular basis for carbenicillinase activity of PSE-4 beta-lactamase from crystallographic and kinetic studies. *Biochemistry* **40**, 395-402.

Lo, N. M., Holliger, P., and Winter, G. (1996). Mimicking somatic hypermutation: Affinity maturation of antibodies displayed on bacteriophage using a bacterial mutator strain, *J. Mol. Biol*. **260**, 359-368.

Lobkovsky, E. *et al*. (1993). Evolution of an enzyme-activity - crystallographic structure at 2-angstrom resolution of cephalosporinase from the AmpC gene of Enterobacter-cloacae-P99 and comparison with a class-A penicillinase. *Proc. Natl. Acad. USA* **90**, 11257-11261.

Lockless, S. W., and Ranganathan, R. (1999). Evolutionary conserved pathways of energetic connectivity in protein families, *Science* **286**, 295-299.

Loeb, D. D., Swanstrom, R., Everitt, L. E., Manchester, M., Stamper, S. E., and Hutchson, C. A. III (1989). Complete mutagenesis of the HIV-1 protease, *Nature* **340**, 397-400.

Loughlin, D. H., and Ranjithan, S. (1997). The neighborhood constraint method: A genetic algorithm-based multiobjective optimization technique, In: *Seventh international conference on genetic algorithms*, Michigan State University, Ed: Thomas Bak, pp. 666-673.

Lutz, S., Ostermeier, M. and Benkovic, S. J. (2001). Rapid generation of incremental truncation libraries for protein engineering using α-phosphothioate nucleotides. *Nucl. Acid. Res*. **29**, e16.

Macken, C. A., Hagan, P. S., and Perelson, A. S. (1991). Evolutionary walks on rugged landscapes, *SIAM J. Appl. Math*. **51**, 799-827.

Macken, C. A. and Perelson, A. S. (1989). Protein evolution on rugged landscapes. *Proc. Natl. Acad. Sci. USA* **86**, 6191-6195.

Malakaukas, S. M. and Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein varient. *Nature Structural Biology* **5,** 470-475.

Manderick, B., de Weger, M., and Spiessens, P. (1991). The genetic algorithm and the structure of the fitness landscape, In: Proceedings of the fourth international conference on genetic algorithms, Below, R. K and Booker, L. B. (eds), UC San Diego, pp. 143-148.

Marshall, S. A., and Mayo, S. L. (2001) Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol*. **305**, 619-631.

Matsumura, I., and Ellington, A. D. (2001). In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates, *J. Mol. Biol*. **305**, 331-339.

Matsumura, M., Wozniak, M., Dao-Pin, S. and Matthews, B. W. (1989). Structural studies of T4 lysozyme that alter hydrophobic stabilization. *J. Biol. Chem.* **264,** 16059-16066.

Matsuura, T., Yomo, T., Trakulnaleamsai, S., Ohashi, Y., Yamamoto, K. and Urabe, I. (1998). Nonadditivity of mutational effects on the properties of catalase I and its application to efficient directed evolution. *Protein Engineering* **11**, 789-795.

May, O., Nguyen, P. T., and Arnold, F. H. (2000). Inverting enantioselectivity by directed evolution of hydantoinase for improved production of L-methionine. *Nature Biotech*. **18**, 317-320.

Mayo, S. L., Olafson, B. D. and III, W. A. G. (1990). DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **94,** 8897-8909.

McAdams, H. H., and Arkin, A. (1997). Stochastic mechanisms in gene expression, *Proc. Natl. Acad. Sci. USA* **94**, 814-819.

McAdams, H. H., and Arkin, A. (1999). It's a noisy business! Genetic regulation at the nanomolar scale, *TIG* **15**, 65-69.

McAdams, H. H., and Arkin, A. (2000). Gene regulation: Towards a circuit engineering discipline, *Curr. Biol.* **10**, R318-R320.

Mélin, R., Li, H., Wingreen, N. S., and Tang, C. (1999). Designability, thermodynamic stability, and dynamics in protein folding, *J. Chem. Phys*. **110**, 1252-1262.

Mendes, J., Soares, C. M. and Carrondo, M. A. (1999). Improvement of side-chain modeling in proteins with the self-consistent mean field theory method based on an analysis of the factors influencing prediction. *Biopolymers* **50**, 111-131.

Mendes, J., Baptista, A. M., Carrondo, M. A., *et al*. (1999). Improved modeling of side-chains in proteins

with rotamer-based methods: A flexible rotamer model. *Proteins* **37**, 530-543.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1092.

Mezey, J. G., Cheverud, J. M., and Wagner, G. P. (2000). Is the genotype-phenotype map modular?: A statistical approach using mouse quantitative trait loci data, *Genetics* **156**: 305-311.

Mitchell, M. (1996). "An Introduction to Genetic Algorithms," The MIT Press, Cambridge, MA.

Mitchell, M., Holland, J. H., and Forrest, S. (1994). When will a genetic algorithm outperform hill climbing?, *Advances in Neural Information Processing Systems*, **6**: 102-118.

Miyazaki, K. and Arnold, F. H. (1999). Exploring nonnatural evolutionary pathways by saturation mutagenesis: Rapid improvement of protein function. *J. Molecular Evolution* **49,** 716-720.

Miyazaki, K., Wintrode, P., Grayling, R., Rubingh, D. and Arnold, F.H. (2000). Directed evolution of temperature adaptation in a psychrophilic enzyme. *J. Mol. Biol.* **297**, 1015-1026.

Moore, J. C. and Arnold, F. H. (1996). Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nature Biotechnology* **14**, 458-467.

Moore, J. C., Jin, H.-M., Kuchner, O. and Arnold, F. H. (1997). Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272**, 336-347.

Mühlenbein, H. (1992). How genetic algorithms really work. I. Mutation and hill-climbing. *Parallel Problem Solving in Nature* **2**, 15-25.

Ness, J. E. *et al*. (1999). DNA shuffling of subgenomic sequences of subtilisin. *Nature Biotechnology* **17**, 893-896.

Neuberger, M. S. and Milstein, C. (1995). Somatic hypermutation. *Curr. Opin. Immunol.* **7**, 248-252.

Nikolova, P. V., Henckel, J., Lane, D. P. and Fersht, A. R. (1998). Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proc. Natl. Acad. Sci USA* **95,** 14675-14342.

Novotny, J., Bruccoleri, R. E. and Saul, F. A. (1989). On the attribution of the binding energy in antigen antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry* **28**, 4735-4749.

Nowak, M. A., Boerlijst, M. C., Cooke, J., and Smith J. M. (1997). Evolution of genetic redundancy, *Nature* **388**, 167-171.

O'Maille, P. E., Bakhtina, M., and Tsai, M-D. (2002). Structure-based combinatorial protein engineering (SCOPE). *J. Mol. Biol*., to be published.

Oprea, M., and Kepler, T. B. (1999). Genetic plasticity of V genes under somatic hypermutation: Statistical analysis using a new resampling-based methodology, *Genome Res*. **9**, 1294-1304.

Orencia, C., Hanson, M. A. and Stevens, R. C. (2000). Structural analysis of affinity matured antibodies and laboratory-evolved enzymes. In *Evolutionary Protein Design* (Arnold, F. H., ed.), Vol. 55, pp. 227-259. Academic Press, New York.

Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P. C., and Stevens, R. C. (2001). Predicting the emergence of antibiotic resistance by directed evolution and structural analysis, Nature *Struct. Biol*. **8**, 238-242.

Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature* **372**, 631-634.

Orengo, C. A., Michie, A. D., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH – a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108.

Ostermeier, M. and Benkovic, S. J. (2000). Evolution of protein function by domain swapping. *Advances in Protein Chemistry* **55**, 29-77.

Ostermeier, M., Shim, J. H., and Benkovic, S. J. (1999). A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotech*. **17**, 1205-1209.

Palzkill, T., and Botstein, D. (1992) Probing β-lactamase structure and function using random replacement mutagenesis. *Proteins* **14**, 19-44.

Palzkill T., Le Q-Q., Venkatachalam, K. V., LaRocco, M., and Ocera, H. (1994). Evolution of antibiotic resistance: several different amino acid substitutions in an active site loop alter the substrate profile of β-lactamase, *Mol. Microbiology* **12**, 217-229.

Panchenko, A. R., Luthey-Schulten, Z., and Wolynes, P. G., (1996) Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci. USA* **93**, 2008-2013.

Parisi, G., and Enchave, J. (2001). Structural constraints and emergence of sequence patterns in protein evolution, *Mol. Biol. Evol*. **18**, 750-756.

Patel, P. H., and Loeb, L. A. (2000). DNA polymerase active site is highly mutable: evolutionary consequences, *Proc. Natl. Acad. Sci. USA* **97**, 5095-5100.

Patten, P. A., *et al*. (1996). The immunological evolution of catalysis. *Science* **271**, 1086.

Pedersen, J. T. and Moult, J. (1996). Genetic algorithms for protein structure prediction. *Current Opinion in Structural Biology* **6**, 227-231.

Petit, A., Maveyraud, L., Lenfant, F., Samama, J-P., Labia, R., and Masson, J-M. (1995). Multiple substitutions at position 104 of β-lactamase TEM-1: assessing the role of this residue in substrate specificity, *Biochem J.* **305**, 33-40.

Petrosino, J., Cantu, C. III, and Palzkill, T. (1998). β-lactamases: protein evolution in real time, *Trends in Microbiology* **6**, 323-327.

Petrounia, I. P., and Arnold, F. H. (2000). Designed evolution of enzymatic properties. *Curr. Opin. Biotech.* **11**, 325-330.

Pierce, N. A., Spriet, J. A., Desmet, J., and Mayo, S. L. (2000). Conformational splitting: A more powerful criterion for dead-end elimination, *J. Comp. Chem.* **21**, 999-1009.

Pjura, P., Matsumura, M., Baase, W. A. and Matthews, B. W. (1993). Development of an in vivo method to identify mutants of phage T4 lysozyme of enhanced thermostability. *Protein Science* **2,** 2217-2225.

Ponder, J. W. and Richards, F. M. (1987). Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791.

Pons, J., Rajpal, A. and Kirsch, J. F. (1999). Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the HyHEL-10/lysozyme interaction. *Protein Science* **8**, 958-968.

Prügel-Bennett, A. and Shapiro, J. L. (1994). Analysis of genetic algorithms using statistical mechanics. *Phys. Rev. Lett.* **72**, 1305-1309.

Raillard, S., *et al*. (2001). Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes, *Chemistry & Biology* **8**, 891-898.

Ranganathan, A. *et al*. (1999). Knowledge-based design of bimodular and trimodular polyketide synthases based on domain and module swaps: a route to simple statin analogues. *Chem. Biol.* **6**, 731-741.

Reidhaar-Olson, J. F. and Sauer, R. T. (1988). Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science* **241**, 53-57.

Reidhaar-Olson, J. F. and Sauer, R. T. (1990). Functionally acceptable substitutions in two alpha-helical regions of lambda-repressor. *Proteins: Struct, Funct, and Gen.* **7**, 306-316.

Rennell, D., Bouvier, S. E., Hardy, L. W., and Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol*. **222**, 67-87.

Riechmann, L. and Winter, G. (2000). Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc. Natl. Acad. Sci. USA* **97**, 10068-10073.

Roberts, R. W., and Szostak, J. W. (1997). RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. USA* **94**, 12297-12302.

Rose, G. D. (1979). Hierarchic organization of domains in globular-proteins *J. Mol. Biol.* **134**, 447-470.

Rose, G. D., and Creamer, T. P. (1994). Protein-folding - predicting predicting. *Proteins* **19**, 1-3.

Rossman, M. G. and Liljas, A. (1974). Recognition of structural domains in globular proteins. *J. Mol. Biol.* **85**, 177-181.

Sánchez, R. and Šali, A. (1997). Comparative protein structure modeling as an optimization problem. *J. Mol. Struct.* **398-399**, 489-496.

Sandberg, W. S., and Terwilliger, T. C. (1993). Engineering multiple properties of a protein by combinatorial mutagenesis, *Proc. Natl. Acad. Sci. USA*, **90** 8367-8371.

Sanschagrin, F., Theriault, E., Sabbagh, Y., Voyer, N. and Levesque, R. C. (2000). Combinatorial biochemistry and shuffling TEM, SHV and Streptomyces albus omega loops in PSE-4 class A beta-lactamase. *J. Antimicrob. Chemo.* **45**, 517-519.

Sasai, M. (1995). Conformation, energy, and folding stability of selected amino acid sequences. *Proc. Natl. Acad. Sci. USA* **92**, 8438-8442.

Savageau, M. (1971). Parameter sensitivity as a criterion for evaluating and comparing the performance of biological systems, *Nature* **229**, 542-544.

Savageau, M. (1972). The behavior of intact biochemical systems, *Curr. Top. Cell. Regul.* **6**, 63-130.

Savageau, M. (1974). Comparison of classical and autogenous systems of regulation in inducible

operons, *Nature* **252**, 546-549.
Savageau, M. (2001). Design principles for elementary gene circuits: Elements, methods, and examples, *Chaos*, **11**, 142-159.
Saven, J. G. and Wolynes, P. G. (1997). Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J. Phys. Chem. B* **101**, 8375-8389.
Schaffer, J. D., and Eshelman, L. J. (1991). On crossover as an evolutionarily viable strategy, In: Proceedings of the fourth international conference on genetic algorithms, Below, R. K and Booker, L. B. (eds), UC San Diego, pp. 63-67.
Schaffer, J. D., and Morishima, A. (1987). An adaptive crossover distribution mechanism for genetic algorithms, *Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms*, pp. 36-40.
Schiffer, C. A., Caldwell, J. W., Kollman, P. A., and Stroud R. M. (1990). Prediction of homologous protein structures based on conformational searches and energetics. *Proteins* **8**, 30-43.
Schmidt-Dannert, C., Umeno, D., and Arnold, F. H. (2000). Molecular breeding of carotenoid biosynthetic pathways, *Nature Biotechnology* **18**, 750-753.
Schultz, P. G., and Lerner, R. A. (1995). From molecular diversity to catalysis: lessons from the immune system, *Science* **269**, 1835-1842.
Shakhnovich, E. I. (1994). Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* **72**, 3907-3910.
Shannon, C. E. (1951). Prediction and entropy of printed English, *Bell System Technical Journal* **27**, 379-343.
Shao, Z., and Arnold, F. H. (1996). Engineering new functions and altering existing functions, *Curr. Opin. Struct. Biol*. **6**, 513-518.
Sharon, J. (1990). Structural correlates of high antibody affinity: three engineered amino acid substitutions can increase the affinity of an anti-p-azophenylarsonate antibody 200-fold. *Proc. Natl. Acad. Sci. USA* **87**, 4814-4817.
Sharon, J., Gefter, M. L., Wyosocki, L. J., and Margolies, M. N. (1989). Recurrent somatic mutations in mouse antibodies to p-azophenylarsonate increase affinity for hapten, *J. Immunol*. **142**, 596-601.
Sherrington, D. and Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Phys. Rev. Lett.* **35**, 1792-1795.
Shindyalov, I. N., and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* **11**, 739-747.
Shoichet, B. K., Baase, W. A., Kuroki, R. and Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. USA* **92**, 452-456.
Sideraki, V., Huang, W., Palzkill, T., and Gilbert, H. F. (2001). A secondary drug resistance mutation of TEM-1 β-lactamase that suppresses misfolding and aggregation, *Proc. Natl. Acad. Sci. USA* **98**, 283-288.
Sieber, V., Martinez, C. A., and Arnold, F. H. (2001). Libraries of hybrid proteins from distantly related sequences. *Nature Biotech*. **19**, 456-460.
Siezen, R. and Leunissen, J. A. M. (1997). Subtilases: The superfamily of subtilisin-like serine proteases. *Protein Science* **6,** 501-523.
Sinha, N., and Nussinov, R. (2001). Point mutations and sequence variability in proteins: Redistributions of preexisting populations, *Proc. Natl. Acad. Sci. USA* **98**, 3139-3144.
Skandalis, A., Encell, L. P. and Loeb, L. A. (1997). Creating novel enzymes by applied molecular evolution. *Chemistry & Biology* **4**, 889-898.
Skinner, M. M., and Terwilliger, T. C. (1996). Potential use of additivity of mutational effects in simplifying protein engineering, *Proc. Natl. Acad. Sci. USA* **93**, 10753-10757.
Smith, J. M. (1970). Natural selection and the concept of a protein space. *Nature* **225,** 563-564.
Soderlind, E., *et al*. (2000). Recombining germline-derived CDR sequences for creating diverse singe-framework antibody libraries, *Nature Biotechnology* **19**, 852-856.
Spears, W. M., and De Jong, K. A. (1991). An analysis of multi-point crossover. *Foundations of Genetic Algorithms* **1**, 301-315.
Spiller, B., Gershenson, A., Arnold, F. H. and Stevens, R. C. (1999). A structural view of evolutionary divergence. *Proc. Natl. Acad. Sci. USA* **96**, 12305-12310.
Spinelli, S. and Alzari, P. M. (1994). Structural implications of somatic mutations during the immune response to 2-phenyloxazolone. *Res. Immunol.* 41-45.
Stemmer, W.P.C (1994). Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389-391.

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

Street, A. G. and Mayo, S. L. (1999). Computational protein design. *Structure* **7,** R105-R109.

Su, A. and Mayo, S. L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Science* **6**, 1701-1707.

Suzuki, M., Christians, F. C., Kim, B., Skandalis, A., Black, M. E., and Loeb, L. A. (1996). Tolerance of different proteins for amino acid diversity, *Molecular Diversity* **2**, 111-118.

Taverna, D. M., and Goldstein, R. A. (2000). The distribution of structures in evolving protein populations. *Biopolymers* **53**, 1-8.

Thayer, A. M. (2001). Biocatalysis. *Chemical & Engineering News*, **May 21**, 27-34.

Tsai, C.-J., Maizel, J. V. and Nussinov, R. (2000). Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc. Natl. Acad. Sci. USA* **97**, 12038-12043.

Tuffery P., Etchebest C., Hazout S., and Lavery R. (1991). A new approach to the rapid determination of protein side-chain conformations. *J. Biomol. Struct. & Dynam*. **8**, 1267-1289.

van Nimwegen, E. (1999). The statistical dynamics of epochal evolution, Thesis, Santa Fe Insitute.

Vásquez, M. (1995). An evaluation of discrete and continuum search techniques for conformational analysis of side chains in proteins. *Biopolymers* **36**, 53-70.

Voigt, C. A., Gordon, B., and Mayo, S. L. (1999). Trading accuracy for speed: comparison of search algorithms in protein design, *J. Mol. Biol*. **299**, 789-803.

Voigt, C. A., Kauffman, S., and Wang, Z-G. (2001a). Rational evolutionary design: the theory of in vitro protein evolution, *Adv. Prot. Chem*. **55**, 79-160.

Voigt, C. A., Martinez, C., Wang, Z-G., Mayo, S. L., and Arnold, F. H. (2002). Protein building blocks preserved by recombination, *Nature Structural Biology* **9**, 553-558.

Voigt, C. A., Mayo, S. L., Arnold, F. H., and Wang, Z-G. (2001b). Computational method to reduce the search space for directed protein evolution. *Proc. Natl. Acad. Sci. USA* **98**, 3778-3783.

Voigt, C. A., Mayo, S. L., Arnold, F. H., and Wang, Z-G. (2001c). Computationally focusing the directed evolution of proteins, *J. Cell. Biol*. **37**, 58-63.

von Dassow, G., Meir, E., Munro, E. M., and Odell, G. M. (2000). The segment polarity network is a robust developmental module, *Nature* **406**, 188-192.

Wagner, A. (2000). Robustness against mutations in genetic networks of yeast, *Nature Genetics* **24**, 355-361.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, *Mol. Biol. Evol*. **18**, 1283-1292.

Wagner, G., and Altenberg, L. (1996). Complex adaptations and the evolution of evolvability, *Evolution* **50**, 967-976.

Wagner, G., Chiu, C-H., and Hansen, T. F. (1999). Is Hsp90 a regulator of evolvability? *J. Exp. Zoology* **285**, 116-118.

Wakasugi, K., Ishimori, K., Imai, K., Wada, Y., and Morishima, I. (1994). "Module" substitution in hemoglobin subunits, *J. Biol. Chem*. **269**, 18750-18756.

Wang, T., Miller, J., Wingreen, N. S., Tang, C., and Dill, K. A. (2000). Symmetry and designability for lattice protein models, *J. Chem. Phys*. **113**, 8329-8336.

Weinberger, E. (1990). Correlated and uncorrelated fitness landscapes and how to tell the difference, *Biol. Cybern*., **63**, 325-336.

Wells, J. A. (1990). Additivity of mutational effects in proteins. *Biochemistry* **29**, 8509-8517

Whitlow, M., Howard, A. J., Wood, J. F., Voss, E. W., and Hardman, K. D. (1995). 1.85-angstrom structure of antifluorescein 4-4-20-FAB. *Protein Engineering* **8**, 749-761.

Wilke, C. O., and Adami, C. (2001). Interaction between directional epistasis and average mutational effects, *Proc. R. Soc. Lond. B* **268**, 1469-1474.

Williams, P. A., Cosme, J., Sridhar, V., Johnson, E. F. and Mcree, D. E. (2000). Mammalian microsomal cytochrome P450 monooxygenase: Structural adaptations for membrane binding and functional diversity. *Mol. Cell*. **93**, 121-131.

Xu, J., Deng, Q., Chen, J., Houk, K. N., Bartek, J., Hilvert, D., and Wilson, I. A. (1999). Evolution of shape complementarity and catalytic efficiency from a primordial antibody template, *Science* **286**, 2345-2348.

Yaoi, T., Miyazaki, K., Oshima, T., Komukai, Y., and Gō, M. (1996). Conversion of the coenzyme

specificity of isocitrate dehydrogenase by module replacement, *J. Biochem* **119**, 1014-1018.
Yasui, H., Ito, W. and Kurosawa, Y. (1994). Effects of substitutions of amino acids on the thermal stability of the Fv fragments of antibodies. *FEBS lett.* **353**, 143-146.
Yi, T-M., Huang, Y., Simon, M. I., and Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control, *Proc. Natl. Acad. Sci. USA* **9**, 4649-4653.
You, L. and Arnold, F. H. (1996). Directed evolution of a subtilisin E in Bacillus subtilis to enhance total activity in aqueous dimethylformamide. *Protein Engineering* **9,** 77-83 (corrigendum in Protein Engineering vol. 9, no. 8, p. 719, 1996).
Yuh, C. H., Bolouri, H., and Davidson, E. H. (1998). Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene, *Science* **279**, 1896-1902.
Zehfus, M. H. and Rose, G. D. (1986). Compact domains in proteins. *Biochemistry* **25**, 5759-5765.
Zhang, Y. X., Perry, K, Vinci, V. A., Powell, K., Stemmer, W. P. C., and del Cardayre, S. B. (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* **415**, 644-646.
Zhao, H. and Arnold, F. H. (1999). Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Engineering* **12,** 47-53.