Characterization of the Functional Diversity and Evolvability of Chimeric Enzymes Assembled by Structure-Guided SCHEMA Recombination

Thesis by

Martina N. Carbone

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2010

(Defended May 18, 2010)

Acknowledgments

I would like to begin by thanking my advisor, Dr. Frances Arnold, for encouraging me to pursue my ideas, for her guidance, and for her support.

Throughout my graduate career I received much help from graduate students and postdoctoral fellows in my laboratory. They laid the foundations of much of the work presented in my thesis and none of it would have been possible without their help and advice. I would like to thank Marco Landwehr, Yougen Li, and Chris Otey for my early work with the P450 enzymes (presented in the first chapter of this thesis). This work would not have been possible without Chris Otey who constructed the SCHEMA library of P450s, Yougen Li who reconstituted the monooxygenase P450s from their heme and reducatse domains, and Marco Landwehr who purified the constructs and performed all of the activity assays with me. Midway through my graduate career I began working on a theoretical project on protein evolution (presented in the second chapter of this thesis). I would like to thank Jesse Bloom and Phil Romero for providing me with lattice protein Python scripts to test my ideas. Late in my Ph.D. experience I resumed experimental work to test the theoretical ideas I had developed. This time I worked with cellulases for which I must thank Pete Heinzelman who assembled the SCHEMA cellulase library and provided me with all the genes I needed for my experiments. I would like to give special thanks to Indira Wu who, when I first started the project, guided me through much of the experimental protocols she developed for expressing, screening, and purifying cellulases. I must also address another special thanks to Sabine Bastian who taught me everything I know about manipulating DNA and molecular biology techniques in general.

Recently I have begun another theoretical project addressing the contribution of recombination to adaptation on NK landscapes. This is extremely recent work that I will not be able to include in my thesis although I may present some relevant results during my defense. For this work I would like to thank my committee member, Dr. Chris Adami, and his graduate student, Bjorn Ostman from the Keck Graduate Institute, for

their interest and advice. I must also thank Chris Snow, who helped me translate a script I had written in Matlab to the much faster Python.

Indira Wu, Phil Romero, and Sabine Bastian have kindly accepted and even offered to review my dissertation and provide comments. I would like to thank them for their generosity and valuable insight.

I would also like to thank the remaining members of my committee, Dr. David Tirrell and Dr. Mark Davis for agreeing to be on my committee and attending to all the responsibilities of this assignment.

To conclude I would like to thank my boyfriend Laurent Mathevet. We met early in my third year of my Ph.D. career and since then he has been a source of endless love, support, and encouragement.

Abstract

In nature proteins evolve by a combination of point mutagenesis and recombination. This process has generated hundreds of fascinating and structurally complex protein folds capable of performing a myriad of important and diverse biochemical functions. This has inspired protein engineers to mimic natural protein evolution in the library to construct synthetic proteins with new or improved properties. Here I show that homologous protein recombination can be used in the library to engineer novel enzymes with new catalytic activities and altered substrate specificities. I also propose that homologous recombination can be used in the laboratory to overcome the challenge of improving the native activities of wild-type proteins. In nature recombination may have helped proteins escape local maxima of the fitness landscape by introducing many homologous mutations to which proteins are highly tolerant. Protein engineers can possibly use it for the same purpose. I validate this hypothesis computationally with highly simplified protein models, and I attempt an experimental verification of this theory with cellulases.

Table of Contents

Introduction

The most intricate human-designed machines pale in comparison to the complexity and stunning functionality of the proteins created by evolution. The bewildering complexity of how protein primary sequences encode these remarkable functions, such as catalyzing in a few seconds chemical reactions that would otherwise take millions of years, reveals the extraordinary capability of natural evolution to seek out protein sequences encoding highly functional molecules from an immense sequence space representing mostly unfolded and dysfunctional proteins.

Natural evolution moves about protein sequence space by single mutational steps and by long jumps spanning many mutations via homologous or nonhomologous recombination. Evidence of the effectiveness of these mutational moves is abundant and ubiquitous in nature: hundreds of different protein folds accounting for an innumerable number of biochemical functions makes up much of our living world. The trophies of natural protein evolution have inspired engineers to borrow nature's algorithm to create new proteins with novel or improved properties. Here, I focus on homologous recombination, and I show that it can be exploited in the laboratory to engineer enzymes with novel activities and specificities. Also, I propose that the products of homologous recombination, chimeras, may be more evolvable than their parents with respect to the native activity because their mutational neighborhood has never been searched by evolutionary processes.

Homologous recombination distinguishes itself from other protein engineering strategies (such as point mutagenesis) in that it explores distant regions of sequence space: proteins that differ in many tens or even hundreds of amino acids from known proteins yet still fold and function can be constructed. Drummond et al. [1] compared random mutation to recombination, investigating how the probability of retaining fold (or parental function) depends on the number of mutations introduced. Random mutations cause a steep, exponential decay in this probability: as is well appreciated by protein engineers and protein scientists, most mutations are deleterious. As chimeras migrate

from one functional native sequence to the next, however, the likelihood of preserving structure or function follows a parabolic curve whose initial slope is much less steep (Figure 1). With data from chimeric and randomly-mutated ß-lactamases, Drummond et al. showed that recombination is much more conservative than random mutation, leading to a folding probability that is many orders of magnitude greater at the highest mutation levels. By exploiting the conservative nature of mutations introduced into a structure that has already proven to tolerate them, recombination creates chimeric enzymes that are distant from one another in sequence with minimal loss in their probability of folding.

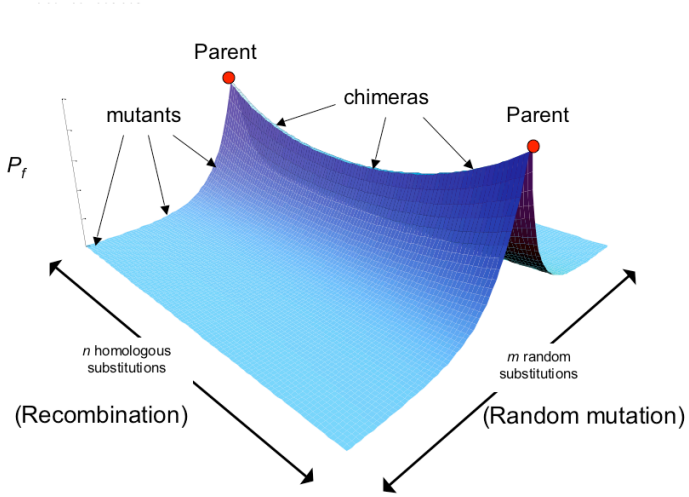


Figure 1: Chimeras occupy a functionally enriched ridge in sequence space. Surface height represents the probability of retaining fold as a function of random and homologous substitutions. Substituting amino acids that already exist in a homologous protein is much more conservative of structure and function than random substitutions. Figure reproduced from [1].

In chapter 1 I describe the first attempt to characterize the diversification of catalytic function within a library of SCHEMA [2] chimeras. SCHEMA is a structure-guided recombination algorithm that selects the crossover locations that maximize the average sequence diversity in the library while minimizing structural disruptions (please refer to *SCHEMA Background* at the end of this section for a brief overview of how

SCHEMA works). Using structure-guided SCHEMA recombination, Otey et al. [3] partitioned the heme domains of cytochrome P450BM3 (CYP102A1) and homologs sharing 61-64% sequence identity into eight blocks and recombined those to make thousands of chimeric P450s. About 47% of the library encodes a properly folded P450, and of those more than 75% are functional. Functional chimeras differ from any known parent by up to 101 amino acid mutations (out of 466).

The inspiration to use homologous recombination to discover new physical and enzymatic properties comes from the observation that proteins with identical folds can diverge greatly not only in sequence, but also in function. The P450 scaffold represents an excellent system to begin this characterization because they comprise a large family of enzymes known to exhibit great diversity at the sequence and functional level. Thousands of P450 sequences exhibiting nearly identical folds and often only 15-20% sequence identity have been reported. They are known to accept many structurally diverse substrates ranging from flexible linear chain molecules like fatty acids to rigid planar molecules like testosterone. P450s are thus naturally malleable to both sequence and functional alterations.

The high sequence diversity among the folded members of the P450 SCHEMA library made this an excellent system to begin probing the functional diversity accessible by recombination. We measured the ability of the parents and fourteen chimeric P450s to hydroxylate a set of eleven substrates, including four human drugs. In chapter 1 I show that the best enzyme on each compound was always a chimera, and some chimeras accepted substrates not accepted by any of the parents. P450s play a major role in drug metabolism and are known to bind and hydroxylate the majority of the drugs we intake. Soluble, bacterial P450 chimeras that can produce drug metabolites may be useful for drug metabolic profiling and lead diversification.

In chapter 2 I present a theory that proposes that chimeras can be expected to be more evolvable than their wild-type parents with respect to the native activity. Improving the native activity of wild-type enzymes is a difficult problem to tackle by directed evolution because the mutational neighborhood of native proteins has already been searched by natural evolution. I propose that chimeras have access to a greater number of beneficial mutations than their native parents because their mutational neighborhood is

unexplored. This argument trivially holds true for chimeras that are less fit than their parents but should also hold for chimeras that are as fit as their parents. Since homologous recombination can introduce many mutations without disrupting folding and function, chimeragenesis may help resolve the problem of improving native activities. The underlying assumption of this theory is that the constraints that prevent the improvement of native activities are evolutionary rather than biophysical or biochemical (i.e., native enzymes are locally rather than globally optimized). The hypothesis is that chimeragenesis provides a means of escape from these local optima and gives chimeras access to beneficial mutations that are not accessible to their wild-type counterparts. I validate this theory in the context of lattice proteins which are highly simplified models of a protein consisting of a chain of 20 monomers on a two-dimensional lattice, and I discuss the requirements that must be satisfied for these results to hold true in the context of real enzymes.

In chapter 3 I test the theory of chapter 2 on real cellulase chimeras assembled by SCHEMA recombination of the catalytic domains of Cel6A from *Trichoderma reesei* and its homologs from *Humicola insolens* and *Chaetomium thermophilum*. Cellulases represent a good system to begin testing this theory because while it is extremely desirable to improve their native cellulolytic activity, to date, no one has reported significant enhancements of their specific activity suggesting that the mutational neighborhood of these enzymes does not contain beneficial mutations. Furthermore, the existence of other glycoside hydrolases performing similar chemistry but exhibiting k_{cat} values that are several orders of magnitude greater than those of cellulases suggests that these enzymes may be locally rather than globally optimized. The SCHEMA library represents a great opportunity to test the theory of chapter 2 because it contains many members that are heavily mutated and yet retain wild-type activities.

The mutational neighborhood of several cellulase chimeras was explored by random point mutagenesis to determine whether beneficial mutations not accessible to their parents could be found. Unfortunately only weakly beneficial mutations representing specific activity improvements comparable to those already reported in the literature were found. The lack of beneficial mutations in the mutational neighborhood of the selected chimeras my reflect either 1) an unlucky choice of chimeras, 2) a high degree

of amino acid conservation in the functionally important regions of the parental enzymes, 3) a low frequency of beneficial mutations in the entirety of sequence space, and 4) a physical limitation to further improvements (i.e., the native enzymes are globally optimized). The various scenarios are discussed in more detail in the discussion of chapter 3.

SCHEMA Background

Proteins are naturally robust to homologous mutations (Figure 1). Computational methods that exploit structural information can be used to further increase the probability that homologous mutations are tolerated and thus optimize the design of recombination libraries. This is effectively equivalent to raising the ridge connecting the two parents of Figure 1. These algorithms generally aim to simultaneously maximize the sequence diversity and the structural integrity of chimeric proteins.

My work is based on chimeras designed using the structure-guided recombination algorithm, SCHEMA [2]. SCHEMA is an algorithm that scores chimeras based on the assumption that nonnative contacts are, on average, deleterious to structure and function. The SCHEMA score, E, is thus equal to the number of nonnative contacts in any given chimera. The algorithm uses the high-resolution crystal structure of one parent to identify all amino acid pair-wise contacts as defined by a 4.5 Å structural cutoff. For any given chimera, the algorithm assumes that its three-dimensional structure will be identical to that of the parent and counts the number of non-native contacts. Nonnative contacts can only form when the two contacting residues are inherited from different parents and when both residues are not perfectly conserved among the parents as shown in Figure 2. An optimization algorithm, RASPP, then directs crossovers to locations that minimize the average disruption in the library while maintaining high diversity [4]. The crossover locations are fixed, such that there exist $3^8 = 6{,}561$ possible sequences in a design based on three parents and seven crossovers. According to this framework, the interfaces between the recombination fragments are composed primarily of conserved residues. Meyer et al. [5] showed that among chimeras with similar numbers of mutations, those

with lower SCHEMA scores are more likely to function validating the physical significance of the scoring method.

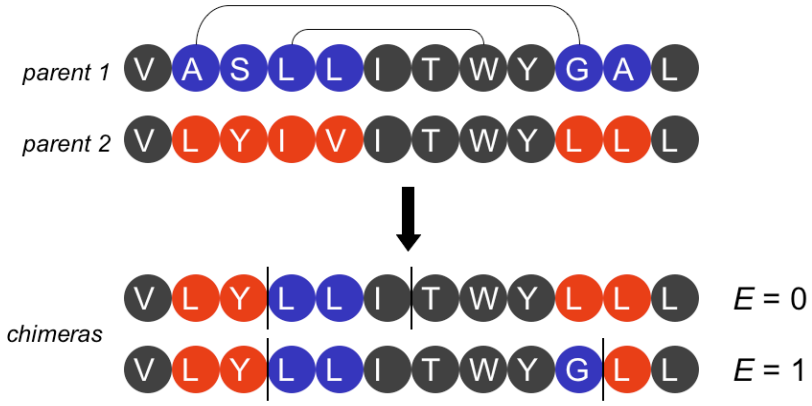


Figure 2: Demonstration of how SCHEMA scores chimeras in a hypothetical 12-residue peptide. Grey residues are conserved in the parents, blue and red residues represent non-conserved amino acids in parents 1 and 2 respectively. Contacting residues (based on the 4.5 Å cutoff of the hypothetical crystal structure of the peptides) are connected by a black solid line shown only in parent 1. Crossovers locations are marked by a short black segment in the chimeras. The SCHEMA score, E, is equal to the number of nonnative contacts in each chimera.

1 Diversification of Catalytic Function in a Synthetic Family of Cytochrome P450s

A version of the chapter has been published in [6]

1.1 Abstract

We report initial characterization of a synthetic family of more than 3,000 cytochrome P450s made by SCHEMA recombination of three bacterial CYP102s. Sixteen heme domains and their holoenzyme fusions with each of the three parental reductase domains were tested for activity on eleven different substrates. The results show that the chimeric enzymes have acquired significant functional diversity, including the ability to accept substrates not accepted by the parent enzymes. K-means clustering analysis of the activity data allowed the enzymes to be classified into five distinct groups based on substrate specificity. The substrates can also be grouped, such that one can be a 'surrogate' for others in the group. Fusion of a functional chimeric heme domain with a parental reductase domain always reconstituted a functional holoenzyme, indicating that key interdomain interactions are conserved upon reductase swapping.

1.2 Introduction

Enzymes with altered activities and specificities can be generated in the laboratory by processes that mimic mechanisms of natural evolution. Directed evolution combining recombination and random point mutation (e.g. DNA shuffling) is effective in generating both genotypic and phenotypic novelty [7-13]. Although recombination can make many mutations with relatively little structural disruption [14], we do not know the degree of functional diversity that is accessible to a process that only explores combinations of mutations already accepted during natural evolution.

We recently reported construction of a synthetic family of more than 3,000 properly folded cytochrome P450 heme domains [15]. Assembled by structure-guided

recombination of the heme domains of CYP102A1 from *Bacillus megaterium* (A1) and its homologs CYP102A2 (A2) and CYP102A3 (A3) that exhibit ~65% amino acid identity, the chimeric proteins differ from the parent sequences by 72 out of 463–466 amino acids on average. Our current goal is to understand how this sequence diversification relates to diversification of function. Initial studies [15,16] demonstrated that recombination, in the absence of point mutations, can generate functional features outside the range exhibited by the parental P450s. For example, a chimeric heme domain significantly more thermostable than any of the parents was identified (T_{50} = 62°C versus 55°C for the most stable parent) [15]; subsequent analysis of more than 200 chimeric heme domains identified many thermostable proteins [17]. Our previous study of selected chimeras of the A1 and A2 heme domains showed that chimeragenesis could also generate activities not exhibited by the parents [16], as has also been reported for recombination of mammalian P450s [18,19].

The biological functions of cytochrome P450s include key roles in drug metabolism, breakdown of xenobiotics, and steroid and secondary metabolite biosynthesis [20]; members of the P450 superfamily catalyze hydroxylation and demethylation reactions on a vast array of substrates [21]. Enzymes from the synthetic P450 family could be useful catalysts for synthesis of biologically-active compounds if they have acquired the ability to accept substrates not accepted by the parent enzymes (which are all fatty acid hydroxylases). Identifying particular desired products, however, usually requires protein purification and HPLC and/or MS analysis, methods that are cumbersome when testing hundreds of biocatalysts. Thus, in addition to exploring the range of catalytic activities in the chimeric P450 family, a second goal of the current study is to determine to what extent 'surrogate' substrates can be used to identify likely catalyst candidates for a particular reaction in a high-throughput screening mode. Can substrates be grouped in such a way that activity towards one member of a group can be used to predict activity towards another?

Enzymes of the CYP102 family are comprised of a reductase domain and a heme domain connected by a flexible linker [22,23]. With a single amino acid substitution (F87A in A1 and F88A in A2 and A3), the heme domains can function alone as peroxygenases, catalyzing oxygen insertion in the presence of hydrogen peroxide [24].

The synthetic CYP102A family was constructed from parental sequences containing this mutation; all of the chimeric proteins can therefore potentially function as peroxygenases. We are also interested in their ability to be reconstituted into functional monooxygenases, utilizing NADPH and molecular oxygen for catalysis, by fusion to a reductase domain. The reductase domain of CYP102A1 (R1) spans ~585 amino acids and encodes a ~20 amino acid linker and the binding domains for the FMN, FAD and NADPH cofactors [23]. The reductases from CYP102A2 and CYP102A3 (R2 and R3) share 52-55% sequence identity with R1 and are comparable in size, the only notable difference being a linker region that is extended in R2 by seven amino acids [25]. Because the chimeric heme domains comprise sequences from three different parents, it is not obvious that fusion to wildtype reductase will generate a catalytically active holoenzyme, nor is it clear which reductase, R1, R2 or R3, should be used. For this initial characterization we therefore selected a set of 14 chimeric heme domains, reconstituted them with all three parental reductase domains, and determined peroxygenase and monooxygenase activities on eleven substrates. These activities have been analyzed to 1) assess the functional diversity of the chimeric enzymes, 2) determine whether substrates fall into groups for the purposes of predicting activities, and 3) compare the activities and specificities of the chimeric peroxygenases with those of their reconstituted monooxygenases.

1.3 Results

1.3.1 Cloning and Expression of P450 Heme Domains and Holoenzymes

Seventeen heme domains, including the three parent heme domains, were chosen for holoenzyme construction by fusion to a wild-type CYP102A reductase domain. For each heme domain, four proteins were examined—the heme domain and its fusion to each of the three reductase domains—for a total of 68 constructs. Heme domains contain the first 463 amino acids for A1 and the first 466 amino acids for A2 and A3. The reductase domains start at amino acid E464 for R1, K467 for R2 and D467 for R3 and encode the linker region of the corresponding reductase. A3 and its fusions with R1 and R2

expressed very poorly, yielding only a very small amount of protein after purification, and were therefore not analyzed further.

The chimeric sequences are reported in terms of the parent from which each of the eight sequence blocks is inherited (Supplemental Table 1.S1). Twelve of the fourteen chimeras were selected because they displayed relatively high activities on substrates in preliminary studies (data not shown). Chimera 23132233 was chosen because it displayed *low* peroxygenase activity, while 22312333 was selected because it is more thermostable than any of the parents (T_{50} = 62°C) [15]. For the constructs studied here, the reductase identity is indicated as the ninth sequence element, with R0 referring to no reductase (i.e., heme domain peroxygenase).

1.3.2 Activity Assays

To assess the functional diversity of the chimeric P450s, we measured their activities on the eleven substrates shown in Figure 1.1. Propranolol (PR), tolbutamide (TB) and chlorzoxazone (CH) are drugs that are metabolized by human P450s [9,26,27]. 12-p-nitrophenoxycarboxylic acid (PN) is a long-chain fatty acid surrogate; parent A1-R1 holoenzyme and the A1 heme domain (with the F87A mutation) both show high activity on this substrate. Previous work showed that A1 has weak peroxygenase activity on some of the aromatic substrates [16]. Aromatic hydroxylation products of all substrates can be detected quantitatively using the 4-amino antipyrine assay [28]. PN hydroxylation can be monitored spectrophometrically [29].

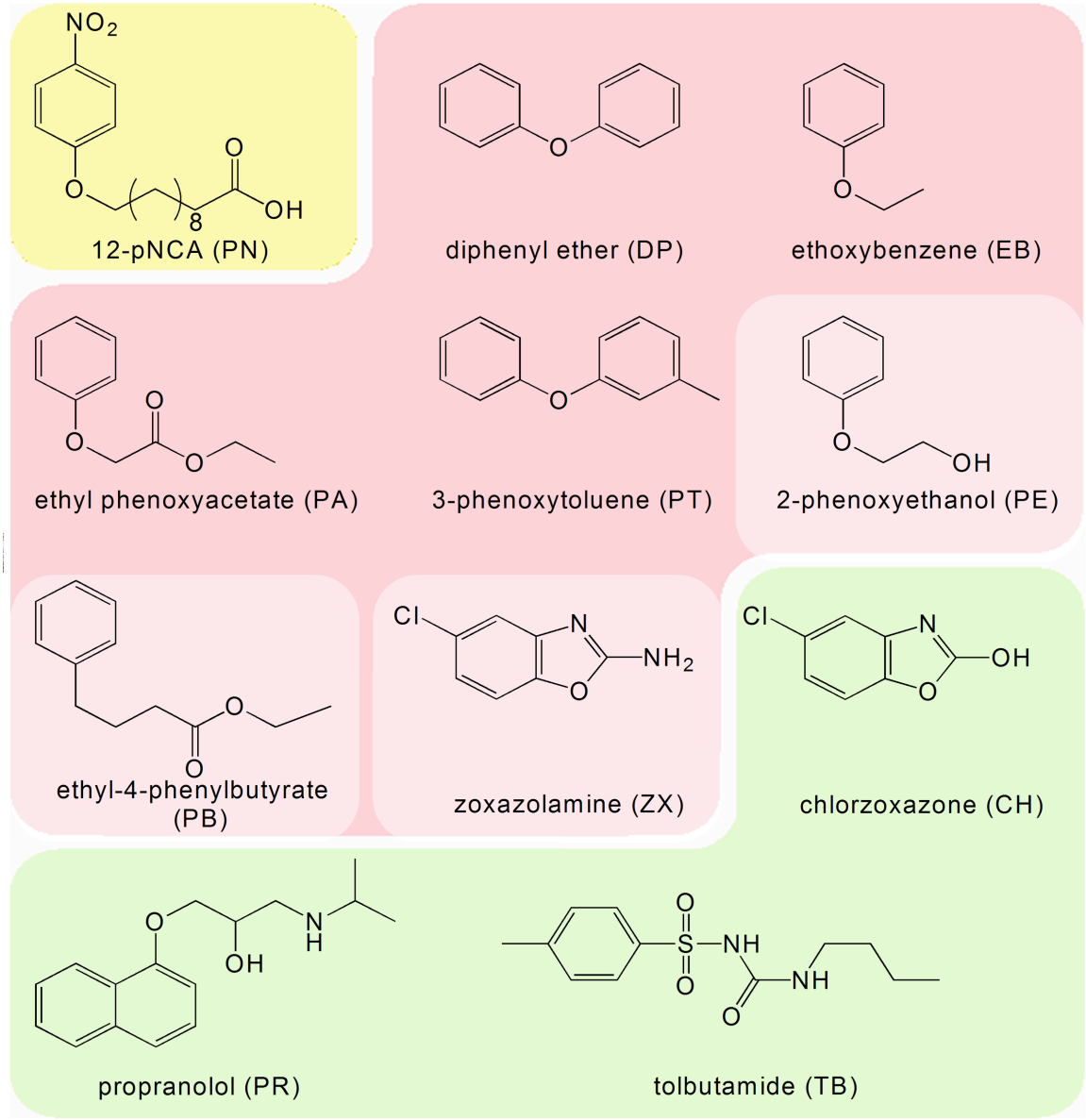


Figure 1.1: Chemical structures and abbreviations. Substrates are grouped according to the pair-wise correlations (see text for details). Members of a group are highly correlated; intergroup correlations are low.

Peroxygenase activities of the 16 heme domains (all except A3) were determined by assaying for product formation after a fixed reaction time in 96-well plates (see Experimental). Similar assays were used to determine monooxygenase activities for each of the fusion proteins. Final enzyme concentrations were fixed to 1 μM in order to reduce large errors associated with low expression and to allow us to compare chimera activities

using absorbance values directly. Protein concentrations were reassayed in 96-well format and determined to be 0.88 μM +/- 13% (SD/average). All samples were prepared and analyzed in triplicate, and outlier data points were eliminated. Supplemental Tables 1.S2 and 1.S3 report the averages and standard deviations for each of the assays. More than 85% of the data for each substrate was retained, and more than 95% was retained for 6 of the 11 substrates (Supplemental Table 1.S4).

Because extinction coefficients are not known for the reaction products, we do not report absolute enzyme activities, nor do we report substrate specificities, which are ratios of enzyme activity on one substrate to activity on another. Our data nonetheless allow us to compare the chimeras with respect to their activities on a given substrate and also to compare their activity profiles and therefore their specificities. Chimeras having a similar profile form the same relative amounts of products from all substrates and are therefore likely to have similar specificities. To better visualize differences among chimeras, the highest average absorbance value for a given substrate was set to 100%, and all other absorbances for the same substrate, but different chimeras, were normalized to this. Figure 1.2 is a heat plot of the complete data set of normalized absorbances, while Supplemental Figure 1.S2 shows the substrate-activity profiles in the form of bar plots.

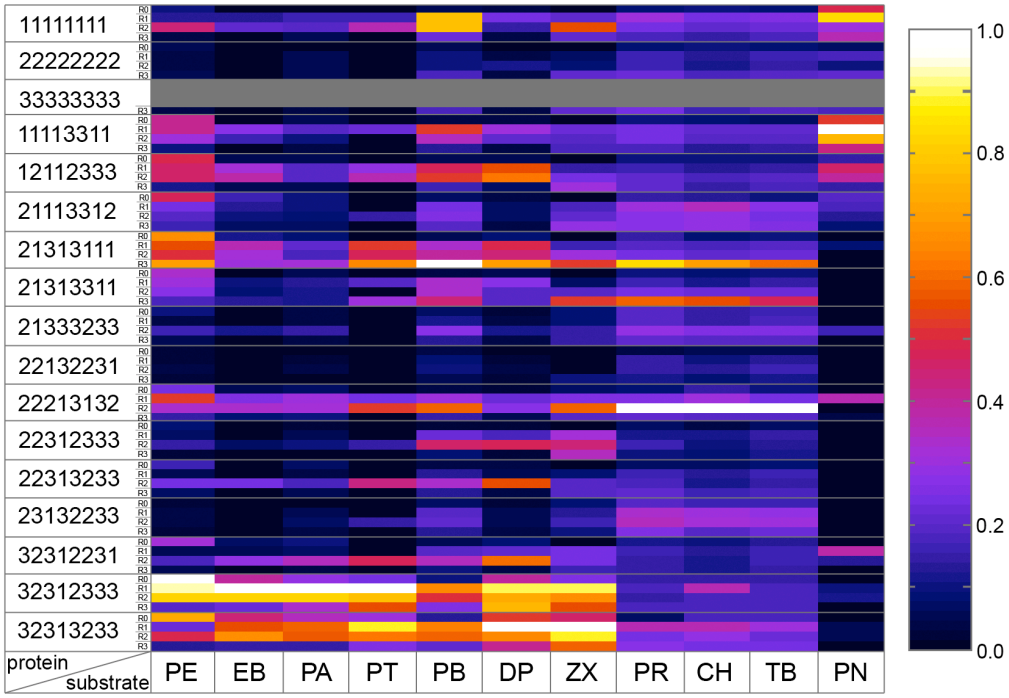


Figure 1.2: Summary of normalized activities for all 56 enzymes acting on 11 substrates. Activities are shown using a color scale (white indicating highest and black lowest activity), with columns representing substrates and rows representing proteins. Not-analyzed A3, A3-R1 and A3-R2 proteins are shown in grey. Protein rows are ordered by their chimeric sequence first, and then by heme domain (R0) and R1-, R2- and R3-fusions.

1.3.3 Activities of Parent Enzymes

Figure 1.3A shows the normalized substrate-activity profiles of the A1 and A2 peroxygenases. Both have relatively low or no activity on any of the substrates except PN, where A1 makes about an order of magnitude more product than does A2. Profiles for the reconstituted parent holoenzymes are shown in Figure 1.3B. Fusion of A1 and R1 generated an enzyme with profile peaks on ethyl 4-phenylbutyrate (PB) and PN. A1 is in fact the second-best-performing enzyme on PB. The A1 peroxygenase activity on this substrate, however, is among the worst, showing that peroxygenase specificity does not necessarily predict that of the monooxygenase. Fusion of A2 to R2 slightly increased

activity relative to A2, but did not alter the profile. The A3-R3 holoenzyme exhibits some activity on the drug-like substrates (PR, TB, CH) as well as PN and PB.

Fusion of the A1 and A2 heme domains to other reductase domains yields holoenzymes that are active on some substrates (Figures 1.3C and 1.3D). The A2 fusions have relatively low activities. A1 fusions with R1 and R2, on the other hand, created highly active enzymes with different specificities: the A1-R1 profile has peaks on PN and PB, while that of A1-R2 has peaks on PB, phenoxyethanol (PE) and zoxazolamine (ZX). The A1-R3 fusion is less active on nearly all substrates.

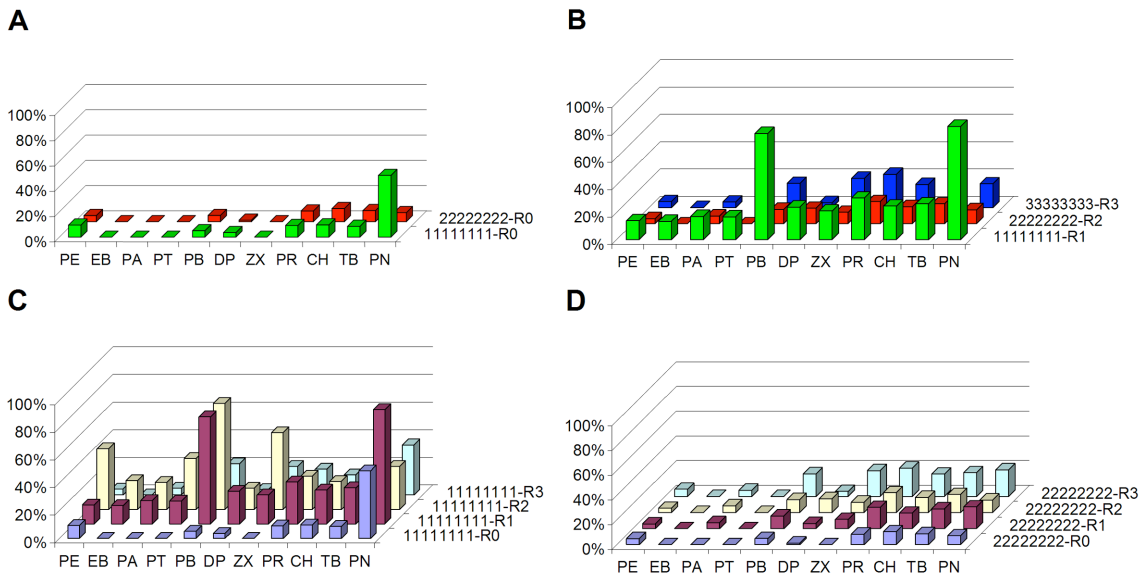


Figure 1.3: Substrate-activity profiles for parent heme domain mono- and peroxygenases. Panel A shows parent peroxygenases, panel B parent holoenzyme monooxygenases profiles, panel C the A1 protein set and panel D the A2 protein set. In A and B the color indicates the origin of the heme domain (Green = A1; Red = A2; Blue = A3). The protein set in panel C includes the heme domain A1 (blue) or its R1- (purple), R2- (yellow) or R3-fusion (turquoise) protein. Panel D depicts the A2 protein set.

1.3.4 Activities of Chimeras and Identification of Chimera Clusters

The fourteen chimeric heme domains generated 56 chimeric peroxygenases and monooxygenases. Nearly all the chimera fusions outperformed even the best parent

holoenzyme, and chimeric peroxygenases consistently outperformed the parent peroxygenases (Figure 1.2 and Supplemental Figure 1.S2). The best enzyme for each substrate is listed in Supplemental Table 1.S5. All the best enzymes are chimeras. Most of the best enzymes are also holoenzymes—only PE has a peroxygenase as the best catalyst.

We now show that there exists a discrete set of characteristic substrate-activity profiles to which each chimera can be uniquely assigned. A k-means clustering analysis was applied to the normalized absorbance data to better understand the functional diversity. K-means clustering, a statistical algorithm that partitions data into clusters based on data similarity [30], has been used by Mannervik and co-workers to identify groups of mutants exhibiting similar substrate specificities [31] and by others to identify protein fragments (4-7 residues) of similar structure [32] and interacting nucleotide pairs with similar three dimensional structures [33]. For our analysis, the normalized data were used to ensure that each of the 11 dimensions is given equal weight by the clustering algorithm. The clustering was performed over values of k (number of clusters) ranging from k = 2 to k = 8. The highest silhouette value (see Experimental) was observed at k = 5.

The cluster composition for k=5 is depicted in Figure 1.4. Cluster 1, consisting of chimeras 32312333-R1/R2 and 32313233-R1/R2 (Figure 1.4B), is characterized by low relative activities on CH, TB, PR and PN and high relative activities on all other substrates. In fact, two of these chimeras are the best enzymes on all the remaining substrates except PB and PE.

Cluster 2 is made up of 22213132-R2, 21313111-R3, 21313311-R3, which are the most active enzymes on TB, CH, and PR (Figure 1.4C). Cluster 2 enzymes are entirely inactive on PN and show low activity on most of the substrates that cluster 1 enzymes accept (PE, DP, PA, and EB). Relative activities on the remaining substrates (i.e., PB, ZX, and PT) are moderate (although lower than cluster 1 chimeras). An exception is 21313111-R3, which is the best enzyme for PB and also fairly good on PE and DP.

Cluster 3 contains chimeras A1-R1/R2, 12112333-R1/R2, 11113311-R1/R2, and 22213132-R1 (Figure 1.4D). The A1-like sequences are characterized by high relative

activity on PN (on which 11113311-R1/R2 and A1-R1 are the three top-ranking enzymes), and moderate to high relative activity on PB and moderate activity on PE.

Cluster 4 contains 21313111-R1/R2, 22313233-R2, 22312333-R2, 32312231-R2, 32312333-R0, 32312333-R3, 32313233-R0, and 32313233-R3 (Figure 1.4E). This cluster is characterized by having the highest relative activity on PE, in addition to moderate activities on PT, DP and ZX. The remaining chimeras appear in a fifth cluster with relatively low activity on everything except PN and PE (Figure 1.4F). This cluster contains parental sequences A1-R0, A1-R3, A2-R0, A2-R1/R2/R3 and A3-R3. Native sequences are thus only found in two of the clusters. The remaining clusters (1, 2 and 4) are made up of highly active chimeras that have acquired novel profiles.

The partition created by the clustering algorithm shows that the presence and identity of the reductase can alter the activity profile and thus the specificity of a heme domain sequence. For example, the R1 and R2 fusions of 32312333 and 32313233 appear in cluster 1, whereas their R0 and R3 counterparts are in cluster 4. Sequences 22213132 and 21313111 also behave differently when fused to different reductases. 22213132-R2, for example, displays pronounced peaks on substrates TB, CH and PR that are not present in the corresponding peroxygenase and R1/R3 profiles (Supplemental Figure 1.2E) and is thus the only member with this heme domain sequence appearing in cluster 2. 21313111-R3 and 21313111-R2/R1 have nearly opposite profiles (Supplemental Figure 1.S2J) and consequently appear in different clusters. Thus the best choice of reductase depends on both the substrate and the chimera sequence.

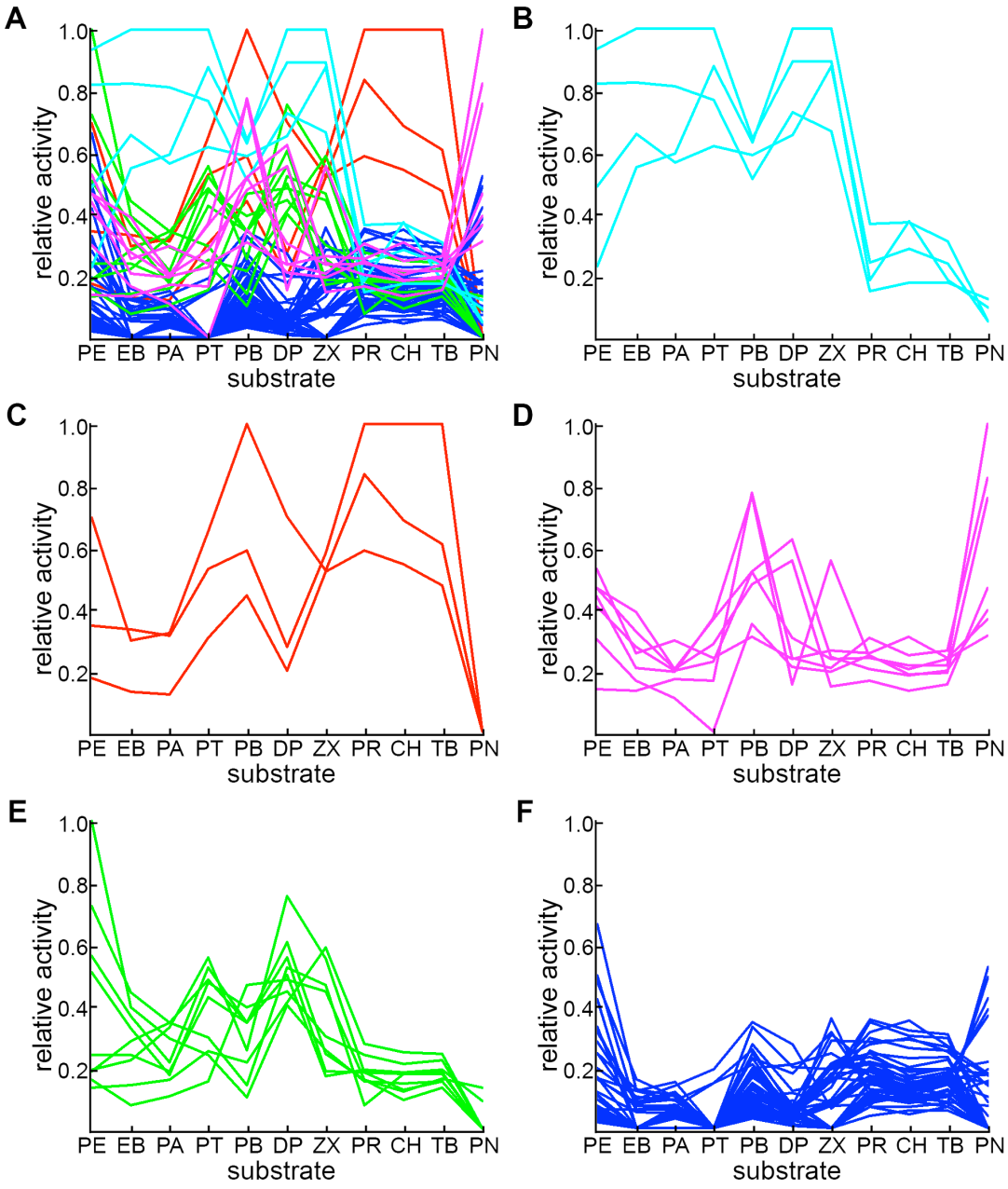


Figure 1.4: K-means clustering analysis separates chimeras into five clusters. All protein-activity profiles are depicted in A, where the color identifies the cluster. Panels B through F show profiles for sequences within each cluster. Panel B depicts 32312333-R1/R2, 32313233-R1/R2. Panel C depicts 22213132-R2, 21313111-R3, 21313311-R3. Panel D depicts A1-R1/R2, 12112333-R1/R2, 11113311-R1/R2 and 22213132-R1. Panel E depicts 21313111-R1/R2, 22313233-R2, 22312333-R2, 32312231-R2, 32312333-R0, 32312333-R3, 32313233-R0, and 32313233-R3. Panel F depicts the remaining sequences.

1.3.5 Peroxygenase Versus Monooxygenase Activities

As shown in Figure 1.2, each of the 14 chimeric heme domains can be fused to a parental reductase to generate a functional monooxygenase. The resulting monooxygenases are generally more active under these conditions than the corresponding peroxygenases (see Supplemental Figure 1.S2). The R1 and R2 fusions tend to outperform R3 fusions. While altering reductase identity never completely deactivates the protein, it does affect specificity in some cases. To quantify the differences between the profiles of the four different enzymes that can be made from a given chimera, the pair-wise linear coefficients (R^2) of the R0/R1, R0/R2, R0/R3, R1/R2, R1/R3 and R2/R3 profiles were determined for each heme domain sequence (with the exception of A3). The results are shown in Supplemental Table 1.S1. High correlations represent enzyme pairs with similar specificities. The results show that peroxygenase and monooxygenase specificities are usually different, R1/R2 fusions of a chimera are often very similar (five pairs have R^2 values above 0.9), and the R1 and R2 fusions are less similar to the R3 enzymes.

1.3.6 Identification of Substrate Groups

To understand whether a chimera's activity on one substrate predicts activity on another, the pair-wise correlations of the absorbances of all the possible substrate pairs were determined (Supplemental Table 1.S5). Mannervik and co-workers used correlations between activities on substrate pairs to identify enzyme variants with novel substrate specificities [13]. Here we use these correlations instead to identify substrates having similar *chimera* profiles. This analysis led to the identification of three substrate clusters characterized by high values of the correlation coefficients. Members of different clusters are poorly correlated. DP, PT, PA and EB all exhibit high correlations with each other (R^2 = 0.71-0.92, see Supplemental Figure 1.S1A for an example) and were grouped into the core of substrate group A. Group B consists of CH, TB and PR. The categorization of this group is clearly defined: its members show high correlations with each other (R^2 above 0.9, see Supplemental Figure 1.S1B for an example), but correlate very poorly with the other substrates (R^2 = 0.01-0.37). PN does not correlate significantly with any of the other substrates tested (R^2 = 0.00-0.08) and is its own substrate group C.

ZX, PB and PE show moderate correlation to members of the group A core (R^2 = 0.56-0.66, 0.39-0.56 and 0.35-0.61, respectively). These substrates are considered loosely associated with group A since they do not belong to any other group due to poor correlation with each other and the remaining substrates.

There exists a correspondence between the chimera clusters and the substrate groups. Group A core substrates have cluster 1 chimeras as their top-performing enzymes, whereas substrates of group B have cluster 2 chimeras as their top-performing enzymes. The top catalysts for group C are three of the cluster 3 chimeras. Members of a substrate group thus share the same best-performing enzymes.

1.4 Discussion

1.4.1 SCHEMA Recombination Creates a Family of Functionally Diverse Enzymes

We have begun to characterize the functional diversity in a synthetic P450 family created by structured-guided recombination of bacterial fatty acid hydroxylases. The folded P450s, which make up almost 50% of the 6,561 sequences in the SCHEMA library, contain an average of 72 mutations from their closest parent. A large fraction of the folded P450s were shown to be catalytically active [15], but they had been systematically studied on only a single substrate (PN). We therefore selected 11 substrates for this initial characterization of 14 of the active chimeric heme domains and their fusions with each of the three parental reductase domains. Although most of the parental enzyme constructs are poorly active on the selected substrates, many of the chimeras are significantly more active. In fact, for every single substrate, including one widely used to assay CYP102A1 (PN), the top-performing enzyme is a chimera. Recombining mutations already accepted in natural homologs thus leads to a family of highly active enzymes that accept a broader range of substrates.

1.4.2 Chimeras Can be Clustered by Substrate Specificity

We further showed that the chimeric enzymes exhibit distinct specificities and that they can be partitioned into clusters based on their specificity. One cluster contains parent A1-R1 and all chimeras with A1-like profiles. Another cluster contains low activity chimeras and includes all remaining parental sequences. The remaining clusters represent highly active chimeras that have acquired new specificities. Members of a cluster are likely to exhibit common structural, physical or chemical features that account for their similar catalytic properties. If the library is large enough, statistical techniques can be used to determine how sequence elements relate to the observed profiles. In particular, if there are sufficient numbers of chimeras in each cluster, then powerful tools such as logistic regression or machine learning can be used to predict which cluster an untested sequence belongs to [15]. This type of analysis would enable the prediction of substrate profiles of untested chimeras based on sequence information alone. The functionally diverse enzymes generated by SCHEMA-guided recombination can therefore be used to probe the sequence and structural basis of enzyme specificity. We recently observed the success of such an approach in predicting the thermostabilities of untested chimeras [Yougen Li, et al. unpublished data]. Although the current data set does not contain enough sequences for a comprehensive analysis of sequence-function relationships, anecdotal observations can be used to generate hypotheses for further testing. For example, the chimeras in the library with parent A1 in blocks 1, 3 and 4 are all among the best enzymes for PN. These same enzymes display low relative activity on all the remaining substrates except for PB. This suggests that having parent A1 sequence at one or more of these blocks improves PN activity and specificity.

1.4.3 Substrates Fall into Groups that Correlate with Chimera Clusters

We were also able to partition the substrates into groups based on the linear correlations of substrate pairs. An enzyme active on one member of a substrate group is therefore likely to be active on another member of the same group. One group consists of the drug-like substrates TB, PR and CH (Figure 1.1). Another consists of PT, PA, EB and DP. If these correlations hold for the larger library of chimeric enzymes, we should be able to

predict with reasonable accuracy the relative activities of a chimera on all the substrates in a group by testing activity on only one. This type of analysis can be expanded to a larger collection of substrates to identify additional groups or additional members of an existing group.

The observed correspondence between the three substrate groups and chimera clusters 1, 2 and 3 illustrates that each group can be associated with a cluster made up of or containing the top-performing enzymes for the substrates in that group. Some degree of correspondence can be expected, given how the partitions were constructed. However, because intra-group correlations are not one and inter-group correlations are not zero, the correspondence is not perfect. For this reason there exist chimeras whose profiles exhibit peaks on only certain members of a group (cluster 4) and others that exhibit peaks on members of different groups (cluster 2 and 3 chimeras). Cluster 4 chimeras have peaks on only certain members of group A and are thus responsible for the lower correlations among group A substrates. Some cluster 2 and cluster 3 chimeras exhibit peaks on PB (on the edge of group A) as well as group B and C, respectively. In fact although PB correlates mostly with group A core substrates it shares its top-performing enzymes with groups B and C and thus displays a hybrid behavior. This is why PB correlates less with group A than core substrates do and why it has higher correlations with group B and C members than any other substrate not belonging to these groups.

Because chimeras displaying high relative activity have more weight in determining the correlation coefficients, the top enzymes for one member of a substrate group will usually be among the top ones for all members of that group. The clearer the definition of the substrate groups, the more likely this is to hold. Given the many important applications of P450s in medicine and biocatalysis, and the lack of high-throughput screens for many compounds of interest, an approach to screening that is based on carefully chosen ‘surrogate’ substrates could significantly enhance our ability to identify useful catalysts. Clearly, any member of a well-defined substrate group can be a surrogate for other members of that group. Further analysis may also help to identify the critical physical, structural or chemical properties of substrates belonging to a known group. This will make it possible to predict which chimeras will be most active on a new untested substrate.

1.4.4 Swapping Reductase Domains Consistently Yields Active Monooxygenases and Conserves Key P450-Reductase FMN Domain Interactions

The literature reports multiple cases in which functional P450s have been reconstituted with new reductase domains. In several studies, swapping reductases improved mammalian P450 activity [34-36]. A self-sufficient chimeric mammalian P450 2E1 enzyme was constructed by fusing the 2E1 heme domain to the CYP102A1 reductase [37]. Functional chimeras of CYP102A1 and the flavocytochrome nitric oxide synthase (nNOS) have been generated [38]. Another study reported the functional expression of CYP153A genes by incorporating them into a framework consisting of the N- and C-termini of homolog CYP153A13a and fusion to the reductase domain of CYP116B2 [39].

Reconstitution of the chimeric CYP102A heme domains with the three parental reductases generated functional monooxygenases in all cases. Although their specificities were often different (particularly when fused to R3), fusion to a reductase was never detrimental to activity, and swapping the reductase never completely inactivated the enzyme (Supplemental Figure 1.S2). Subtle changes in the structure and coupling behavior that affect total product formation may account for specificity differences. The fact that the parental reductase domains are accepted without loss of function, however, suggests that key domain-domain interactions are conserved upon reductase swapping.

Although a complete crystal structure of a CYP102A holoenzyme is not available, a partial CYP102A1 structure (1BVY) includes the interface between the heme and the reductase FMN domains. Only a few direct contacts, including one hydrogen bond, one salt bridge and several water-mediated contacts, make up this A1-R1 interface [40]. We aligned the parental sequences using ClustalW [41] and found that the interactions depicted in the 1BVY crystal structure involve amino acids that are mostly conserved in the parent proteins. Figure 1.5 displays the interface between the heme and reductase domains of CYP102A1 and highlights the amino acids involved in key interactions. The salt bridge is formed between reductase residue E494 and heme domain residue H100, both of which are conserved in all three parents. Thus this key interaction would be retained upon reductase swapping that conserves the orientation of the two domains.

The direct hydrogen bond occurs between the reductase backbone carbonyl of N573 and the side-chain hydroxyl group of heme domain residue S383. N573 is only conserved in R1 and R2, but because the interaction involves the backbone oxygen, the reductase side of the interface is not affected by changes in the side-chain identity. S383 is only conserved in parents A1 and A3. However, the corresponding residue in A2, D385, may also be capable of forming the hydrogen bond. This interaction may therefore be present in all the chimeras.

There are two water-mediated hydrogen bonds between the hydrogen of the indole nitrogen of reductase residue W574 and the backbone carbonyl of S383 and I385. W574 was earlier shown to be crucial for electron transfer from the FMN to the heme [42] and is conserved in R1, R2 and R3. S383 and I385 are conserved in A1 and A3 but not A2, where the corresponding residues are D385 and V387. Because the hydrogen bonds involve the backbone oxygens of these residues, these interactions may be retained upon domain substitution. Also, all possible pair-wise interactions that can be formed at these positions by domain swapping already exist in at least one of the parental sequences and are thus likely not to be destabilizing. Finally, the substitutions that do occur are conservative, replacing a hydrophilic residue with another hydrophilic residue and a hydrophobic residue with another hydrophobic residue. The third water-mediated hydrogen bond between the side chains of reductase residue R498 and heme domain residue E244 (block 5) is conserved in A1-R1, A2-R2 but not A3-R3, where the corresponding residues are G501 and V246. A3-R3 thus cannot form this interaction nor can any chimera that inherits A3 sequence at block 5 and/or is fused to R3.

In summary, it appears that the direct hydrogen bond, two of the three water-mediated hydrogen bonds and the salt bridge are all conserved in the chimera-reductase fusions. The third water-mediated hydrogen bond is conserved only in R1/R2 fusions that do not have parent A3 in block 5 (8 out of 17 sequences). Thus the activities of the reconstituted monooxygenases are consistent with their sequences, the domain-domain interactions identified in the 1BVY structure and the assumption that the overall structures and orientations are conserved upon reductase swapping. These results demonstrate the highly conservative nature of mutation by recombination of protein

domains: as long as key interactions are retained, the remaining sequences can vary extensively.

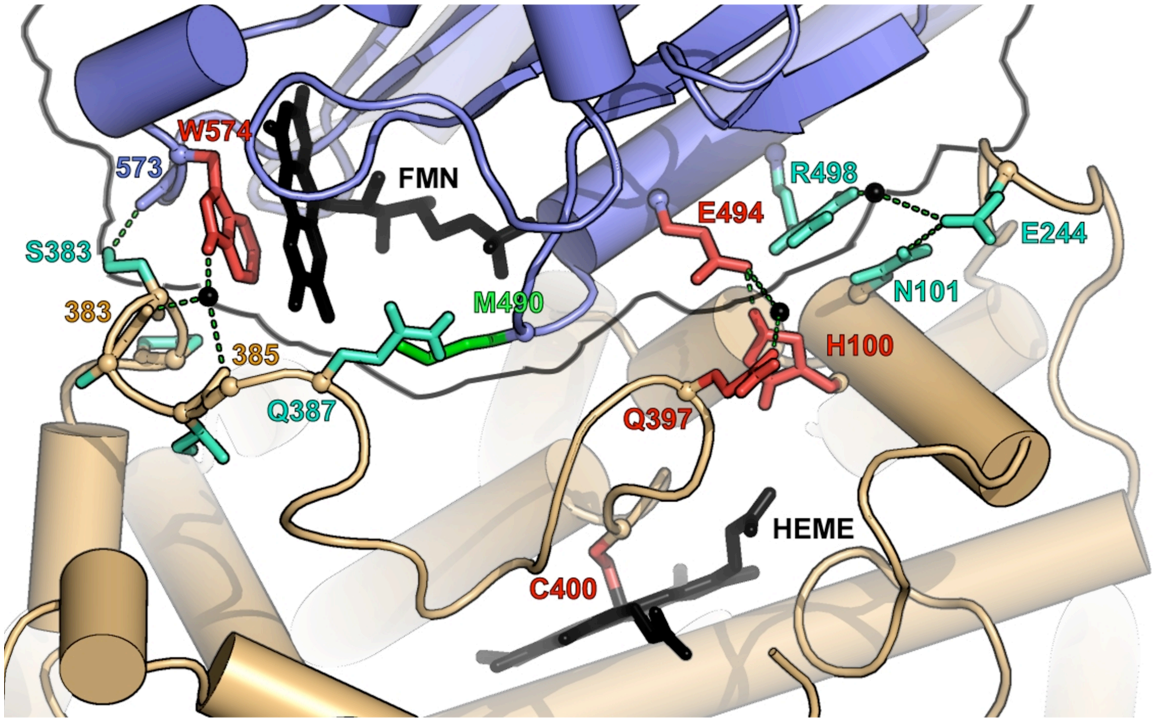


Figure 1.5: Interface between the FMN (blue backbone) and heme domain (brown backbone) based on the 1BVY structure redrawn according to Sevrioukova et al. [40]. Residue colors indicate the degree of conservation: red (three parents), turquoise (two parents) and green (not conserved). Hydrogen bonds are shown as dashed lines. The amino acids correspond to CYP102A1 numbering. PyMOL was used to create this figure [43].

1.5 Conclusions

The evolvable cytochrome P450 scaffold has diversified over millions of years of mutation and natural selection to exhibit the myriad activities of the natural enzyme family, of which more than 4,500 sequences are known [44]. We constructed a large synthetic P450 family by recombining sequence elements from three bacterial P450s [15]. We have now shown that members of this synthetic family exhibit diverse activities and specificities, including activities towards substrates that are not accepted by the parent P450s and drug-like compounds that are substrates of human P450s. Thus

enzymes in this family have acquired the ability to mimic important reactions in human drug metabolism. The grouping of substrates according to likelihood that a given chimera will accept them, as has been demonstrated here, will aid in the identification of useful catalysts from this synthetic family by high-throughput screening of substrate 'surrogates'. We anticipate that these enzymes will be useful for synthesis of drug metabolites [26], as needed for toxicity testing and drug discovery.

1.6 Experimental Methods

1.6.1 Nomenclature and Construction of Holoenzymes from Chimeric Heme Domains

Details of chimera construction have been reported previously [15]. Sequences are given an eight-digit number, where each digit indicates the parent from which each of the eight blocks was inherited. The identity of the reductase is indicated by R0 (for no reductase) or R1, R2 or R3 for the CYP102A1, A2, or A3 reductases, respectively.

To construct the holoenzymes, the chimeric heme domains were fused to each of the three wild-type reductase domains after amino acid residue 463 when the last block originates from CYP102A1 and 466 for CYP102A2 and CYP102A3. The holoenzymes were constructed by overlap extension PCR [45] and/or ligation and cloned into the pCWori expression vector [46]. All constructs were confirmed by sequencing.

1.6.2 Protein Expression and Purification

Proteins were expressed in *E. coli* as described previously and purified by anion exchange on Toyopearl SuperQ-650M from Tosoh [47]. After binding of the proteins, the matrix was washed with a 30 mM NaCl buffer, and proteins were eluted with 150 mM NaCl (all buffers used for purification contained 25 mM phosphate buffer pH 8.0). Proteins were rebuffered into 100 mM phosphate buffer and concentrated using 30,000 MWCO Amicon Ultra centrifugal filter devices (Millipore). Proteins were stored at -20°C in 50% glycerol.

Protein concentration was measured by CO absorption at 450 nm as described [48]. A protein concentration of 1 μM was chosen for the activity assays. Protein concentrations were reassayed in 96-well format and determined to be 0.88 μM +/- 13% (SD/average).

1.6.3 Functional Assays

Proteins were assayed for mono- or peroxygenase activities in 96-well plates as described [15,49]. Heme domains were assayed for peroxygenase activity using hydrogen peroxide as the oxygen and electron source. Reductase domain fusion proteins were assayed for monooxygenase activity, using molecular oxygen and NADPH. Reactions were carried out in 100 mM EPPS buffer pH 8, 1% acetone, 1% DMSO, 1 μM protein in 120 μl volumes. Substrate concentrations depended on their solubility under the assay conditions. Final concentrations were: 2-phenoxyethanol (PE), 100 mM; ethoxybenzene (EB), 50 mM; ethyl phenoxyacetate (PA), 10 mM; 3-phenoxytoluene (PT), 10 mM; ethyl 4-phenylbutyrate (PB), 5 mM; diphenyl ether (DP), 10 mM; zoxazolamine (ZX), 5 mM; propranolol (PR), 4 mM; chlorzoxazone (CH), 5 mM; tolbutamide (TB), 10 mM; 12-p-nitrophenoxycarboxylic acid (PN), 0.25 mM. The reaction was initiated by the addition of NADPH or hydrogen peroxide stock solution (final concentration of 500 μM NADPH or 2 mM hydrogen peroxide) and mixed briefly. After two hours at room temperature, reactions with substrates 1-10 were quenched with 120 μl of 0.1 M NaOH and 4 M urea. Thirty-six μl of 0.6% (w/v) 4-aminoantipyrine (4-AAP) was then added. The 96-well plate reader was zeroed at 500 nm and 36 μl of 0.6% (w/v) potassium persulfate was added. After 20 min, the absorbance at 500 nm was read [28]. Reactions on PN were monitored directly at 410 nm by the absorption of accumulated 4-nitrophenol. All experiments were performed in triplicate, and the absorption data were averaged.

1.6.4 Data Analysis

The background absorbance (BG) was subtracted from the raw data. BG reactions contained buffer, cofactor and substrate in the absence of protein sample and were done

in triplicates. All absorbance measurements were done once on three separate samples (triplicate sampling). Data points with a SD/average ≥ 20% that did not lie within the average ± 1.1*SD were eliminated. 1.1*SD was chosen so that for each substrate at least 85% of the points were retained. This never resulted in the elimination of more than one point from each triplicate set of measurements. All points with an average absorbance < BG were set to zero, because they are assumed to belong to inactive proteins. The absorbance matrix thus obtained for all 68 proteins on all 11 substrates is displayed in Supplemental Table 1.S2. The SD/average matrix is displayed in Supplemental Table 1.S3. SD/average was calculated ignoring values for inactive enzymes.

1.6.5 Cluster Analysis

K-means clustering is a partitioning method that divides a set of observations into k mutually exclusive clusters. K-means treats each data point as an object having a location in m-dimensional space (m=11 in this analysis) [30]. It then finds a partition such that members of the same cluster are as close as possible to each other and as far as possible to members of other clusters. For this reason, a measure of the meaningfulness of a partition is given by the silhouette value $s = avg\left(\frac{b(i)-a(i)}{\max[a(i),b(i)]}\right)$, where a(i) is the average distance of point i to all other points in its cluster and b (i) is the average distance of point i to all points in the closest cluster. It is evident that $-1 \le s \le 1$ and the quality of the clustering increases as s -> 1 [50]. Distances are measured by the square of the Euclidean distance.

1.7 Acknowledgments

This work is supported by the National Institutes of Health (R01 GM068664-0) and a Naional Science Foundation Predoctoral Fellowship (to MC). The authors thank Sally A. Kim for critically reading the manuscript, Daniela C. Dieterich for help with Figures 1.2 and 1.4 and Christopher Snow for Figure 1.5. ML, CRO and FHA designed and planned the project, YL performed all the cloning involved in attaching the reductase to the heme

domain, ML purified all the enzymes, ML and MC performed all the activity asasays, MC processed and analyzed the data. MC, ML, and FHA prepared the manuscript.

1.8 Supplementary Material

Table 1.S1: Pair-wise correlations of normalized activities for monooxygenases (R1, R2, R3) and peroxygenases (R0) of fourteen chimeras and the A1 and A2 parents. R^2 values are reported. Bold and underlined=0.7-1.0; Underlined=0.4-0.7; Regular=0.0-0.4.

Heme sequence	R0/R1	R0/R2	R0/R3	R1/R2	R1/R3	R2/R3
11111111	0.49	0.00	0.53	0.21	0.66	0.11
22222222	**0.70**	0.53	0.49	**0.75**	**0.83**	0.66
11113311	0.61	0.65	0.49	**0.90**	0.59	**0.78**
12112333	0.11	0.04	0.00	**0.91**	0.11	0.10
21113312	0.14	0.01	0.00	**0.73**	**0.76**	**0.77**
21313111	0.24	0.19	0.05	**0.84**	0.15	0.39
21313311	0.25	0.28	0.00	0.41	0.01	0.34
21333233	**0.90**	0.64	**0.87**	**0.72**	**0.95**	0.66
22132231	**0.80**	**0.85**	0.56	**0.98**	0.64	0.60
22213132	0.46	0.08	0.37	0.11	0.01	0.54
22312333	0.01	0.02	0.00	0.69	0.69	0.25
22313233	0.17	0.01	0.08	0.02	**0.85**	0.07
23132233	**0.96**	**0.89**	**0.97**	**0.90**	**0.99**	**0.90**
32312231	0.14	0.06	0.02	0.07	0.04	0.21
32312333	0.33	0.41	0.02	**0.97**	0.40	0.33
32313233	0.15	0.44	0.09	**0.74**	0.60	0.38

Table 1.S2: Average activity in absorbance units for each substrate-construct pair (maximal value for each substrate in bold/italic).

	2-phenoxyethanol	ethoxybenzene	ethyl phenoxyacetate	3-phenoxytoluene	ethyl 4-phenylbutyrate	diphenyl ether	2-amino-5-chloro-benzoxazole	propranolol	chlorzoxazone	tolbutamide	12-pNCA
11111111-R0	0.105	0.000	0.000	0.000	0.013	0.027	0.000	0.011	0.013	0.011	0.178
11111111-R1	0.152	0.115	0.136	0.053	0.202	0.177	0.055	0.037	0.032	0.033	0.302
11111111-R2	0.484	0.179	0.157	0.118	0.200	0.114	0.146	0.029	0.026	0.029	0.114
11111111-R3	0.048	0.000	0.038	0.000	0.059	0.030	0.054	0.023	0.019	0.022	0.132
22222222-R0	0.054	0.000	0.000	0.000	0.013	0.009	0.000	0.010	0.014	0.011	0.026
22222222-R1	0.042	0.000	0.038	0.000	0.027	0.031	0.020	0.021	0.016	0.020	0.064
22222222-R2	0.039	0.000	0.045	0.000	0.027	0.083	0.022	0.020	0.016	0.018	0.037
22222222-R3	0.065	0.000	0.040	0.000	0.048	0.031	0.055	0.028	0.024	0.024	0.079
33333333-R3	0.049	0.000	0.033	0.000	0.046	0.026	0.056	0.030	0.022	0.024	0.063
11113311-R0	0.463	0.000	0.046	0.000	0.011	0.031	0.000	0.013	0.012	0.009	0.190
11113311-R1	0.448	0.238	0.160	0.072	0.135	0.225	0.061	0.029	0.028	0.027	***0.364***
11113311-R2	0.329	0.145	0.087	0.000	0.091	0.159	0.051	0.030	0.024	0.024	0.277
11113311-R3	0.118	0.000	0.033	0.000	0.032	0.028	0.047	0.022	0.017	0.019	0.155
12112333-R0	0.544	0.053	0.048	0.000	0.013	0.038	0.000	0.012	0.014	0.013	0.056
12112333-R1	0.513	0.282	0.163	0.091	0.124	0.414	0.038	0.020	0.017	0.019	0.170
12112333-R2	0.511	0.334	0.163	0.116	0.135	0.462	0.063	0.025	0.024	0.025	0.143
12112333-R3	0.129	0.044	0.039	0.000	0.043	0.058	0.080	0.025	0.019	0.022	0.053
21113312-R0	0.522	0.135	0.078	0.000	0.017	0.034	0.000	0.017	0.017	0.013	0.069
21113312-R1	0.269	0.107	0.084	0.000	0.063	0.056	0.046	0.038	0.045	0.034	0.065
21113312-R2	0.213	0.085	0.073	0.046	0.066	0.047	0.055	0.033	0.038	0.031	0.050
21113312-R3	0.179	0.063	0.058	0.000	0.049	0.034	0.075	0.034	0.037	0.033	0.031
21313111-R0	0.731	0.105	0.073	0.000	0.016	0.058	0.000	0.018	0.012	0.013	0.000
21313111-R1	0.617	0.313	0.173	0.167	0.089	0.370	0.044	0.024	0.024	0.024	0.033
21313111-R2	0.560	0.282	0.139	0.152	0.102	0.332	0.079	0.029	0.027	0.028	0.000
21313111-R3	0.767	0.256	0.258	0.207	***0.260***	0.518	0.137	0.102	0.089	0.076	0.000
21313311-R0	0.365	0.000	0.046	0.000	0.009	0.038	0.000	0.012	0.011	0.012	0.000
21313311-R1	0.343	0.082	0.109	0.061	0.089	0.202	0.017	0.019	0.015	0.019	0.000
21313311-R2	0.306	0.074	0.092	0.000	0.086	0.149	0.050	0.030	0.029	0.029	0.000
21313311-R3	0.190	0.109	0.098	0.097	0.115	0.150	0.136	0.072	0.071	0.060	0.000
21333233-R0	0.113	0.000	0.036	0.000	0.020	0.016	0.023	0.025	0.020	0.020	0.000
21333233-R1	0.046	0.000	0.035	0.000	0.029	0.026	0.022	0.024	0.019	0.022	0.000
21333233-R2	0.180	0.104	0.119	0.000	0.070	0.090	0.039	0.036	0.034	0.031	0.062
21333233-R3	0.057	0.000	0.035	0.000	0.036	0.028	0.040	0.026	0.025	0.024	0.000
22132231-R0	0.034	0.000	0.000	0.000	0.009	0.006	0.000	0.005	0.008	0.007	0.000
22132231-R1	0.025	0.000	0.024	0.000	0.023	0.018	0.000	0.018	0.014	0.018	0.000
22132231-R2	0.045	0.000	0.035	0.000	0.026	0.033	0.000	0.018	0.016	0.020	0.000
22132231-R3	0.022	0.000	0.000	0.000	0.016	0.015	0.025	0.014	0.012	0.015	0.000
22213132-R0	0.269	0.051	0.061	0.000	0.010	0.017	0.020	0.010	0.019	0.013	0.000
22213132-R1	0.584	0.217	0.238	0.076	0.081	0.172	0.068	0.031	0.040	0.030	0.133
22213132-R2	0.377	0.289	0.253	0.169	0.153	0.206	0.152	***0.122***	***0.130***	***0.126***	0.000
22213132-R3	0.172	0.070	0.077	0.000	0.038	0.043	0.051	0.026	0.025	0.024	0.015
22312333-R0	0.103	0.000	0.024	0.000	0.008	0.017	0.000	0.009	0.006	0.009	0.000
22312333-R1	0.080	0.000	0.044	0.000	0.058	0.132	0.082	0.015	0.015	0.018	0.000
22312333-R2	0.172	0.067	0.084	0.049	0.121	0.356	0.117	0.019	0.012	0.017	0.000
22312333-R3	0.034	0.000	0.000	0.000	0.022	0.019	0.093	0.012	0.011	0.015	0.000
22313233-R0	0.185	0.000	0.050	0.000	0.011	0.029	0.000	0.008	0.009	0.010	0.000
22313233-R1	0.064	0.000	0.036	0.000	0.033	0.044	0.023	0.021	0.018	0.021	0.000
22313233-R2	0.260	0.204	0.150	0.137	0.089	0.415	0.049	0.022	0.016	0.019	0.000
22313233-R3	0.077	0.000	0.041	0.000	0.034	0.031	0.053	0.026	0.020	0.023	0.000
23132233-R0	0.024	0.000	0.000	0.000	0.019	0.019	0.022	0.025	0.021	0.021	0.000
23132233-R1	0.044	0.000	0.043	0.000	0.051	0.037	0.035	0.042	0.039	0.036	0.000
23132233-R2	0.049	0.000	0.055	0.046	0.054	0.044	0.043	0.043	0.041	0.038	0.000
23132233-R3	0.030	0.000	0.031	0.000	0.034	0.024	0.025	0.031	0.026	0.028	0.000
32312231-R0	0.354	0.065	0.085	0.000	0.016	0.067	0.000	0.015	0.013	0.018	0.000
32312231-R1	0.067	0.053	0.055	0.000	0.051	0.156	0.063	0.021	0.016	0.021	0.139
32312231-R2	0.204	0.245	0.277	0.154	0.090	0.448	0.063	0.019	0.016	0.020	0.048
32312231-R3	0.064	0.000	0.035	0.000	0.025	0.024	0.044	0.018	0.015	0.018	0.000
32312333-R0	***1.101***	0.338	0.236	0.076	0.025	0.297	0.067	0.019	0.019	0.019	0.000
32312333-R1	1.030	***0.860***	***0.803***	***0.320***	0.167	0.664	0.233	0.022	0.048	0.023	0.034
32312333-R2	0.907	0.712	0.653	0.246	0.133	0.538	0.174	0.018	0.023	0.022	0.044
32312333-R3	0.212	0.189	0.264	0.178	0.066	0.561	0.145	0.023	0.023	0.023	0.000
32313233-R0	0.796	0.383	0.276	0.095	0.036	0.389	0.121	0.009	0.023	0.023	0.000
32313233-R1	0.249	0.471	0.476	0.280	0.163	***0.742***	***0.261***	0.044	0.048	0.039	0.018
32313233-R2	0.535	0.566	0.454	0.197	0.153	0.485	0.229	0.029	0.037	0.029	0.017
32313233-R3	0.147	0.123	0.125	0.081	0.056	0.304	0.153	0.034	0.032	0.031	0.000

Table 1.S3: Standard deviations/ average of absorbance for each substrate-construct pair. Blanks indicate where the average absorbance equals zero.

	2-phenoxyethanol	ethoxybenzene	ethyl phenoxyacetate	3-phenoxytoluene	ethyl 4-phenylbutyrate	diphenyl ether	2-amino-5-chloro-benzoxazole	propranolol	chlorzoxazone	tolbutamide	12-pNCA
11111111-R0	0.091				0.233	0.735		0.162	0.148	0.098	0.052
11111111-R1	0.093	0.183	0.058	0.128	0.033	0.118	0.364	0.054	0.128	0.106	0.076
11111111-R2	0.039	0.020	0.118	0.135	0.041	0.030	0.112	0.113	0.120	0.067	0.159
11111111-R3	0.054		0.031		0.029	0.066	0.189	0.092	0.082	0.118	0.083
22222222-R0	0.089				0.156	0.264		0.261	0.005	0.159	0.125
22222222-R1	0.128		0.074		0.077	0.119	0.255	0.076	0.144	0.144	0.040
22222222-R2	0.071		0.054		0.113	0.081	0.251	0.085	0.108	0.099	0.011
22222222-R3	0.053		0.111		0.084	0.070	0.058	0.155	0.123	0.086	0.096
33333333-R3	0.134		0.126		0.017	0.094	0.082	0.110	0.155	0.088	0.068
11113311-R0	0.092		0.097		0.086	0.370		0.117	0.083	0.000	0.058
11113311-R1	0.045	0.158	0.124	0.092	0.159	0.032	0.622	0.084	0.127	0.079	0.007
11113311-R2	0.046	0.018	0.113		0.035	0.079	0.177	0.130	0.102	0.038	0.012
11113311-R3	0.103		0.093		0.033	0.065	0.110	0.110	0.176	0.022	0.102
12112333-R0	0.012	0.046	0.045		0.159	0.034		0.193	0.114	0.067	0.073
12112333-R1	0.092	0.014	0.114	0.107	0.029	0.104	0.065	0.177	0.137	0.069	0.075
12112333-R2	0.054	0.118	0.094	0.021	0.024	0.081	0.115	0.160	0.019	0.073	0.129
12112333-R3	0.039	0.016	0.057		0.020	0.035	0.064	0.082	0.066	0.115	0.133
21113312-R0	0.129	0.076	0.126		0.074	0.176		0.156	0.053	0.156	0.118
21113312-R1	0.065	0.049	0.060		0.045	0.046	0.075	0.156	0.051	0.058	0.250
21113312-R2	0.024	0.190	0.114	0.150	0.064	0.182	0.183	0.182	0.088	0.051	0.379
21113312-R3	0.094	0.147	0.087		0.051	0.044	0.005	0.350	0.121	0.110	0.080
21313111-R0	0.078	0.177	0.142		0.038	0.092		0.138	0.167	0.107	
21313111-R1	0.116	0.046	0.019	0.088	0.055	0.032	0.239	0.135	0.107	0.083	0.095
21313111-R2	0.012	0.084	0.076	0.039	0.037	0.069	0.424	0.083	0.106	0.088	
21313111-R3	0.038	0.200	0.092	0.034	0.034	0.107	0.195	0.035	0.145	0.127	
21313311-R0	0.065		0.143		0.162	0.078		0.041	0.168	0.105	
21313311-R1	0.026	0.051	0.166	0.178	0.086	0.024	0.448	0.029	0.097	0.072	
21313311-R2	0.137	0.141	0.169		0.018	0.049	0.020	0.183	0.084	0.049	
21313311-R3	0.012	0.053	0.038	0.075	0.010	0.111	0.131	0.148	0.091	0.040	
21333233-R0	0.062		0.242		0.110	0.188	0.377	0.159	0.133	0.128	
21333233-R1	0.095		0.049		0.038	0.192	0.189	0.085	0.074	0.120	
21333233-R2	0.036	0.183	0.135		0.016	0.044	0.026	0.119	0.117	0.062	0.105
21333233-R3	0.043		0.044		0.044	0.182	0.067	0.043	0.082	0.041	
22132231-R0	0.002				0.180	0.398		0.677	0.060	0.189	
22132231-R1	0.052		0.041		0.051	0.077		0.183	0.166	0.110	
22132231-R2	0.063		0.067		0.019	0.092		0.063	0.148	0.073	
22132231-R3	0.080				0.061	0.014	0.137	0.142	0.160	0.044	
22213132-R0	0.153	0.128	0.058		0.081	0.147	0.156	0.166	0.073	0.137	
22213132-R1	0.077	0.118	0.104	0.053	0.066	0.058	0.339	0.098	0.147	0.030	0.048
22213132-R2	0.065	0.091	0.059	0.075	0.050	0.039	0.070	0.124	0.120	0.005	
22213132-R3	0.097	0.061	0.116		0.061	0.052	0.119	0.144	0.111	0.114	0.000
22312333-R0	0.023		0.173		0.181	0.387		0.151	0.132	0.170	
22312333-R1	0.103		0.110		0.046	0.068	0.266	0.098	0.085	0.076	
22312333-R2	0.060	0.191	0.108	0.050	0.047	0.059	0.042	0.160	0.091	0.016	
22312333-R3	0.101				0.077	0.127	0.153	0.121	0.264	0.038	
22313233-R0	0.100		0.158		0.080	0.134		0.334	0.246	0.127	
22313233-R1	0.055		0.023		0.158	0.034	0.154	0.101	0.079	0.104	
22313233-R2	0.076	0.245	0.144	0.062	0.079	0.019	0.118	0.006	0.134	0.106	
22313233-R3	0.028		0.005		0.036	0.141	0.155	0.040	0.081	0.104	
23132233-R0	0.056				0.013	0.095	0.058	0.092	0.182	0.086	
23132233-R1	0.050		0.109		0.045	0.050	0.060	0.012	0.116	0.078	
23132233-R2	0.042		0.009	0.178	0.076	0.067	0.078	0.122	0.091	0.118	
23132233-R3	0.061		0.052		0.028	0.047	0.146	0.053	0.089	0.098	
32312231-R0	0.119	0.119	0.019		0.085	0.034		0.167	0.105	0.177	
32312231-R1	0.114	0.046	0.133		0.108	0.074	0.531	0.050	0.102	0.064	0.190
32312231-R2	0.088	0.061	0.062	0.146	0.107	0.058	0.174	0.096	0.191	0.088	0.085
32312231-R3	0.036		0.014		0.031	0.118	0.054	0.055	0.117	0.051	
32312333-R0	0.081	0.074	0.089	0.034	0.071	0.015	0.056	0.137	0.077	0.125	
32312333-R1	0.068	0.111	0.045	0.020	0.056	0.113	0.014	0.052	0.102	0.042	0.457
32312333-R2	0.051	0.107	0.035	0.019	0.049	0.097	0.150	0.173	0.023	0.068	0.139
32312333-R3	0.107	0.070	0.079	0.133	0.030	0.075	0.095	0.050	0.078	0.069	
32313233-R0	0.090	0.149	0.049	0.120	0.031	0.140	0.050	1.863	0.074	0.067	
32313233-R1	0.143	0.105	0.036	0.011	0.063	0.089	0.184	0.147	0.078	0.044	0.062
32313233-R2	0.064	0.053	0.033	0.020	0.083	0.113	0.102	0.122	0.072	0.035	0.346
32313233-R3	0.064	0.093	0.073	0.034	0.013	0.034	0.005	0.132	0.133	0.039	

Table 1.S4: Summary of error statistics for collected absorbance data sorted by substrates. The percentage of the standard deviation divided by the average value and the percentage of data points retained for the analysis are measures of data quality. For each substrate, 65 data points were collected. The Triplicates/Duplicates column indicates how many of those data points were used for the analysis performed here.

Substrate	% SD/avg (mean)	% points retained	Triplicates/ Duplicates
2-phenoxyethanol (PE)	7.1	99	63/2
ethoxybenzene (EB)	10.2	87	39/26
ethyl phenoxyacetate (PA)	8.5	95	56/9
3-phenoxytoluene (PT)	8.0	94	53/12
ethyl 4-phenylbutyrate (PB)	6.7	100	65/0
diphenyl ether (DP)	10.9	95	56/9
zoxazolamine (ZX)	16.0	87	40/25
propranolol (PR)	15.6	90	45/20
chlorzoxazone (CH)	11.2	99	63/2
tolbutamide (TB)	8.5	99	63/2
12-p-nitrophenoxycarboxylic acid (PN)	11.8	87	40/25

Table 1.S5: Summary of most active chimeric proteins for each substrate. Pair-wise correlation matrix of the activities on all substrates.). R^2 values are reported. Bold and underlined=0.7-1.0; Underlined=0.4-0.7; Regular=0.0-0.4

Protein		PE	EB	PA	PT	PB	DP	ZX	PR	CH	TB	PN
32312231-R0	PE	N.A.	0.61	0.48	0.37	0.18	0.35	0.15	0.01	0.05	0.02	0.01
32312231-R1	EB		N.A.	**0.92**	**0.80**	0.41	**0.73**	0.56	0.04	0.13	0.06	0.00
32312231-R1	PA			N.A.	**0.81**	0.39	**0.71**	0.62	0.04	0.14	0.06	0.00
32312231-R1	PT				N.A.	0.56	**0.85**	0.66	0.14	0.24	0.16	0.00
21313111-R3	PB					N.A.	0.49	0.49	0.36	0.37	0.33	0.08
32313233-R1	DP						**N.A.**	0.58	0.05	0.10	0.06	0.00
32313233-R1	ZX							N.A.	0.18	0.29	0.21	0.00
22213132-R2	PR								N.A.	**0.91**	**0.95**	0.00
22213132-R2	CH									N.A.	**0.94**	0.00
22213132-R2	TB										N.A.	0.00
11113311-R1	PN											N.A.

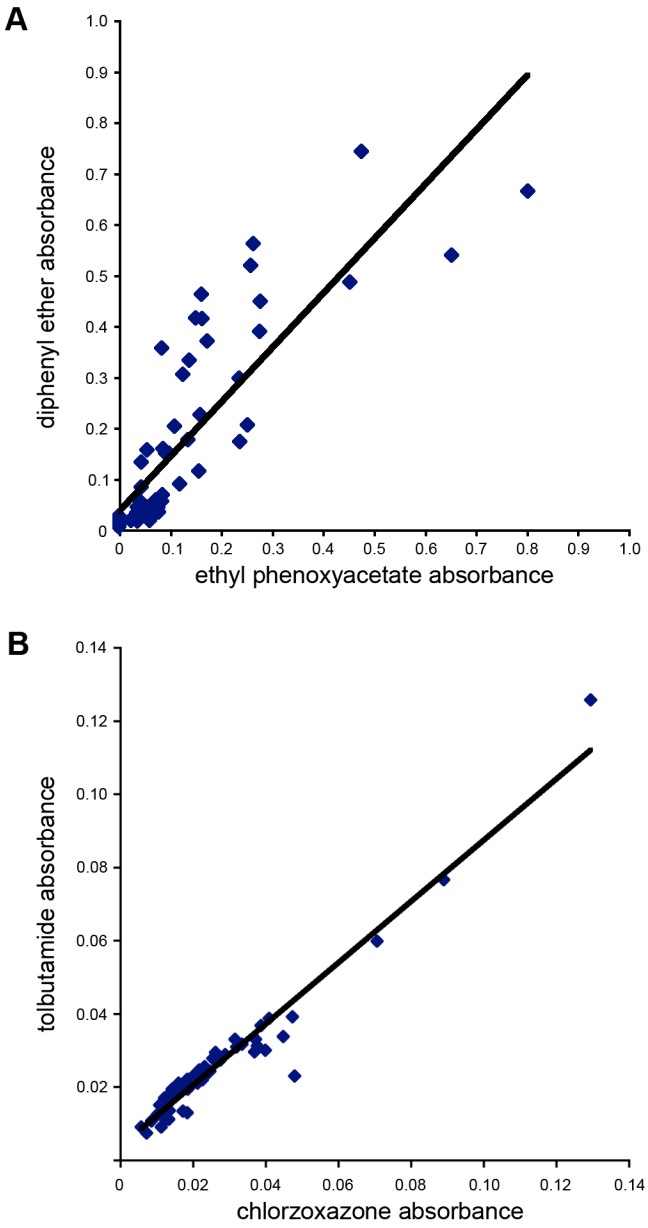
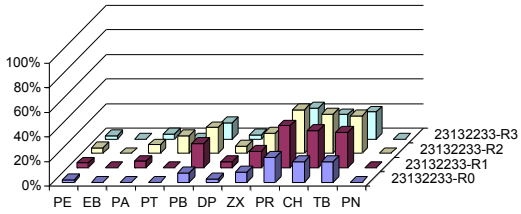
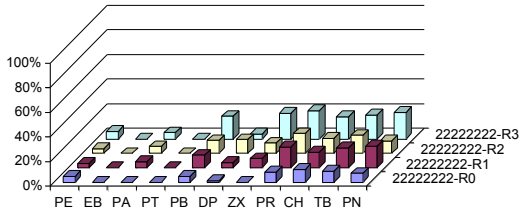


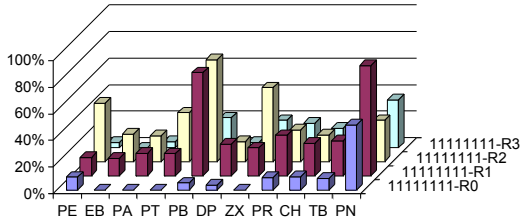
Figure 1.S1: Examples of the correlation of absorbances values measured within substrate Group A and Group B. Panel A shows the correlation between diphenyl ether (DP) and ethyl phenoxyacetate (PA) with a R^2=0.71. Panel B shows the correlation between tolbutamide (TB) activity and chlorzoxazone (CH) activity with R^2=0.94.


100%
80%
60%
40%
20%
0%
PE EB PA PT PB DP ZX PR CH TB PN
23132233-R3
23132233-R2
23132233-R1
23132233-R0

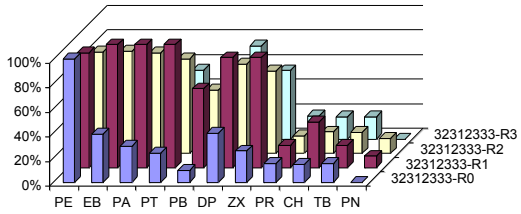
A


100%
80%
60%
40%
20%
0%
PE EB PA PT PB DP ZX PR CH TB PN
22222222-R3
22222222-R2
22222222-R1
22222222-R0

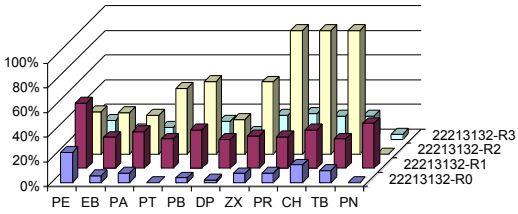
B


100%
80%
60%
40%
20%
0%
PE EB PA PT PB DP ZX PR CH TB PN
11111111-R3
11111111-R2
11111111-R1
11111111-R0

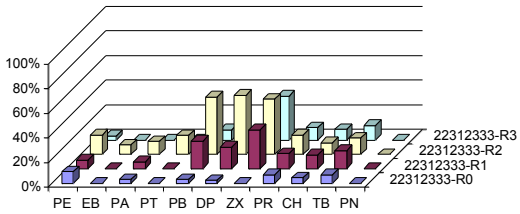
C


100%
80%
60%
40%
20%
0%
PE EB PA PT PB DP ZX PR CH TB PN
32312333-R3
32312333-R2
32312333-R1
32312333-R0

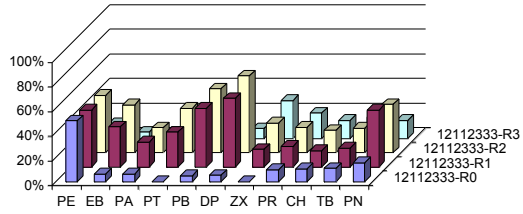
D


100%
80%
60%
40%
20%
0%
PE EB PA PT PB DP ZX PR CH TB PN
22213132-R3
22213132-R2
22213132-R1
22213132-R0

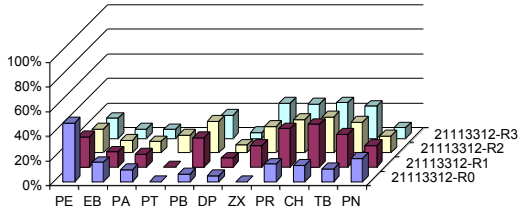
E


100%
80%
60%
40%
20%
0%
PE EB PA PT PB DP ZX PR CH TB PN
22312333-R3
22312333-R2
22312333-R1
22312333-R0

F


100%
80%
60%
40%
20%
0%
PE EB PA PT PB DP ZX PR CH TB PN
12112333-R3
12112333-R2
12112333-R1
12112333-R0

G


100%
80%
60%
40%
20%
0%
PE EB PA PT PB DP ZX PR CH TB PN
21113312-R3
21113312-R2
21113312-R1
21113312-R0

H

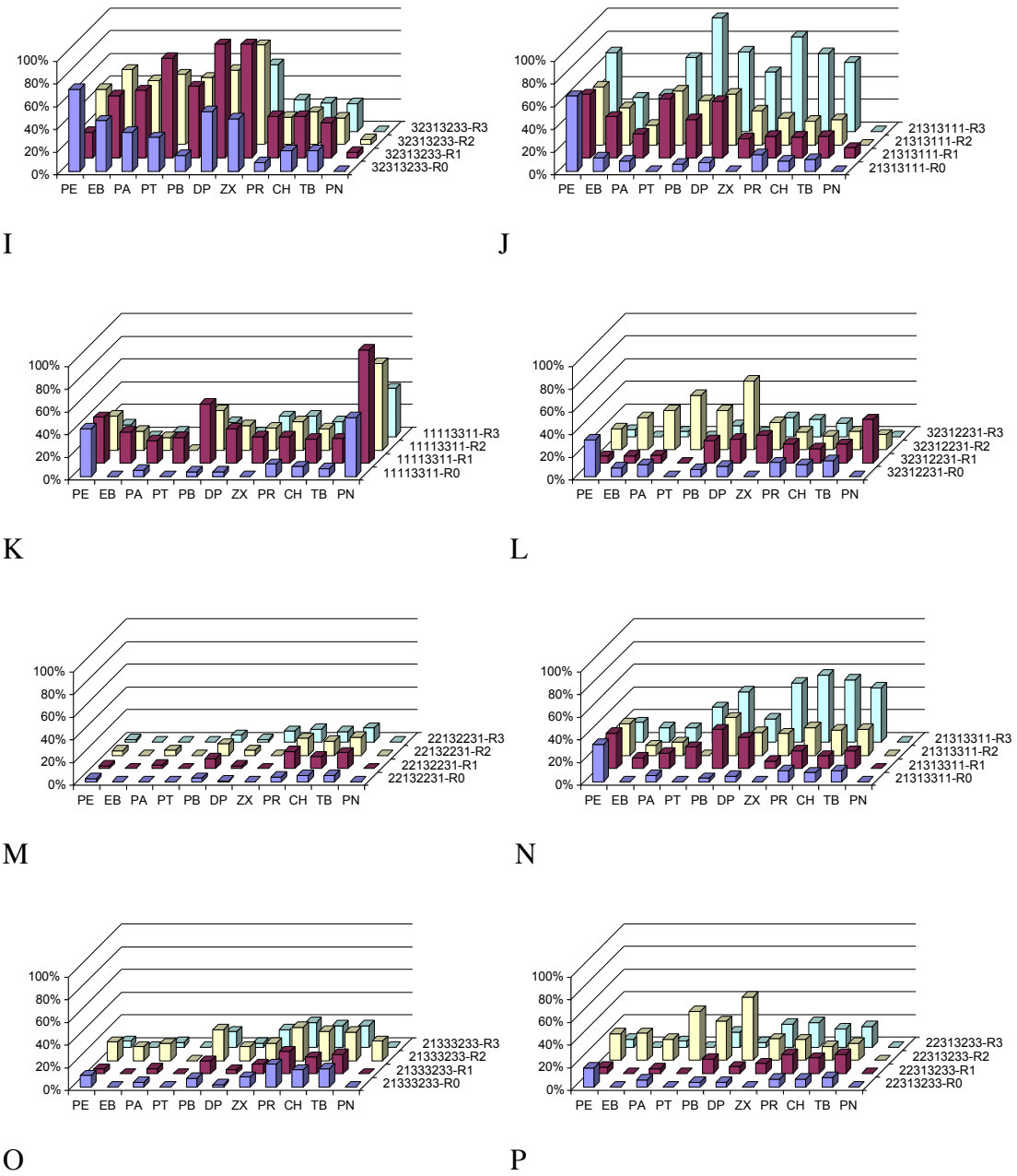


Figure 1.S2: Substrate-activity profiles of all chimeras. The columns are color coded as follows: heme domain (R0, blue), R1- (purple), R2- (yellow), R3-fusion (turquoise) protein.

2 Evolvability of Evolutionarily Young Enzymes

2.1 Abstract

Native enzymes have been highly optimized by natural evolution to perform their biological function. For this reason, improving the native activities of wild-type enzymes is challenging and often unsuccessful. Yet there are scientific and industrial applications that would benefit from an understanding of how to do this. Here, I propose that proteins with mutational neighborhoods that have never been searched by evolutionary processes are more evolvable than native proteins *of equal fitness*. I propose that homologous recombination can be used to design proteins with unexplored mutational neighborhoods because it affords the simultaneous incorporation of numerous neutral mutations. I validate this hypothesis in the context of lattice proteins, which are highly simplified models of a protein on a two-dimensional lattice. The underlying assumption of my hypothesis is that the constraints that prevent improving the activities of native enzymes are evolutionary rather than biochemical or biophysical (i.e., native enzymes are locally rather than globally optimized).

2.2 Introduction

Native enzymes are the products of millions of years of evolution. Evolutionary pressure fine-tuned their amino acid sequence to optimize biological function. This may translate to maximizing catalytic activity, resistance to high temperatures or extremely acidic environments, regioselectivity, stereospecificity, and more. As a consequence, experimental efforts to further improve phenotypic properties that underwent selection during natural evolution, such as the thermostability of an enzyme from a thermophilic organism or the catalytic activity of an enzyme on its native substrate, are often very laborious and yield small improvements. Yet, overcoming these difficulties could potentially have a tremendous impact on certain scientific applications. As an example, cellulases, a class of enzymes the catalyze the hydrolysis of cellulose to sugar, could play

a significant role in the development of an environmentally friendly alternative to gasoline and the attenuation of the energy crisis, but their specific activity is too low [51,52]. Protein engineers have devoted much effort to improving the activities of these enzymes without significant success.

In general we do not know whether the constraints that prevent improving the native activities of wild-type enzymes are physical or evolutionary. In some cases natural evolution has driven native enzymes to be so efficient that they are binding substrate and releasing product as fast as diffusion allows. These enzymes are globally optimized and cannot be engineered to perform better. In most cases, however, there is no evidence of physical limitations constraining the activities of native enzymes. In fact it is not unlikely that many natural enzymes are only locally optimized (i.e., none of the possible single mutational steps lead to an increase in fitness despite the existence of better enzymes) and need many amino acid substitutions to escape the local optima.

In nature recombination may have aided proteins escape local maxima of the fitness landscape (fitness as a function of sequence) by introducing many homologous mutations to which proteins are highly tolerant. With data from chimeric and randomly mutated ß-lactamases, Drummond et al. [1] showed that recombination is much more conservative than random mutation, leading to a probability of folding and retaining function that is many orders of magnitude greater at the highest mutation levels. In fact, Heinzelman et al. recently designed a chimeric library of cellulases containing members with wild-type levels of cellulolytic activity and over 50 mutations relative to their closest parent [53]. Before them, others were able to achieve similar results with ß-lactamases and P450s [3,5]. The dozens of neutral mutations afforded by recombination may allow protein engineers to bypass the local maxima of native enzymes.

I propose that chimeras, on average, are more evolvable than their parents because evolutionary processes have not searched their mutational neighborhood. This argument trivially holds true for chimeras that are less fit than their parents, but also applies to chimeras that are as fit as their parents. Here, an enzyme is evolvable in the sense that beneficial mutations can be found in its mutational neighborhood. The basic intuition is that the probability of finding beneficial mutations is higher in regions of sequence space that have not already been searched by evolution than in regions that have. Note that this

argument does not require that native enzymes be *strictly* locally maximized. Rather, it only requires that their mutational neighborhood be explored by evolution. The underlying assumption of this argument is that native enzymes are not globally optimized.

Since the building blocks of chimeras are derived from native enzymes, it is unclear whether their mutational neighborhood is effectively unexplored. A mutation is effectively unexplored when the contribution to fitness that it makes in a chimeric background is different from the contribution it makes in a parental background. This occurs when the contribution depends on the amino acid identities of other residues. If it depends on one other residue, the pair forms a second order interaction. If it depends on two other residues the triplet forms a third order interaction and so on. As suggested by Figure 2.1A, mutations must be recruited into locally non-native environments to make different contributions to fitness in a chimera versus a parent, unless they interact with distal residues.

The crossovers of recombination can disrupt native interactions and form new non-native interactions. When a residue interacts with a single other residue, the formation of a new interaction does not grant access to effectively unexplored mutations. This is because all pair-wise combinations of amino acids that are accessible to the chimera are also accessible to one of their parents (Figure 2.1B). However, when a mutation occurs in a network of three or more interacting residues then, provided the network was disrupted by the crossovers of recombination, chimeras can gain access to combinations of amino acids that are not accessible to their parents (Figure 2.1B). In order for chimeras to be more evolvable than their native parents, there must exist mutations that are beneficial in the background of the former but not in that of the latter. This can occur only when a mutation is recruited into a network of three or more interacting residues that was disrupted by the crossovers of recombination. Thus, the neighborhood of chimeric enzymes includes effectively unexplored mutations when 1) third and higher order interactions contribute to fitness, and 2) the crossovers of recombination disrupt the interactions.

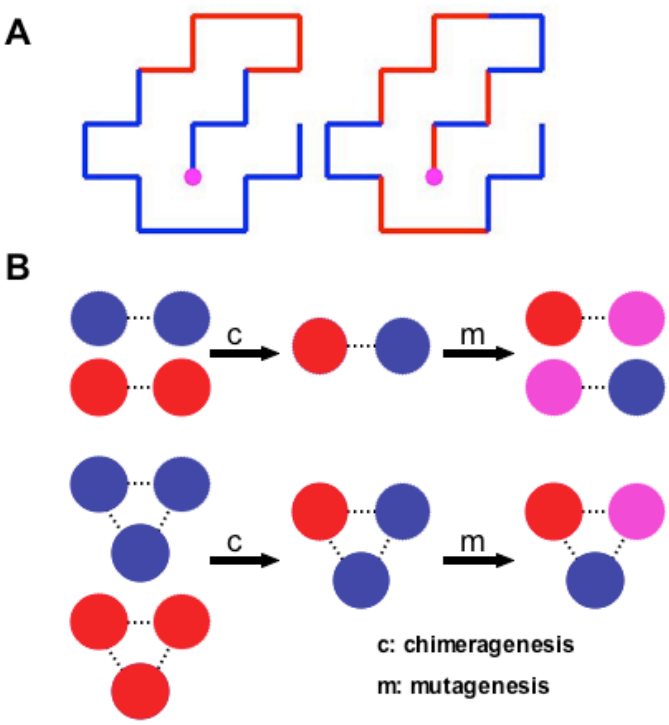


Figure 2.1: A. Effect of mutations in different chimeric lattice protein backgrounds. The red and blue segments represent fragments derived from different native proteins and the magenta circle represents a point mutation. When mutations are recruited into locally native environments (left) they interact with the same residues and make the same contributions to fitness as they would in a parental background unless they also interact with distal residues. Instead, when they are recruited into less native environments (right) they are less likely to make the same contribution to fitness in native and chimeric backgrounds. **B**. Chimeras can access combinations of residues not accessible to their parents only when third or higher order interactions are disrupted by crossovers. Red and blue circles represent residues from different parents and dotted lines represent interactions. Magenta circles represent point mutations. When a mutation occurs in the context of a pair-wise interaction that was disrupted by a crossover, it leads to a combination of amino acids that was accessible by a single mutation in the native background. Instead, when three or more residues are interacting, mutations lead to combinations of amino acids that are not attainable by single mutational steps in the native background.

Here, I use the lattice protein framework to investigate the relationship between the order of the interactions contributing to fitness and the evolvability of chimeric lattice proteins relative to native ones. Lattice proteins are highly simplified models of a protein consisting of a chain of 20 monomers on a two-dimensional lattice. Lattice proteins have been widely used to address questions of general principle related to protein folding, structure, and evolution. For example, lattice proteins have been used to propose statistical explanations to the marginal stability of real proteins [54] and the apparent anti-correlation between stability and activity [55]. In some cases, the results from lattice protein simulations have been validated by observations made on real proteins and by direct experimentation. For example, lattice protein simulations predict that sequences enriched in consensus amino acids are highly stable and robust to mutations [56,57]. Consensus mutations have been widely used to stabilize proteins [58-61] and Bloom and co-workers showed that stable enzymes are more robust to mutations [62]. A comprehensive review on lattice proteins can be found in [63].

I show that lattice proteins whose mutational neighborhoods have not been searched by evolutionary processes are more evolvable than native lattice proteins having equal fitness. I show that the mutational neighborhood of chimeric lattice proteins is effectively unexplored only when high order interactions contribute to fitness and are broken by the crossovers of recombination. Here, the evolvability of a lattice protein is evaluated according to three measures: 1) the number of improved single-mutant neighbors, 2) the greatest improvement in fitness among the improved neighbors, and 3) the fitness attained after a steepest ascent walk. A steepest ascent walk is one in which after each step the fitnesses of all the single-mutant neighbors are enumerated and the walk moves to the sequence bearing the greatest improvement in fitness until a local maximum is reached. To a first approximation, directed evolution is a steepest ascent walk.

Proteins with unexplored mutational neighborhoods can be expected to have a greater number of improved single-mutant neighbors because the probability that a mutation is beneficial given that it has never been tested by evolution (as is the case for mutations occurring in enzymes with unexplored mutational neighborhood) is higher than the probability that a mutation is beneficial given that it has been tested but not selected

by evolution (as is the case for mutations occurring in native proteins). Therefore, on average, I expect proteins with unexplored mutational neighborhoods to have access to a greater number of beneficial mutations than native proteins having equal fitness (first measure of evolvability). Likewise, strongly beneficial mutations are unlikely to be found in the neighborhood of native enzymes because if they existed evolution would have selected them. Thus, I expect proteins with unexplored mutational neighborhoods to have access to more strongly beneficial mutations than native proteins having equal fitness (second measure of evolvability). Finally, after each step of a steepest ascent walk, I expect chimeras to continue encountering more and better beneficial mutations than native proteins (for the same reasons supporting the first two measures of evolvability) and thus attain a higher fitness at the end of the walk (third measure of evolvability). This requires that the mutational neighborhood of native enzymes be searched beyond the one-mutant neighbors.

2.3 Methods

2.3.1 Lattice Proteins

The lattice proteins [55,62-66] used in the simulations are highly simplified models of a protein consisting of a chain of 20 monomers on a two-dimensional lattice that can occupy any one of 41,889,578 possible compact or non-compact conformations. The monomers can be of 20 types corresponding to the 20 amino acids. Each monomer on the lattice has four nearest-neighbor sites, of which as many as two can be occupied by nonbonded neighboring residues (three in the case of terminal residues). The energy of a target conformation C_T is given by,

$$E(C_T) = \sum_{i=1}^{20} \sum_{j=i}^{20} C_{ij}(C_T) \times \varepsilon(P_i, P_j),$$

where $C_{ij}(C_T)$ is one if residues i and j are nonbonded nearest neighbors in conformation C and zero otherwise, and $\varepsilon(P_i, P_j)$ is the interaction energy between amino acid P_i and P_j based on a widely used statistical analysis of real proteins by Miyazawa and Jernigan [67]. The free energy of folding of a lattice protein is related to the difference between

the free energy of the target conformation and the free energy of the ensemble of all other conformations,

$$\Delta G_f(C_T) = E(C_T) + T\ln\left\{Q(T) - \exp\left[\frac{-E(C_T)}{T}\right]\right\},$$

where C_T is the target conformation and $Q(T)$ is the partition function:

$$Q(T) = \sum_{\{C_i\}} \exp\left[\frac{-E(C_i)}{T}\right].$$

All simulations were performed at a reduced temperature of $T = 1.0$. Proteins are defined to be folded if their free energy of folding is less than or equal to zero.

For those proteins that stably fold activity is modeled as the binding energy (BE) of a small rigid peptide ligand to the active site of a folded lattice protein. The basic idea is that if a protein folds with at least the minimal required stability, then evolution selects for a protein's function and is indifferent to the actual stability. This model of lattice protein folding and function has been used by others to investigate the evolvability of new functions in stable proteins [62] and to investigate the correlation between activity and stability [55]. The models of ligand binding in the present study are different from those reported previously because they include high order contributions (up to fifth order).

In the simplest model, the BE (BE and activity are used interchangeably; fitness refers to the BE and/or the ΔG_f) is the summation of adjacent protein-ligand residue interactions as shown by the red dotted lines in Figure 2.2 (model 1). In this model, protein residues make independent contributions to the BE (first order), and higher order contributions to fitness are introduced solely by the requirement that $\Delta G_f \leq 0$ in order for the lattice protein to bind the ligand. In the remaining models, the BE is the summation of first order protein-ligand interactions and second (model 2), third (model 3), or fifth (model 5) order intra-protein interactions.

In real proteins, residues have been observed to make both first and high order contributions to the activity of enzymes. In homologous enzymes, for example, residues that are directly involved in a specific function are often conserved despite the great

sequence diversity that can be observed in their vicinity. Their contributions are thus largely independent of sequence and largely first order. The conserved cysteine that is responsible for the proper positioning of the heme in the P450 family represents a good example of this. At the same time, high order contributions to fitness have frequently been reported, in particular when catalytic activity is being studied rather than stability [68-71]. Thus, my models include both first and high order contributions to the BE. The BE function of models 2, 3, and 5 is composed of a first order term and a high order term. Since each first order protein-ligand interaction is, on average, equal in magnitude to each high order interaction (see below), the relative number of first and high order interactions indicates the relative contribution that each of these terms makes to the BE. Thus, the contributions to the BE from high order interactions are 33%, 33%, and 50% in models 2, 3, and 5 respectively. Model 5 is composed of six first order interactions and six fifth order interactions involving most of the 20 residues of the lattice protein. The frequency of high order interactions in this model is very high and not intended to depict realistic models of ligand binding. Instead, this model was included to elucidate a qualitative trend.

In the second order model, interactions were assigned manually, and protein-protein interactions involve only active site residues. In the third order model, all adjacent protein-ligand residues are interacting, and third order interactions involve one active site residue and two residues randomly chosen *in silico* with a probability inversely proportional to their distance from the active site residue. Likewise, in the fifth order model, all adjacent protein-ligand residues are interacting and the fifth order interactions involve one active site residue and four residues randomly chosen *in silico* with a probability inversely proportional to their distance from the active site residue. There are numerous reports of non-active site mutations that alter catalytic activity in real enzymes [72-75] and for this reason I included long-range interactions in some of my models.

All binary interactions (protein-ligand and protein-protein) are those proposed by Miyazawa and Jernigan [67] (Table 5). Since Miyazawa and Jernigan limited their analysis to pair-wise interaction energies, the third and fifth order interactions were selected randomly from the same distribution [67] (Table 5). This was done to ensure that, on average, the contribution of each third and fifth order interaction was equal in

magnitude to the contribution of a single pair-wise interaction between the lattice-protein and the ligand. This is important to ensure that the high order contributions do not overwhelm the first order protein-ligand contributions and viceversa. Besides the relative magnitudes of the first and high order interaction energies, I do not expect the exact nature of their distributions to affect the qualitative features of my results.

Unlike the pair-wise interactions, the third and fifth order interactions are position dependent (i.e., the interaction Met-Arg-Tyr is different from Arg-Met-Tyr). The Miyazawa and Jernigan potentials reflect the average energetic contributions that contacting amino acid pairs make to real protein stability and they inherently do not depend on position. However, since the third and fifth order interactions involve distal residues in the lattice protein, position-dependant energies are more appropriate and more realistic.

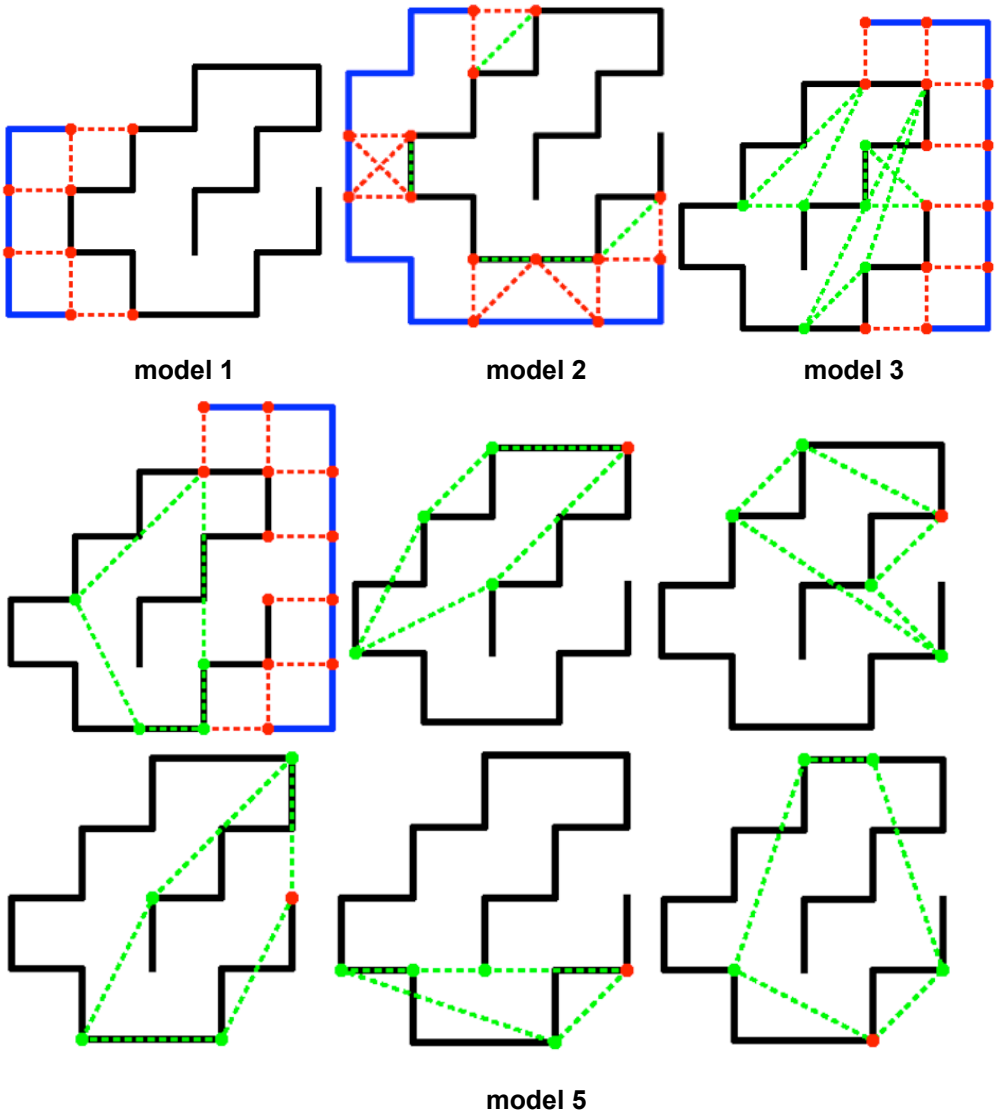


Figure 2.2: Lattice protein models. Black solid line: lattice protein; blue solid line: ligand; red dotted lines: first order protein-ligand interactions; and green dotted lines: second, third or fifth order intra-protein interactions. Note that in model 1 the BE is a purely additive function and the ruggedness of the fitness landscape arises from the folding constraint. The BE of models 2, 3, and 5, instead, has second, third, and fifth order contributions. Model 5 is composed of six first order and six fifth order interactions. Each fifth order interaction is shown separately so they can be visualized more clearly. In every model, the BE energy is equal to the summation of all first and high order interactions.

2.3.2 Evolutionary Simulations

A random search was performed to identify a set of 1,143 sequences, S_{ran}, that stably fold into the native structure shown in Figure 2.2. To create a set of "native" sequences, S_{nat}, each of these 1,143 random sequences was evolved seven independent times for 1,000 generations for improved BE with the ligands shown in Figure 2.2 (FGLLGD for model 1, AMHYRTFGLLGDTE for model 2, and LGNVAELLK for model 3 and 5) (a total of $1{,}143 \times 7 = 8{,}001$ evolutionary runs for each model). The qualitative features of my results were not found to depend on the sequence of ligands (data not shown). For each evolutionary run, the starting population is composed of ten identical lattice protein sequences equal to one of the random sequences. At each generation ten offspring are produced. These offspring are identical copies of existing members of the population, with the probability of a copy being from a member of the population being proportional to the quantity, e^{-BE}, associated with that member of the population. These new sequences then replace the existing sequences in the population (note that the population size is kept constant). All members of the population are then mutated with a per site mutation rate of 0.005. The final BE is that of the most abundant lattice protein sequence in the population after 1,000 generations. Native real enzymes are probably not strictly locally maximized. Rather they may have access to a few mildly beneficial mutations. Like wise, in the present work, native lattice proteins are not constrained to local maxima. This is achieved by carrying out evolutionary simulations for a fixed number of generations. As a result, some lattice proteins will be strictly locally maximized and many will not.

2.3.3 Creation of Chimeric Lattice Proteins

The sets of native sequences were used to create hundreds of chimeric families for each of the four models. One hundred families containing 50 unique chimeras were made for model 1, 283 families containing 50 unique chimeras were made for model 2, and 313 families containing 25 unique chimeras were made for model 3 and 5. Each chimeric family consists of a collection of unique chimeras derived from the same three parents. The parents of each family are selected randomly from the seven possible native sequences that evolved from the same random sequence. This was done because

homologous enzymes in nature share a common ancestor. In fact, the average difference in sequence between parents of the same family is lower than what one would expect if the two sequences were drawn at random. Chimeras were accepted into their family provided they differed by at least $n^{\min}$ residues from their closest parent. Each chimera was constructed by randomly selecting the positions and the number of crossovers (between 1 and $c^{\max}$) and then randomly selecting the parents that went into the segments defined by the crossovers. The parameters used to create chimeras for model 1, 2, 3, and 5, respectively, are: $n^{\min} = 10$ and $c^{\max} = 7$, $n^{\min} = 11$ and $c^{\max} = 7$, $n^{\min} = 2$ and $c^{\max} = 1$, and $n^{\min} = 5$ and $c^{\max} = 1$. Higher values of $n^{\min}$ and $c^{\max}$ were used in the lower order models to increase the probability that crossovers diversify the residues contributing to the BE. This is because in model 1 only four residues contribute to the BE and in model 2 all the intra-protein pair-wise interactions are adjacent in sequence (and thus less likely to be separated by crossovers than interactions that are not adjacent in sequence).

2.4 Results

2.4.1 Proof of Principle: Lattice Proteins with Unexplored Mutational Neighborhoods are more Evolvable than their Native Counterparts

To obtain a proof of principle that lattice proteins whose mutational neighborhood has not been searched by evolutionary processes can be more evolvable than native lattice proteins *of equal fitness*, the evolvabilities of the random proteins in S_{ran} were compared to those of the native proteins in S_{nat} having equal fitness. The native proteins are the products of 1,000 generations of evolution. Their mutational neighborhood has thus been searched by evolutionary processes. On the other hand, the random sequences were generated randomly and their mutational neighborhood is entirely unexplored. Two sets, s_{ran} and s_{nat}, were extracted from S_{ran} and S_{nat}, respectively, such that s_{ran} and s_{nat} have indistinguishable distributions of free energies and BEs (please refer to the next section for the details on how this is done). It is important to control for the free energy of folding because sequences having greater stability can tolerate a greater number of destabilizing mutations and are thus more evolvable [62]. It was not possible to generate s_{ran} and s_{nat} for models 3 and 5 because the random sequences had considerably worse

BEs than the native ones. This is to be expected for more complex fitness functions because it becomes less likely to randomly generate highly fit sequences. In the next section, models 3 and 5 will be used to generate chimeric lattice sequences, which, unlike random sequences will exhibit BEs comparable to their native parents. The cumulative distribution functions (CDF) of the free energies and BEs of s_{ran} and s_{nat} are shown to be statistically indistinguishable in Supplementary Figure 2.S1 and their average values are summarized in Supplementary Table 2.S1.

Here, the evolvability of a lattice protein is evaluated according to three measures: 1) the number of improved single-mutant neighbors, 2) the greatest improvement in BE among the improved neighbors, and 3) the BE attained after a steepest ascent walk. A steepest ascent walk is one in which after each step the fitness values of the $19 \times 20 = 380$ single-mutant neighbors are enumerated and the walk moves to the sequence bearing the greatest improvement in BE until a local maximum is reached. The stabilities and BEs of all the single-mutant neighbors of the members of s_{ran} and s_{nat} (consisting of 380 mutants for each sequence) were characterized. The number of neighbors that stably fold and exhibit an improvement with respect to the BE was determined for each sequence in s_{ran} and s_{nat} to compare their evolvabilities according to the first measure of evolvability. Following this calculation, the greatest increment in BE among the neighbors exhibiting improvement was determined for each sequence in s_{ran} and s_{nat} to compare their evolvabilities according to the second measure of evolvability. Finally, each sequence in s_{ran} and s_{nat} was subjected to a steepest ascent walk to compare their evolvabilities according to the last measure of evolvability. The CDFs of the number of improved neighbors, the greatest improvement in BE among the improved neighbors, and the BEs attained after a steepest ascent walk of the sequences in s_{ran} and s_{nat} are shown for each model in Figure 2.3. Their average values are summarized in Supplementary Table 2.S1.

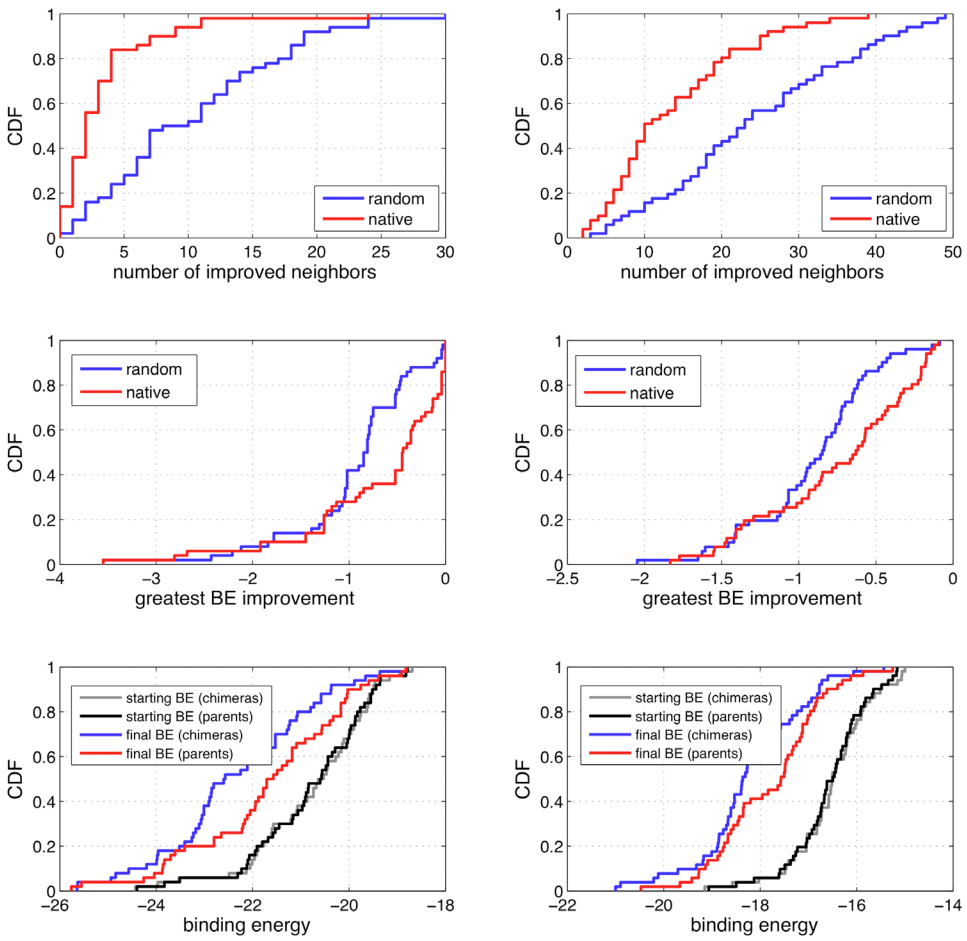


Figure 2.3: CDF of the number of improved neighbors (top), the greatest BE improvement among the improved neighbors (center), and the BE attained after a steepest ascent walk (bottom) (the black (s_{nat}) and grey (s_{ran}) lines represent the starting BEs before the steepest ascent walk and are the same distributions shown in the right panels of Supplementary Figure 2.S1). Left panels: model 1 ($p = 10^{-8}$; $p = 10^{-4}$; $p = 0.03$), and right panels: model 2 ($p = 10^{-4}$; $p = 0.02$, $p = 0.03$). The p-values here and elsewhere in this chapter are based on the two-sample Kolmogorov-Smirnov test [76] and represent the probabilities that the CDFs associated with s_{ran} and s_{nat} would look the way they do if they were drawn from identical distributions. Thus, p-values that are close to zero indicate that the distributions are statistically different.

Figure 2.3 shows that in both models random lattice proteins are more evolvable than native ones of equal fitness according to all three measures of evolvability. Random lattice proteins have access to more and better beneficial mutations and can attain lower BEs after a steepest ascent walk than native lattice proteins of equal fitness. For each steepest ascent walk, the number of improved neighbors after each step of the walk and the total number of steps taken were recorded. As anticipated, after each step of the walk random proteins continue to encounter a greater number of improved neighbors than native proteins and can thus take a greater number of steps before reaching a local maximum (data not shown). Presumably, this is because evolution searched the mutational neighborhood of the native proteins beyond the one one-mutant neighbors. This allows them to walk, on average, to lower BEs than native lattice proteins.

The results obtained in this section serve as a proof of principle that sequences with unexplored mutational neighborhoods can be more evolvable than native sequences of equal fitness.

2.4.2 Chimeric Lattice Proteins are more Evolvable than their Native Lattice Proteins when their Mutational Neighborhood is Effectively Unexplored

The building blocks of chimeric lattice proteins are derived from native lattice proteins. Thus, unlike the random lattice proteins, it is unclear whether their mutational neighborhood is effectively unexplored by evolutionary processes. As described in the introduction, mutations occurring in chimeric backgrounds are effectively unexplored when they occur in a network of three or more interacting residues that are disrupted by the crossovers of recombination. To characterize the dependence of the evolvabilities of chimeric lattice proteins (relative to those of native lattice proteins) on the order of the interactions contributing to the BE, BE models composed of interactions ranging from first to fifth order were studied and compared.

The free energies of folding and the BEs of the chimeric lattice proteins were calculated and are shown in Figure 2.4 for the chimeras that fold with $\Delta G_f \leq 0$. Figure 2.4 shows that, as the order of the interactions in the four different models increases, the

average BE of the chimeras relative to that of their parents worsens. This is consistent with experimental observations made on real chimeras that show that, on average, newly formed interactions are deleterious to function [77]. Therefore, chimeras will suffer a greater loss in fitness when the opportunity to form new interactions is higher.

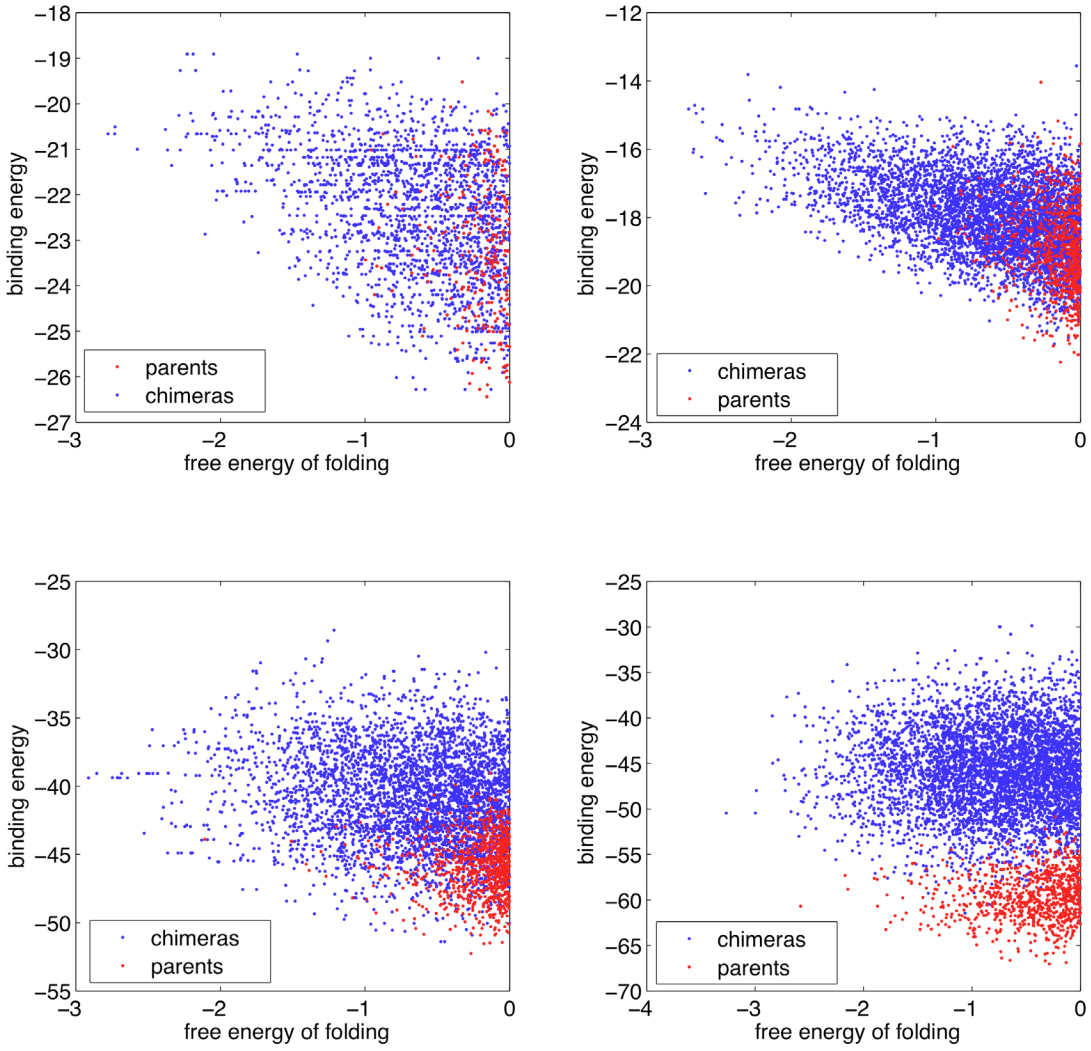


Figure 2.4: BEs and free energies of folding of the chimeras and native lattice proteins. Each blue dot represents a chimera and each red dot represents a native protein.

The chimeras of model 2, 3, and 5 are trivially more evolvable than their parents because, on average, they have lower BEs. In order to compare the evolvabilities of parental and chimeric lattice proteins having equal fitness, two subsets, s_{chi} and s_{nat}, were extracted from the full set of chimeric and native lattice proteins, S_{chi} and S_{nat},

respectively, such that s_{chi} and s_{nat} have indistinguishable distributions of free energies of folding and BEs. This was done in the following way. For each point representing a chimera in Figure 2.4, a single point representing a parent from the set of points lying within a cutoff radius of that chimera was randomly chosen and added to s_{nat}. The chosen parent was then removed from S_{nat} to ensure that it would not be selected again for a different chimera. This guarantees that s_{chi} and s_{nat} have the same size and contain only unique sequences. The resulting parents and chimeras in s_{chi} and s_{nat} are shown in Supplementary Figure 2.S2. The length of the cutoff radius was chosen in such a way that the distributions of BEs and free energies of folding associated with s_{chi} and s_{nat} were statistically indistinguishable according to the two-sample Kolmogorov-Smirnov test [76]. The sizes of s_{chi} and s_{nat} were 144, 473, 223, and 75 for models 1, 2, 3, and 5 respectively. As long as the distributions of binding and free energies remain indistinguishable, the qualitative nature of my results is not affected by variations of the cutoff radius. The CDFs of the BEs and free energies of folding associated with s_{chi} and s_{nat} are shown to be statistically indistinguishable in Supplementary Figure 2.S3.

The evolvabilites of the lattice proteins in s_{chi} and s_{nat} were determined in the same way described in the previous section. The CDFs of the number of improved neighbors and the CDFs of the greatest improvement in BE among the improved neighbors are shown for each model in Figure 2.5. Their average values are summarized in Supplementary Table 2.S2.

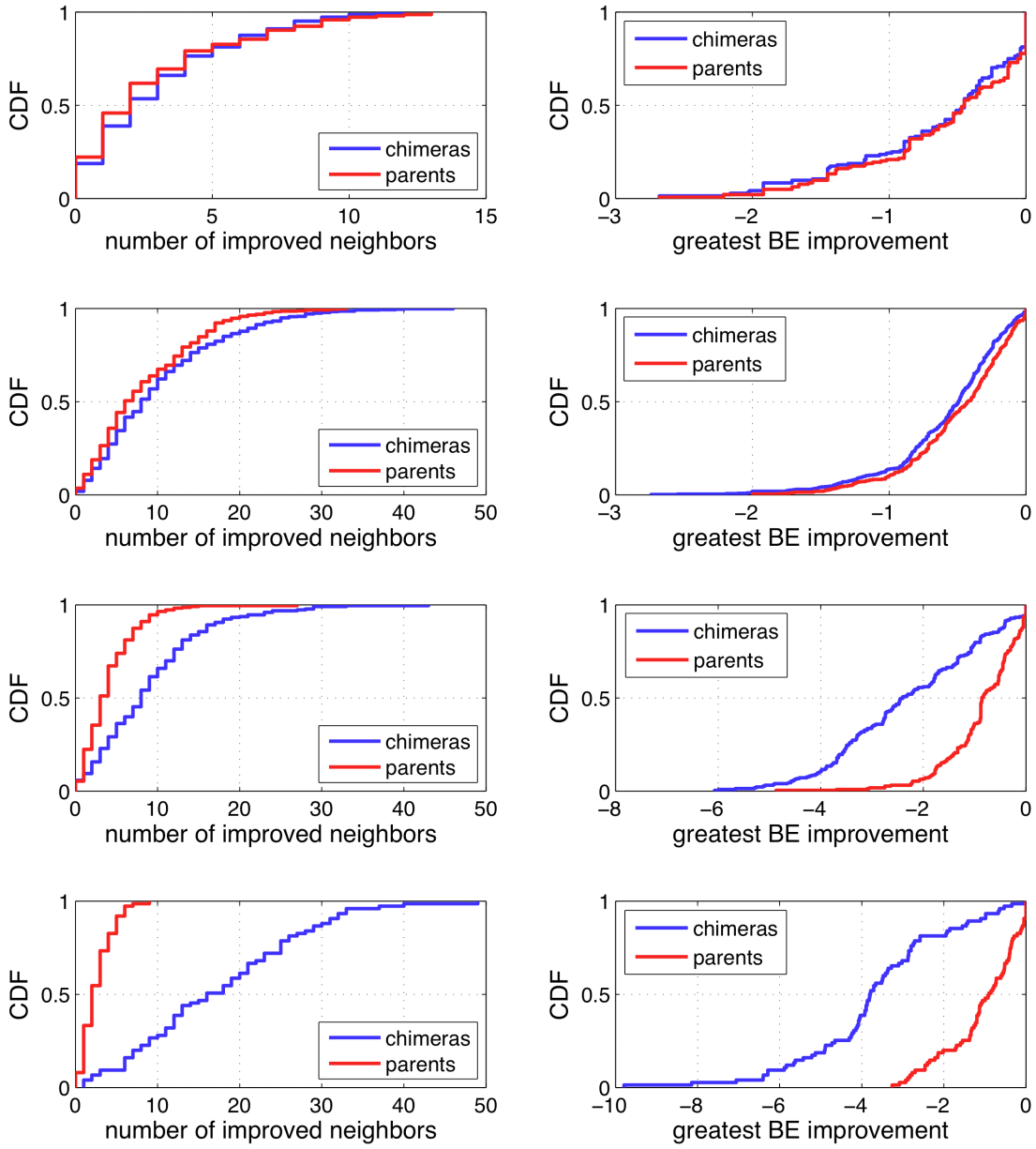


Figure 2.5: CDF of the number of improved neighbors (left) and the greatest BE improvement (right) among the improved neighbors of sequences in s_{chi} and s_{nat}. First row: model 1 ($p = 0.7$; $p = 0.7$); second row: model 2 ($p = 0.006$; $p = 0.006$); third row: model 3 ($p = 10^{-18}$; $p = 10^{-26}$); last row: model 5 ($p = 10^{-23}$; $p = 10^{-18}$).

It is immediately apparent from Figure 2.5 that the differences between the evolvabilities of chimeric and native lattice proteins increase as the order of the

interactions that contribute to the BE increases. In the case of model 1, chimeric and native lattice proteins exhibit indistinguishable evolvabilities. This is not surprising because residues make independent contributions to the BE. Any mutations that are beneficial in the context of a chimera are also beneficial in the context of a native parent. Thus, evolution has effectively searched their mutational neighborhood and they cannot be more evolvable than native proteins. In contrast, in models 3 and 5, the chimeras are significantly more evolvable than native proteins. When a mutation occurs in a network of three or more interacting residues then, provided the network is disrupted by the crossovers of recombination, chimeras can gain access to combinations of amino acids that are not accessible to their parents (Figure 2.1B). Thus, mutations occurring at these sites are effectively unexplored by evolution and chimeras can exhibit greater evolvability than their native counterparts. In model 2 chimeras are more evolvable than the native proteins but the differences are substantially less pronounced. Mutations occurring in networks of only two interacting residues lead to combinations of amino acids that are already accessible to the parents (Figure 2.1B). Therefore, in order for them to make beneficial contributions in the chimeric but not the parental backgrounds, they must occur in the context of a non-native interaction (formed by recombination) that is deleterious with respect to the native interaction. However, when too many deleterious interactions are formed by recombination the chimera will suffer a significant loss in BE. Since this study is only concerned with the evolvabilities of chimeras having comparable fitnesses as their native counterparts, such a chimera would not be included in s_{chi}. Therefore, as verified by the results, chimeras from model 2 are not expected to be significantly more evolvable than native proteins.

Each of the sequences in s_{chi} and s_{nat} from models 2, 3, and 5 was subjected to a steepest ascent walk to compare their evolvabilities according to the last measure of evolvability. This analysis was not carried out on the sequences of model 1 because the evolvabilities of the chimeric and native lattice proteins were already shown to be indistinguishable based on the first two measures. The CDFs of the BEs attained after the steepest ascent walk and the number of steps taken before reaching a local maximum are shown in Figure 2.6. Their average values are summarized in Supplementary Table 2.S2.

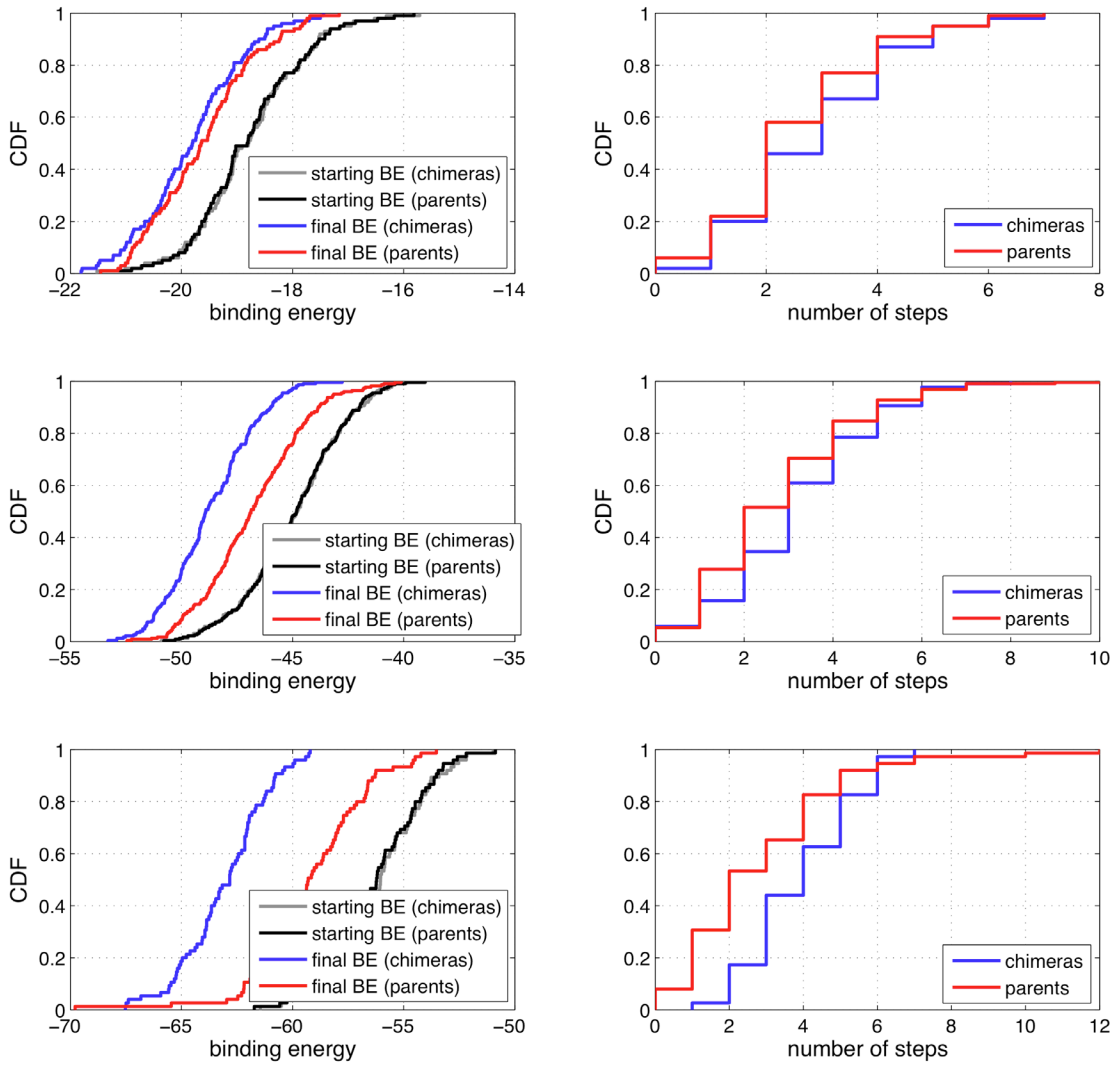


Figure 2.6: CDF of the BE attained after a steepest ascent walk by the sequences in s_{chi} and s_{nat} (left panels) (the color scheme is the same as that in Figure 2.3), and CDF of the number of steps taken before reaching a local maximum (right panels). Top: model 2 (p = 0.34; p = 0.44); middle: model 3 ($p = 10^{-12}$; p = 0.003); bottom: model 5 ($p = 10^{-16}$; $p = 10^{-5}$).

The left panels of Figure 2.6 show that when chimeras and native proteins having equal starting BEs and free energies of folding are subjected to a steepest ascent walk, the chimeras attain substantially lower BEs than native lattice proteins provided the BE model includes high order contributions. Likewise, chimeras, on average, take a greater number of steps before attaining a local optimum (right panels) and encounter a greater

number of improved neighbors after each step of the walk than native lattice proteins (data not shown). Presumably, this is because evolution has searched the mutational neighborhood of native enzymes beyond the one-mutant neighbors. Thus, provided the model of BE is composed of high order terms, just like random lattice proteins, chimeric lattice proteins are more evolvable than their native counterparts according to all three measures of evolvability. This is consistent with the hypothesis that chimeras are more evolvable than native enzymes because their mutational neighborhood is unexplored by evolutionary processes, but that their mutational neighborhood is only effectively unexplored when high order interactions contribute to fitness. Since random sequences are not composed of native residues, their mutational neighborhood is effectively unexplored independently of the order of the interactions contributing to their fitness.

I have shown that chimeras are more evolvable than their parents when third or higher order interactions contribute to the BE. However, as suggested by Figure 2.1B, mutations can lead to novel combinations of amino acids only when they occur in a network of three or more interacting residues that was disrupted by recombination. Crossovers do not necessarily cut though interacting residues. For this reason, the chimeras in s_{chi} of models 3 and 5 that do not have any non-native interactions should not exhibit greater evolvability than their native counterparts. To illustrate this, the number of non-native interactions was determined for each chimera in s_{chi}. Forty-one out of the 223 sequences in s_{chi} of model 3 do not have any non-native third order interactions. The evolvabilities of these 41 sequences were compared to those of the native sequences according to the usual three measures of evolvabilty. The same comparison was made between sequences in s_{chi} having a single non-native interaction and the native sequences. The results are shown in Figure 2.7. As anticipated, the evolvabilities of chimeras in model 3 that do not have non-native interactions are indistinguishable from those of the native proteins. Instead, chimeras having a single non-native interaction are significantly more evolvable than native parents according to all three measures of evolvability. This analysis was not possible in the case of model 5 because all the chimeras in s_{chi} have at least one non-native interaction. These results show that the crossovers of recombination must break existing high order interactions in order for

chimeras to gain access to beneficial mutations that are not beneficial in the context of native enzymes.

In summary, chimeric lattice proteins can be more evolvable than native lattice proteins provided, 1) third or higher order interactions contribute to fitness, and 2) the crossovers of recombination disrupt the native interactions and form new, non-native ones.

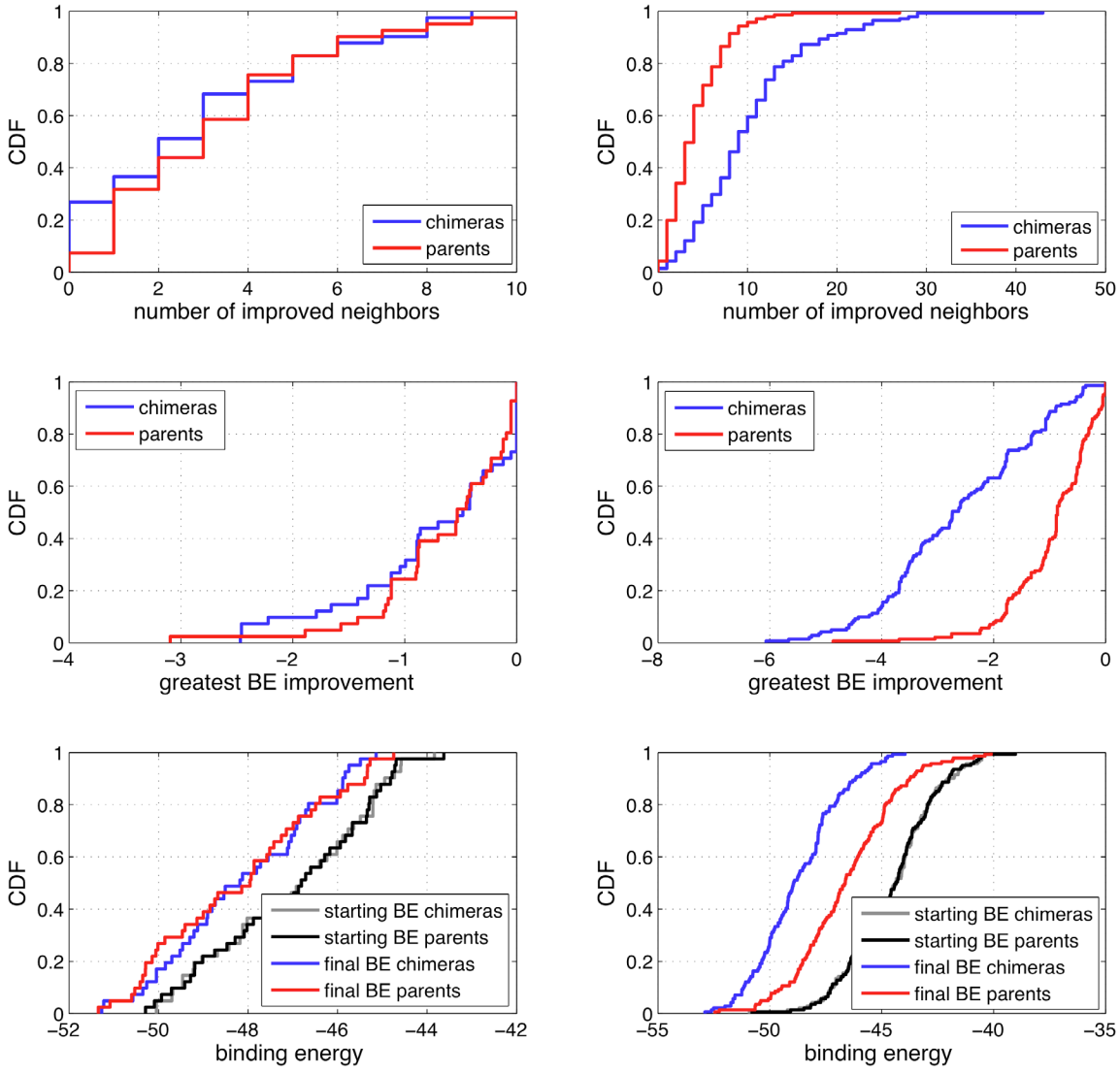


Figure 2.7: Comparison of the evolvabilities of the native sequences in s_{evo} and the chimeric ones in s_{chi} having either zero (left) or one (right) non-native third order interaction(s) (this analysis is based on model 3). As usual the sets s_{chi} and s_{evo} have indistinguishable distributions of free energies and BEs (data not shown). Despite the large differences in evolvabilities among the complete sets of sequences in s_{chi} and s_{evo} of model 3 ($p = 10^{-18}, p = 10^{-26}, p = 10^{-12}$), the evolvabilities become indistinguishable ($p = 0.38, p = 0.38, p = 0.90$) when only chimeric sequences with no broken interactions are allowed in s_{chi}. A single non-native interaction is sufficient to grant chimeras greater evolvabilities with respect to their parents ($p = 10^{-16}$; $p = 10^{-22}$; 10^{-12}).

2.5 Discussion

The present work proposes that chimeric proteins with unexplored mutational neighborhoods are more evolvable then native ones of equal fitness and validates this hypothesis in the context of four different lattice protein fitness landscapes. Here, the evolvability of a lattice protein is evaluated according to three measures: 1) the number of improved single-mutant neighbors, 2) the greatest improvement in BE among the improved neighbors, and 3) the BE attained after a steepest ascent walk. I show that chimeric lattice proteins are more evolvable than native lattice proteins when contributions to the BE include third or higher order terms and the crossovers of recombination disrupt existing native interactions and replace them with new, non-native ones. When the contributions to the BE energy include only first or second order terms the evolvabilities of the chimeras are indistinguishable from those of the native proteins. However, unlike chimeric lattice proteins, random lattice proteins do exhibit significantly greater evolvabilities relative to native ones with these low order models. These results support the argument that non-native sequences must have an effectively unexplored mutational neighborhood to be more evolvable than native ones. Since random sequences are not composed of native residues, their mutational neighborhood is effectively unexplored independently of the order of the interactions contributing to their fitness.

These results have practical relevance when it is desirable to improve the native activities of real enzymes and there is evidence that there exist no beneficial mutations in their near neighborhood. Enzymes with cellulolytic activity represent a relevant example. Efforts to improve their catalytic efficiency on cellulose by directed evolution have not lead to significant success. Recently, Heinzelman et al. [53] constructed a chimeric library of celluases using structure-guided recombination. Several members of this library have over 50 mutations relative to their closest parent and have been shown to exhibit wild-type levels of specific activity on cellulose. Provided the parent cellulases are not globally optimized, my results suggest that the chimeras may represent a better starting point for directed evolution experiments.

The purpose of this study was to elucidate a qualitative trend rather than to build realistic protein BE models. For this reason, it is important to understand the differences between the models of BE used in this work and the functions relating sequence to

activity in real proteins. In real proteins, on average, only 5-15% of residues are directly involved in binding substrates. In the lattice protein models, 20%, 40%, 60%, and 85% of residues make direct contributions to the BE in models 1, 2, 3, and 5, respectively. The greater the number of residues directly involved in function, the greater the likelihood that crossovers break functionally important interactions. Another difference between my models and real proteins is the frequency, spatial distribution, and orders of the interactions that contribute to the BE. The existence of high order interactions affecting catalytic activity in real proteins has been verified experimentally [68-70]. High order interactions are particularly common in functionally important regions of proteins [68,71]. In some of my models, however, high order interactions are frequent relative to the size of the lattice protein and are distributed throughout its structure. Therefore, the most relevant way to interpret the lattice protein models in this study is to view them as models of the functionally important regions of larger proteins. This does not imply that they should be viewed specifically as active site models. Rather, they are models of the collection of residues that contribute to function.

Future studies could possibly use larger lattice proteins to determine whether the nature of my results changes when at most 20-30% of the residues contribute to the BE and third order interactions are localized to the active site. Recall, however, that a single newly formed third order interaction is sufficient to grant chimeras greater evolvability than native lattice proteins (Figure 2.7). Thus, provided interacting residues are sufficiently spaced apart in sequence so that crossovers are likely to separate them, I do not expect the qualitative nature of my results to change significantly with larger lattice proteins and/or less interactions (provided they are at least third order).

The major limitation of this work is that the native lattice proteins used in this study exhibit much greater diversity at their functional residues than is normally observed in the active sites of real homologous enzymes. The average sequence identity at the residues that directly contribute to the BE between the parents of a chimeric family ranges from 12-30% for the different lattice protein models. Instead, real homologous enzymes that perform the same catalytic chemistry and have similar substrate specificities often have nearly perfectly conserved active sites and most of their amino acid differences are found on their surfaces. Therefore, the resulting chimeras also have

conserved active sites [3,5,53]. Instead, the average sequence identity at the functional sites between chimeric lattice proteins retaining native levels of BE and their closest parent ranges from 60-85%. This suggests that the regions of real homologous enzymes that are directly involved in function are less tolerant to change than those of the lattice proteins of the present study. Thus, recombination of real enzymes cannot create non-native interactions in the active sites. Instead, it relies on the existence of functionally significant long-range interactions between the surface and active site residues to form new interactions. Such long-range interactions have been observed in real proteins [72-75] but their frequency and relative contribution to activity are likely to be less significant than functional interactions occurring within the active site.

There are two factors that can contribute to the discrepancy between the diversity that is tolerated at the functional residues of lattice versus real proteins. The first is related to an inherent property of the BE models used in this study that may allow multiple sequences to be compatible with a high level of activity. More significantly, however, the discrepancy may indicate that the native lattice proteins of this study are less optimized relative to their fitness landscape than real native proteins are relative to theirs. A protein that is highly optimized relative to its fitness landscape is one that is almost globally optimized (i.e., better enzymes are extremely rare). I expect that diversifying the functional residues of such a protein without losing activity is extremely hard. Thus, a very reasonable explanation for the discrepancy between the sequence entropy tolerated at the functional residues of real versus native lattice proteins is that real proteins are more optimized than the native lattice proteins of the present study.

To address this issue, future work should investigate how the number of generations used in the evolutionary simulations affects the nature of my results. Increasing the number of generations will produce native lattice proteins that are more optimized relative to their fitness landscape than the ones in the present study. In this scenario, I expect the diversity of native lattice proteins to decrease relative to the diversity observed in the present study. This will lead to a decrease in the number of non-native interactions that can be formed using chimeragenesis and will reduce the likelihood that chimeric lattice proteins exhibit greater evolvability than their native counterparts. Alternatively, if the diversity of lattice proteins does not decrease

considerably, recombining highly diverse and highly optimized lattice proteins will lead to a great loss in BE, and it will not be possible to generate chimeric lattice proteins that have the same fitness as their parents like it was in this study. These arguments emphasize that, ultimately, it is the degree of optimization of native enzymes relative to their fitness landscape that determines whether chimeras can be more evolvable than their native parents. If they are highly optimized relative to their fitness landscape, then either 1) their functional residues will be conserved and it will not be possible to recombine them and generate non-native functionally relevant interactions, or 2) their functional residues will not be conserved but it will be impossible to generate non-native interactions without suffering a big loss in fitness.

These limitations are related to the underlying assumption of my hypothesis which is that native proteins are not globally optimized and that better proteins are frequent enough to be found using recombination and mutagenesis. This assumption is clearly satisfied in the present work. In general, however, we do no know whether it holds in real enzymes and only experiments can shed further light on this issue.

2.6 Acknowledgments

I would like to thank Jesse Bloom and Phil Romero for help with lattice proteins.

2.7 Supplementary Material

Table 2.S1: Summary of the average value of the free energy of folding, $\langle \Delta G_f \rangle$, binding energy, $\langle BE \rangle$, number of improved neighbors, $\langle n_{imp} \rangle$, greatest BE increments among improved neighbors $\langle \Delta BE_{\max} \rangle$, BE after a steepest ascent walk, $\langle BESA \rangle$, and number of steps taken to the nearest local maximum $\langle n_{steps} \rangle$ for the sets s_{ran} and s_{nat}.

	$\langle \Delta G_f \rangle$		$\langle BE \rangle$		$\langle n_{imp} \rangle$		$\langle \Delta BE_{\max} \rangle$		$\langle BESA \rangle$		$\langle n_{steps} \rangle$	
	s_{ran}	s_{nat}	s_{ran}	s_{nat}	s_{ran}	s_{nat}	s_{ran}	s_{nat}	s_{ran}	s_{nat}	s_{ran}	s_{nat}
Model 1	-0.26	-0.25	-20.8	-20.8	10.3	3.3	-0.95	-0.70	-22.3	-21.6	2.3	1.4
Model 2	-0.28	-0.28	-16.5	-16.6	24.3	13.5	-0.91	-0.75	-18.2	-17.8	3.5	4.2

Table 2.S2: Summary of the average value of the free energy of folding, $\langle \Delta G_f \rangle$, binding energy, $\langle BE \rangle$, number of improved neighbors, $\langle n_{imp} \rangle$, greatest BE increments among improved neighbors $\langle \Delta BE_{\max} \rangle$, BE after a steepest ascent walk, $\langle BESA \rangle$, and number of steps taken to the nearest local maximum $\langle n_{steps} \rangle$ for the sets s_{chi} and s_{nat}.

	$\langle \Delta G_f \rangle$		$\langle BE \rangle$		$\langle n_{imp} \rangle$		$\langle \Delta BE_{\max} \rangle$		$\langle BESA \rangle$		$\langle n_{steps} \rangle$	
	s_{chi}	s_{nat}	s_{chi}	s_{nat}	s_{chi}	s_{nat}	s_{chi}	s_{nat}	s_{chi}	s_{nat}	s_{chi}	s_{nat}
Model 1	-0.28	-0.30	-23.1	-23.1	3.0	2.8	-0.65	-0.59	--	--	--	--
Model 2	-0.20	-0.21	-18.8	-18.8	10.5	8.2	-0.57	-0.49	-19.8	-19.6	2.9	2.5
Model 3	-0.39	-0.40	-44.9	-44.9	9.0	4.1	-2.3	-0.90	-48.7	-46.8	3.2	2.7
Model 5	-0.64	-0.66	-56.1	-56.2	17.4	2.6	-3.70	-1.0	-63.0	-59.1	3.9	2.8

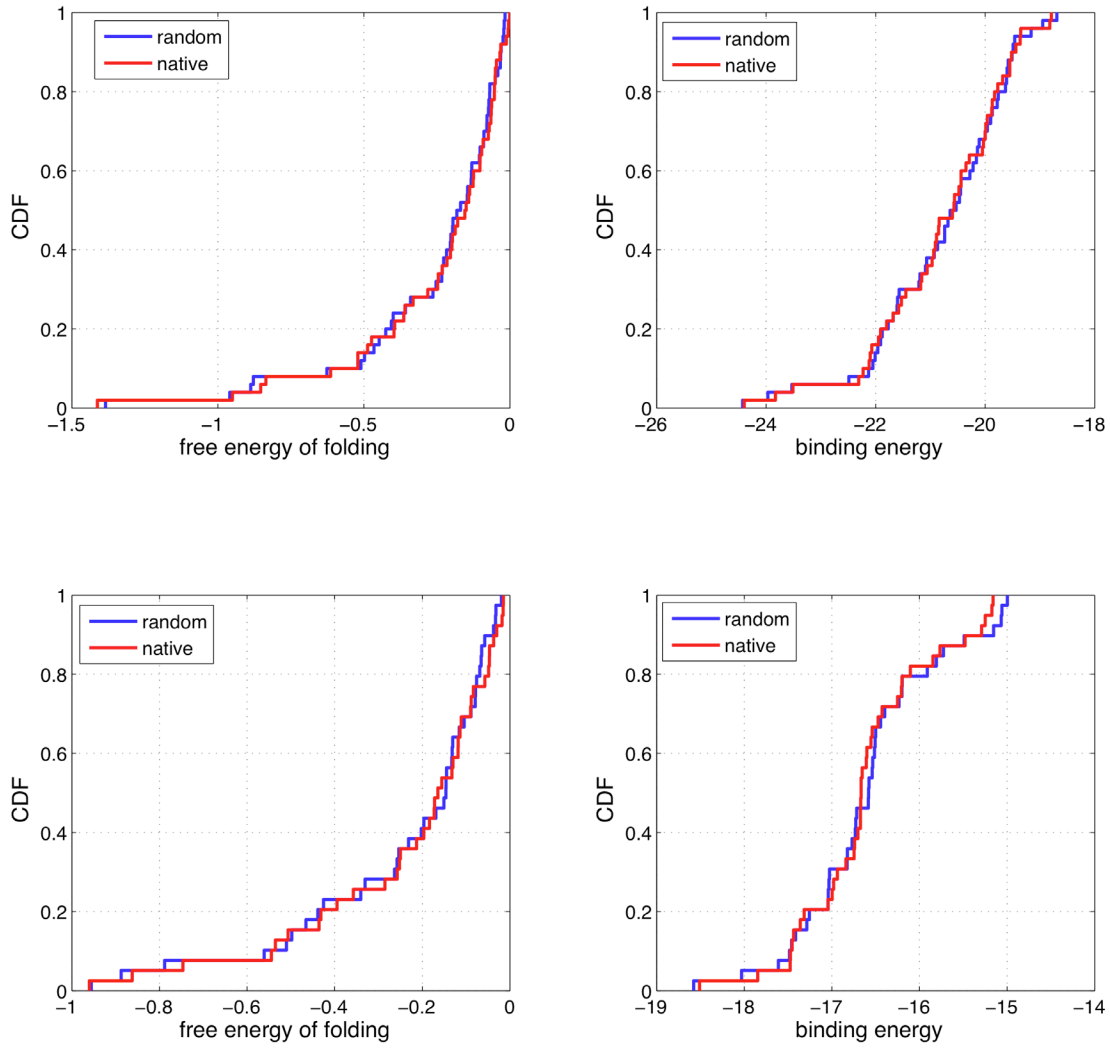


Figure 2.S1: CDF of the free energy of folding and binding energy of sequences in s_{ran} and s_{nat}. The distributions are statistically indistinguishable. Top: model 1 (p = 1.0; p = 1.0); bottom: model 2 (p = 1.0; p = 0.9).

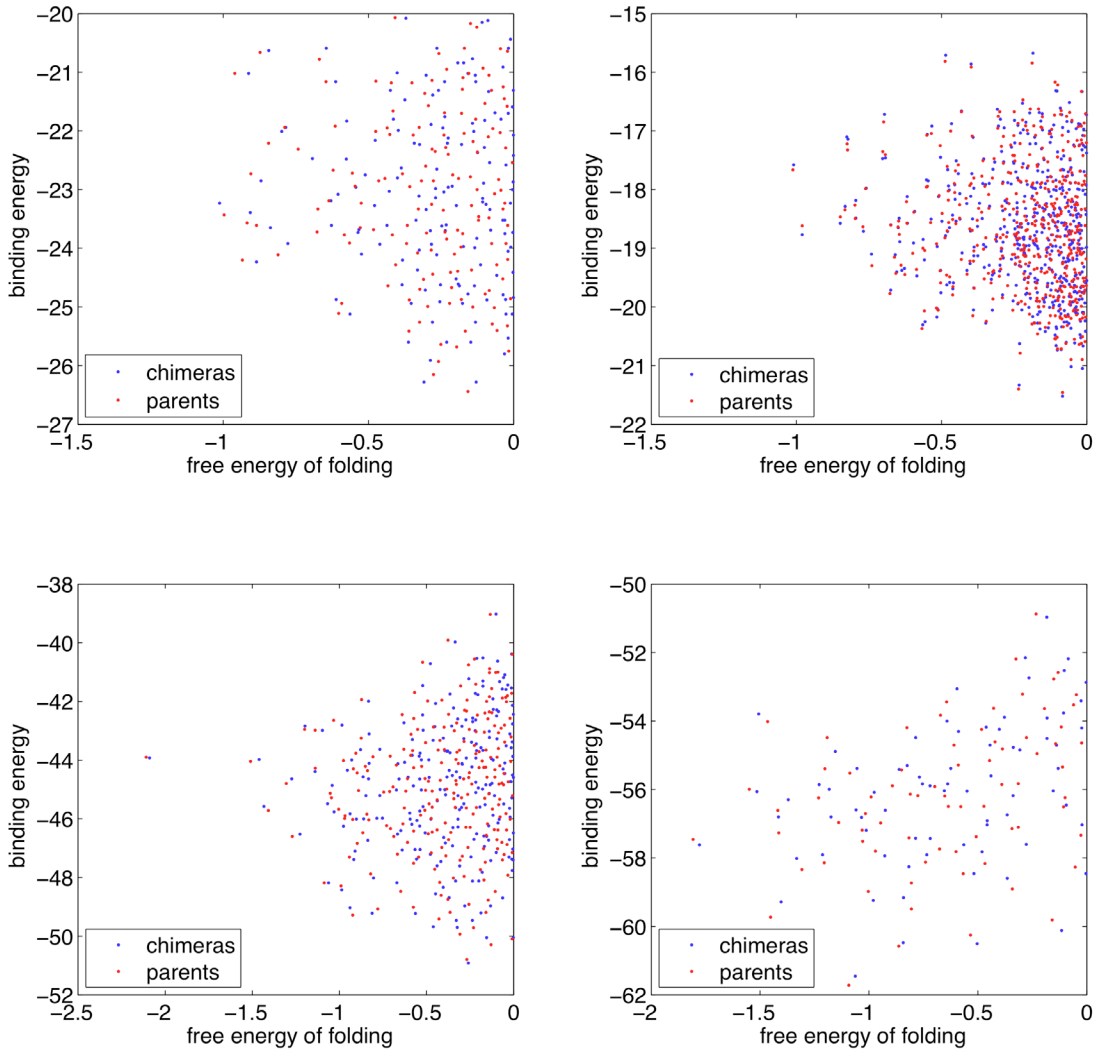


Figure 2.S2: Chimeric lattice proteins (blue dots) and native lattice proteins (red dots) in s_{chi} and s_{nat}.

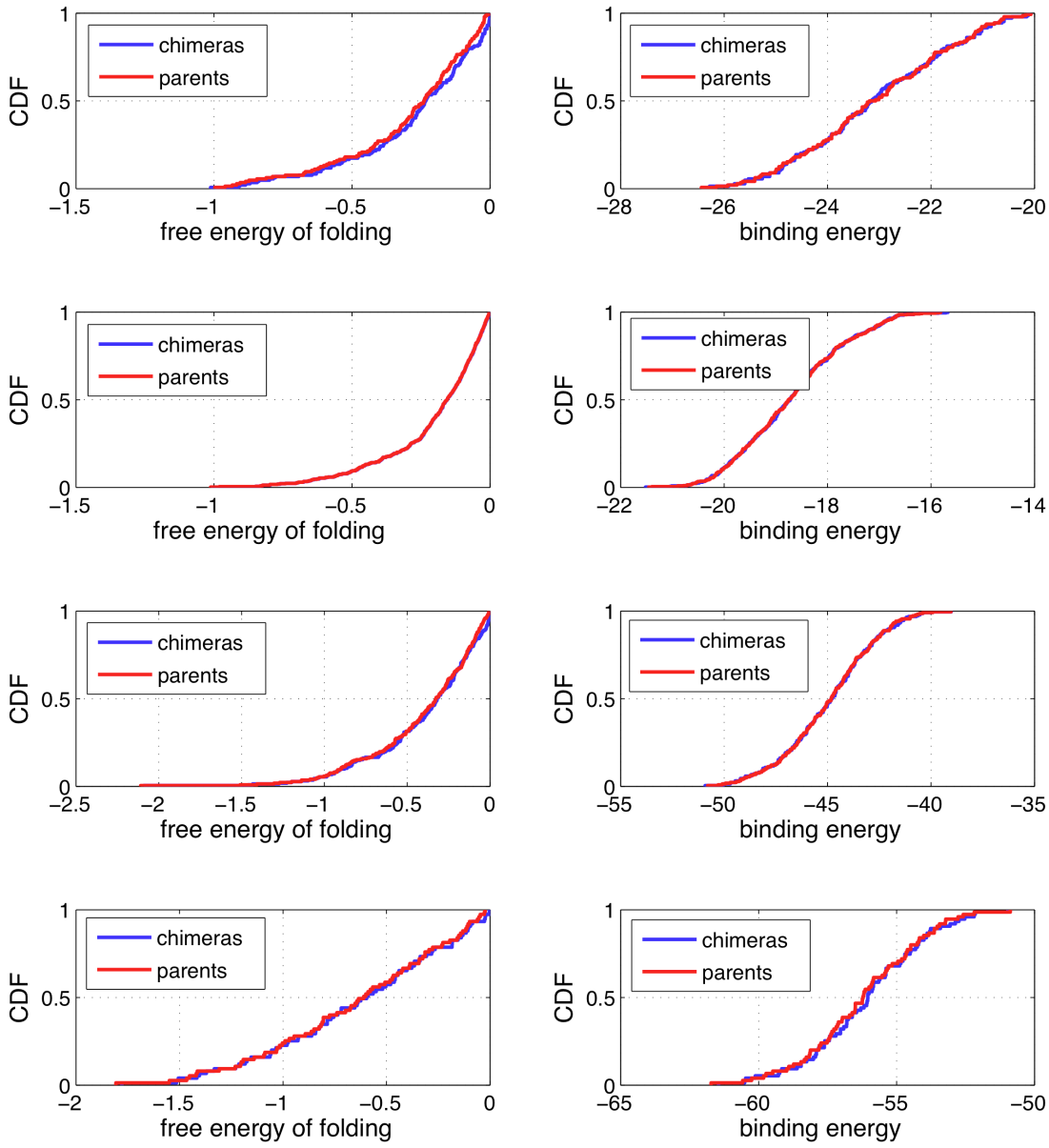


Figure 2. S3: CDF of the free energy of folding of sequences in s_{chi} and s_{nat}. The distributions are statistically indistinguishable. Top row: model 1 ($p = 0.5$; $p = 1.0$); second row: model 2 ($p = 1.0$; $p = 1.0$); third row: model 3 ($p = 0.7$; $p = 1.0$); last row: model 5 ($p = 1.0$; $p = 0.9$).

3 Evolvability of an Evolutionarily Young Chimeric Cellobiohydrolase II Derived from *Trichoderma reesei, Humicola insolens, and Chaetomium thermophilum*

3.1 Abstract

Cellulases are an important class of enzymes that catalyze the hydrolysis of cellulose to sugar. Unfortunately, they have specific activities on cellulose that are too low for the development of economically viable processes that convert biomass to sugar. Despite many efforts to further improve the specific activities of these enzymes, only minimal improvements in activity have been achieved. The present work proposes that highly mutated chimeric cellulases, assembled using homologous recombination, are more likely to have access to beneficial mutations than native ones of comparable activity because evolution has not searched their mutational neighborhood as it has in the case of native cellulases. The underlying assumption is that wild-type cellulases are locally optimized and that the constraints that hinder further improvement of their activities are evolutionary rather than physical. To test this hypothesis the mutational neighborhood of chimeric cellobiohydrolases II generated using SCHEMA structure-guided recombination was searched for beneficial mutations. Unfortunately, no mutants exhibiting a significant improvement in specific activity were found. The failure to identify improved mutants in the neighborhood of the chimeric cellulases can be attributed to 1) an unlucky choice of chimeras, 2) an assay poorly suited to identifying increments in specific activity, 3) a high degree of conservation in the functionally important regions of the cellulases in the SCHEMA library, or 4) a physical or chemical limitation to further improvements (i.e., native cellulases are globally optimized).

3.2 Introduction

Incentives to ameliorate the performance of natural cellulases are not lacking. Over 250 million motor vehicles populate the United States accounting for 28% of our energy use

and 34% of our carbon dioxide emissions [78,79]. Petroleum, derived from depleting fossil fuels, powers most of those vehicles [79], and a significant portion of that petroleum is imported from countries that have unstable political and economic ties with the United States. Producing "clean" energy domestically from renewable resources is thus critical for both securing our supply and salvaging our environment. Cellulose is the most abundant renewable resource, and a study from the Department of Energy and the Department of Agriculture estimates that the United States can produce enough biomass to supply over 30% of our current oil demand without a dramatic negative impact on food supply [78].

Cellulose degradation, however, is not easy. Cellulose is a linear condensation polymer consisting of D-anhydroglucopyranose joined together by ß-1,4-glycosidic bonds. Adjacent chains and sheets of cellulose are held together by hydrogen bonds and van der Waal's forces resulting in a stable crystalline structure of great tensile strength and low accessibility [52]. The crystalline nature of cellulose makes it difficult to break down and requires the concerted attack of a team of enzymes acting synergistically and collectively referred to as cellulases. Endoglucanases hydrolyze accessible ß-1,4-glycosidic bonds in amorphous regions of the polymer, disrupting its crystalline structure and exposing individual chain ends. Two different cellobiohydrolases-- one type working processively from the reducing end and the other from the non-reducing end of cellulose-- then attack these individual chains and break them down to cellobiose. Finally, the cellobiose units are broken down to glucose by the actions of ß-glucosidases. Celluase systems come in two flavors: complexed and non-complexed. In non-complexed systems each of these enzymes is secreted individually from the cells. In complexed systems, (typical of anaerobic microorganisms), the enzymes are grouped into a complex known as the cellulosome that remains attached to the exterior cell wall. The present work is based on non-complexed cellulases.

Grinding and pretreating naturally occurring cellulose with acids to disrupt its crystalline nature relieves some of the burden on cellulases, but introduces additional costs and waste treatment concerns. Furthermore, too much acid can damage the sugars [78]. Therefore, it is desirable to break down cellulose using more biology and less chemistry [78]. Currently, however, the bioconversion of natural cellulose to sugars by

organisms that secrete cellulases is too costly to be implemented for the large-scale production of biofuels. Strategies that will make this technology more economical include increasing the stability of cellulases at elevated temperatures and specific pH regimes and increasing their specific activity [52].

Efforts to improve these enzymes are not lacking either: scientists and engineers have been working on the cellulase system of *Trichoderma reesei (T. reesei*) for over 50 years. Since then, besides the wealth of literature describing the characterization of various natural cellulase systems, many groups have reported novel cellulases with altered pH profiles [53,80-83] or improved thermal tolerance [53,80,84-92]. For example Heinzelman et al. [53,90] recently created about 15 highly diversified chimeric cellobiohydrolases with thermostabilities higher than their most stable parent (which came from a thermophilic organism). A fairly accurate linear regression model predicts that hundreds of the designed (but uncharacterized) chimeras are thermostabilized with respect to the most stable parent.

Despite the strong incentives and the many years of research on cellulases, efforts to improve the specific activities of cellulases have probably been the least successful. Using DNA shuffling Kim et al. [93] found a *Bacillus subtilis* endoglucanase mutant containing seven mutations with a five fold increase in specific activity on carboxymethyl cellulose (CMC) relative to the wild-type. While this may appear as a significant improvement, it has been shown that cellulase activities on soluble derivatives of cellulose such as CMC are poorly correlated with the activities on the naturally occurring insoluble cellulose [52]. McCarthy et al. [94] reported a modest improvement of 31% in the hydrolysis rate of a *Thermotoga neapolitana* ß-glucosidase single mutant on cellobiose found using random mutagenesis. Voutilainen et al. [85] recently reported the rational design of a cellobiohydrolase double mutant from the thermophilic fungus *Talaromyces emersonii* with an 80% improvement in the k_{cat}/K_M with respect to the soluble cellulase substrate, 4-methylumbelliferyl-ß-D-lactoside (mulAC). However, nearly all of the 80% improvement in k_{cat}/K_M can be attributed to a decrease in K_M which translates to an increase in affinity for the non-natural substrate mulAC. In fact when they tested this mutant on microcrystalline cellulose (known commercially as Avicel) which is insoluble and has a significantly higher resemblance to naturally occurring cellulose than

mulAC, they found that the rates of hydrolysis of the mutant and the wild-type were the same. Escovar-Kousen et al. [95] reported the rational design of a single mutant of the cellulase Cel9A from *Thermobifida fusca (T. fusca)* with a 40% activity improvement on CMC and swollen cellulose (insoluble amourphous cellulose obtained using acid treatment). This, however, translated to no improvement on the more crystalline substrates, filter paper and bacterial microcrystalline cellulose (BMCC). Zhang et al. [96] studied the effects of 14 mutations in Cel6A from *T. fusca* involving six non-catalytic active site residues on a series of cellulolytic substrates varying in polymer length, crystallinity, solubility and charge. They only observed improvements on CMC. Zhang et al. [97] performed similar work on Cel6B from *T. fusca* and found a double mutant exhibiting two and three fold improvement on filter paper and swollen cellulose, respectively. They also found four mutants exhibiting up to a four-fold improvement on CMC.

The achievements on the specific activities of cellulases on insoluble cellulose substrates are very modest. The challenge is at least in part a reflection of the fact that native cellulases have already been highly optimized by natural evolution to break down cellulose. While the stability and the pH profiles of an enzyme are not necessarily selected traits-- an enzyme need only be stable enough to function at the biologically relevant temperature and pH-- the high caloric value and the natural abundance of cellulose apply a significant selective pressure on microbes for its utilization. An organism that is well adapted to cellulose utilization will thrive in any habitat [98]. In fact cellulose hydrolysis limits the rate of microbial cellulose utilization as was inferred from the observation that microbial growth rates are several fold higher on soluble sugars than on crystalline cellulose [98]. Hence, evolution may have driven native cellulases to local maxima in their fitness landscape such that their mutational neighborhood does not contain any beneficial mutations, making directed evolution or any low-mutagenesis engineering approach (such as the rational design of a few active site residues) unlikely to succeed at improving the catalytic properties of these enzymes.

The present work uses the cellulase platform and takes advantage of the existing library of highly diversified chimeric cellobiohydrolases II (Cel6A) constructed by Heinzelman et al. [53], to test the theory presented in chapter 2 that proposes that highly

mutated, non-native, chimeras are more evolvable than their native counterparts. The intuition is that non-native enzymes whose mutational neighborhood has not been searched by evolutionary processes are more likely to have access to beneficial mutations. However, whether this holds true for chimeric non-native enzymes whose building blocks are derived from native enzymes is not clear and is the subject of this study. The underlying assumption is that cellulases are not globally optimized.

Random mutagenesis was used to generate mutants from chimeric and parental Cel6As. The number of improved mutants and their increment in specific activity are used as measures of evolvability. The former measure has been previously used by others [62] to quantify evolvability. The chimeric library of Heinzelman et al. [53] consists of three fungal Cel6As from *Humicola insolens (H. insolens)*, *T. reesei*, and *Chaetomium thermophilum (C. thermophilum)* denoted as P1, P2, and P3, respectively, and the possible 6,558 chimeras that can be constructed from the seven crossovers designed by the structure-guided recombination algorithms SCHEMA and RASPP [2,4]. The Cel6As consist of a catalytic domain (CD) and a cellulose binding module (CBM) joined by a flexible linker. The crossovers occur in the CD, while the linker region and the CBM are the same in each member of the library and derived from the Cel6A of *T. reesei*. The CBM, CD and linker region contain a total of about 450 amino acids. Chimeric and native Cel6As were expressed in *Saccharomyces cerevisiae (S. cerevisiae)*.

SCHEMA and RASPP are structured-guided recombination algorithms that select the crossover locations that minimize structural disruption while maintaining a high level of sequence diversity. SCHEMA scores chimeras according to the number of non-native amino acids contacts formed upon recombination. Contacts are defined by a 4.5 Å structural cutoff. RASPP then directs crossovers to locations that minimize the average number of non-native contacts while maintaining a high average mutation rate in the chimeric library (please refer to the last section of the introduction to this thesis for more details on how SCHEMA and RASPP work). Members of the cellulase SCHEMA library have on average 50 mutations from their closest parent and 15 non-native contacts. Several highly mutated members of this library were shown to have the same level of specific activity on Avicel as their parent enzymes making this an excellent system to begin testing the theory of chapter 2 on real proteins.

In order for chimeric Cel6As to have access to more and better beneficial mutations than their wild-type counterparts, there must exist mutations that are beneficial in the context of the former but not the latter. This can occur only if mutations make non-additive contributions to catalytic activity which means that the effect of a mutation depends on the presence or absence of other amino acids in the protein. When mutations make additive contributions to fitness, their effect is the same in any background and chimeras cannot gain access to beneficial mutations that are not beneficial in the context of a parental enzyme (denoted hereafter as new beneficial mutations). In order for chimeras to gain access to new beneficial mutations the effect of mutations must depend on at least two other residues (third order contributions to fitness) and the network of three or more interacting residues must be broken by the crossovers of recombination as illustrated in Figure 2.1.

The existence of high order interactions affecting catalytic activity in real proteins has been verified experimentally [68-70]. Nevertheless, it has been argued that mutations exhibit mostly additive effects in proteins and that high order contributions to fitness are rare [64,71]. These studies are often based on double mutant cycle analyses, in which two mutations are found to be independent of one another (and thus to exhibit additivity) when the contribution to fitness of the double mutant is equal to the sum of the contributions of the single mutants [71]. When this condition is not met the effect of one mutation depends on the presence of the other, and the two residues form a pair-wise interaction. Mutations are likely to exhibit non-additive effects when they are in direct contact in the protein structure [71,99,100]. Since most proteins are large, two randomly chosen amino acids are unlikely to be in contact and thus to exhibit non-additivity [64,101]. Thus if one were to perform double mutant cycle analyses [102] on all possible amino acid pairs in a protein, the outcome would be that most pairs exhibit additivity.

Chimeras, however, already contain many mutations relative to their parents. Any further mutations that are introduced into their structure are likely to occur in the proximity of other mutations and thus to exhibit non-additive effects (i.e., to make different contributions to fitness in chimeric versus native backgrounds). This, however, does not hold true for all heavily mutated chimeras. Consider a chimera with a single crossover that inherits its N-terminal half from one parent and its C-terminal half from

another parent. Such a chimera is heavily mutated (the number of mutations relative to its closest parent is high) but, because it is composed of two large native segments, any mutation will likely occur in a native, non-mutated local environment and exhibit additivity unless it occurs at the interface of the segments. This reasoning reemphasizes the importance of creating non-native interactions to gain access to new beneficial mutations and benefit from non-additivity. Creating too many non-native interactions, however, will compromise the activity of the chimera. Whether chimeras are sufficiently diversified to gain access to new beneficial mutations without significantly disrupting the function of the enzyme is not clear. Finally it is important to keep in mind that SCHEMA [2] is based on the principle of conserving native interactions to preserve structural integrity. This further reduces the probability that the library crossovers create new interactions.

According to the high-resolution crystal structure [103] of the CD of the *H. insolens* parent, the active site is highly conserved among the three parent cellulases. Of the 39 amino acids that are within 4.5 Å of the substrate, only one is not conserved among the three parents. Active site mutations are thus likely to make the same contributions to activity in chimeric and parental backgrounds. Therefore long-range interactions with the active must exist in order for chimeras to gain access to new beneficial mutations with respect to catalytic activity. There are numerous reports in the literature of non active site mutations that can affect catalytic activity via long-range interactions with the active site [72-75], but whether they exist in the Cel6A scaffold is not clear.

To determine whether SCHEMA chimeras can exhibit greater evolvability than their parents despite these limitations, the activities on cellulose of point mutants within the mutational neighborhood of native and chimeric Cel6As were determined and compared. The number of clones exhibiting improved specific activity on Avicel and the magnitude of the activity increment was used as the sole measure of evolvability. Random mutagenesis of two native Cel6As and four chimeric Cel6As was used to explore their mutational neighborhood. The genes of these enzymes were amplified and mutated by error-prone PCR. The plasmids bearing the mutated genes were then transformed into *S. cerevisiae* and their whole cell activities on Avicel were determined

using a high-throughput screen recently developed in our lab. A whole cell activity screen is one in which a fixed amount of the cell culture supernatant is used in the screen and thus does not account for differences in protein expression. For this reason, further characterizations must be performed on the best mutants to verifiy improvements in specific activity (activity per unit mass of enzyme). No clones exhibiting improved whole cell activities were found in the native libraries, whereas several were found in the chimeric libraries. The best clones from one of the chimeric libraries were recombined and the top hits from the recombination library were purified and their specific activities on Avicel measured. Specific activity assays are performed with a fixed amount of protein so that differences in protein expression are accounted for. Unfortunately, the improvements in specific activity were very modest, and most of the increase in whole cell activity could be attributed to an increase in expression level.

3.3 Results

3.3.1 Selection of Chimeras

In order to compare the evolvabilities of chimeric and native Cel6As of equal fitness, it was necessary to first identify a set of chimeras that have thermostabilities and activities comparable to their parent enzymes. It is important to control for thermostability because stable enzymes can tolerate destabilizing mutations that are not accessible to less stable enzymes but may be beneficial with respect to catalytic activity [62]. To this end 38 previously characterized chimeras [53] with stabilities comparable to those of their parents were selected and assayed on Avicel along with P1 and P3 using the whole cell high-throughput Avicel assay described in the *experimental methods (*EM). Avicel is considered by many researchers to be a good substrate for exoglucanase activity because it is highly crystalline and it has a high fraction of free reducing ends relative to the fraction of accessible ß-glucosidic bonds [52,104].

Table 3.S1 in the *supplementary material* (SM) lists the sequences and stabilities of the 38 chimeras and two parents used in this initial characterization. The chimeric sequences are reported in terms of the parent from which each of the eight sequence

fragments is inherited. For example chimera 31111112 inherits its N-and C-terminal fragments from P3 and P2, respectively, and all its internal fragments from P1. Stability is reported in terms of T_{50}, the temperature at which an enzyme loses 50% of its activity after a 10 minute incubation [90]. P2 was omitted from this initial characterization because its extremely low level of expression in *S. cerevisiae* would not allow its activity to be detected by this assay. The cells bearing the P3 gene did not grow for the initial characterization. P3 was thus tested in second assay. The assay was performed at five different temperatures: 45, 50, 55, 60, and 65°C. Each enzyme was assayed twice on two different assay plates.

Six chimeras were found to have whole cell activities comparable to P1 over the entire range of temperatures. These results are shown in Figure 3.1. The selected chimeras, P1, and P3 were then reassayed at 50°C and the results are shown in Figure 3.2. This was done by plating a 10 μl aliquot of the remaining supernatant from the culture onto a fresh SD-URA plate. Colonies were allowed to grow for 72 hours at 30°C and then sixteen individual colonies were used to inoculate sixteen wells of two 96-deep well plates containing 50 μl of SD-URA. Growth, expression, and assaying from this point on are described in the EM. The low temperature of 50°C was chosen as the temperature for all subsequent activity tests to 1) reduce the likelihood of selecting stabilizing mutations over activity enhancing mutations and to 2) reduce the bias in evolvability arising from native and parental enzymes having different thermostabilities. A lower temperature would not have produced enough signal in the Avicel assay.

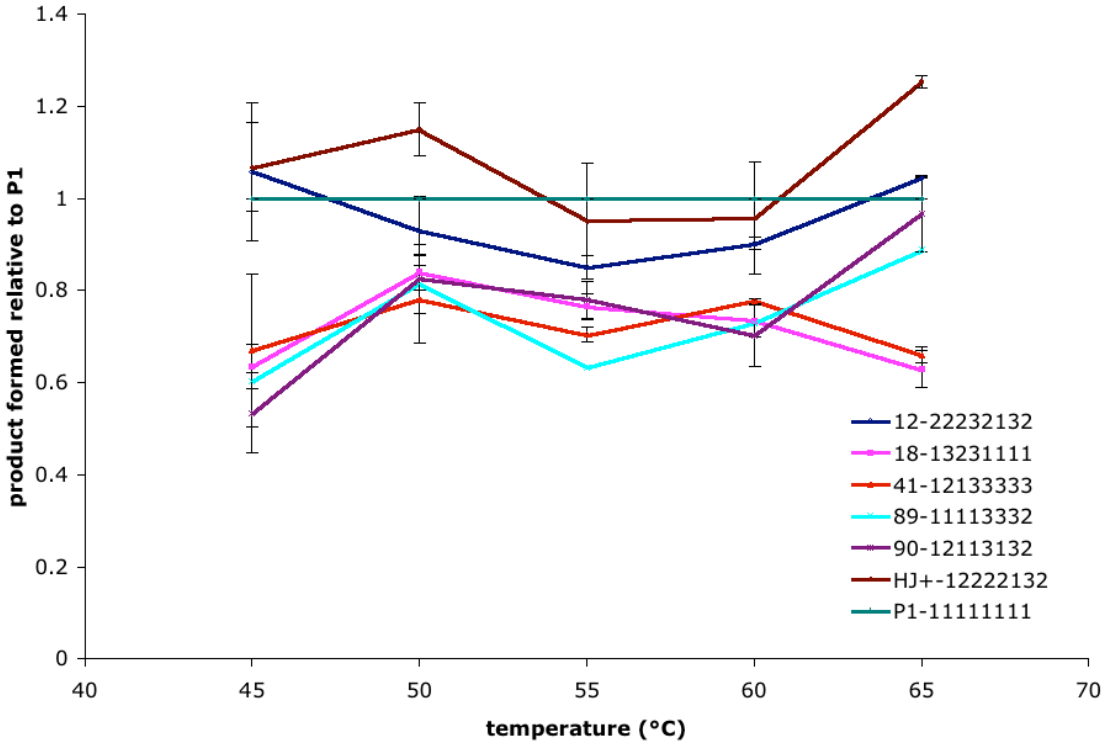


Figure 3.1: Relative amount of cellobiose formed by the six chimeras having the same whole-cell activity as P1. P2 was omitted from this initial characterization because its extremely low level of expression in *S. cerevisiae* would not allow its activity to be detected by this assay. The cells bearing the P3 gene did not grow and this enzyme was tested in a subsequent assay (Figure 3.2). The error bars represent the difference between the two measurements made in different assay plates.

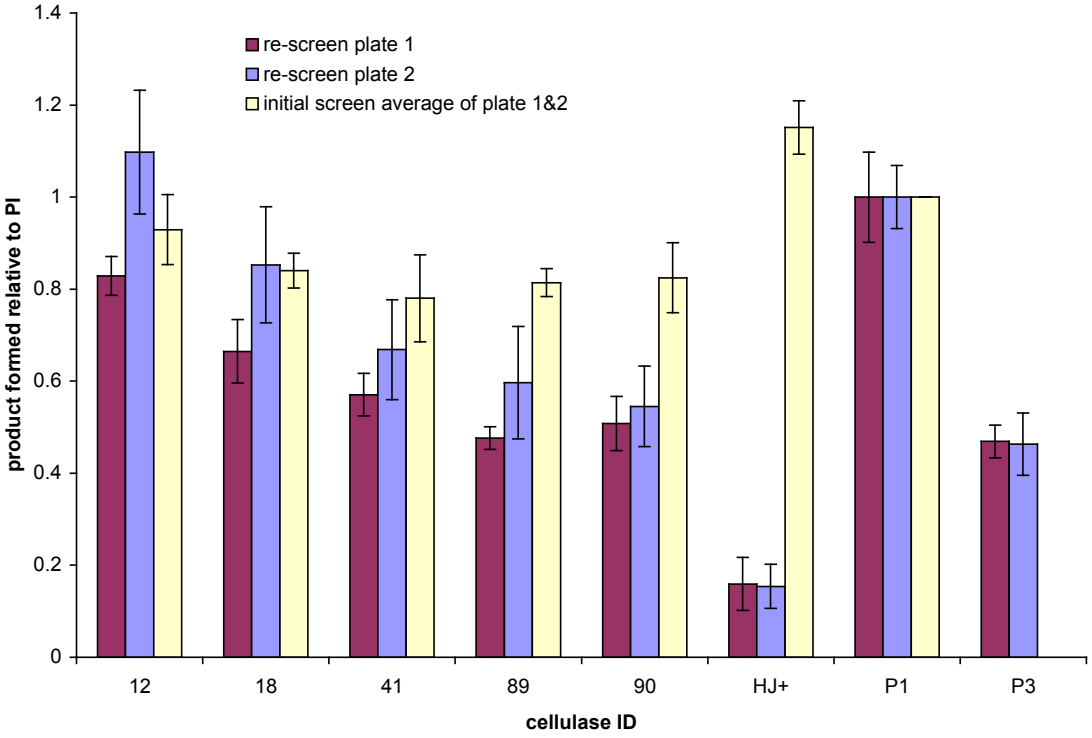


Figure 3.2: Cellobiose formed relative to P1 at 50°C. The yellow bars represent the data from Figure 3.1. The purple and blue bars represent the data from the second assay of plates 1 and 2, respectively. The error bars of the purple and blue bars represent the standard deviations across the 8 measurements of the plate. The discrepancy in HJ^+ activity is related to using the HIS_6-tagged version of this chimera in the second assay.

Another important criterion to consider when selecting the chimeras is the degree to which they differ from their parents. As stated in the introduction and emphasized in chapter 2, chimeras gain access to new beneficial mutations when non-native interactions are formed. For each of the chimeras of Figures 3.1 and 3.2 the number of mutations from their closest parent, m, the number of non-native pair-wise interactions, E, and the number of non-native third order interactions, E^3, were calculated and shown in Table 3.1. Interactions are defined by a 4.5 Å structural cutoff. In a third order interaction each residue is within 4.5 Å of the other two residues. A mutation that occurs in a non-native network of three or more interacting residues can potentially exhibit non-additive effects because it leads to a combination of amino acids that are not accessible to parental

Cel6As (Figure 2.1). The T_{50}s and the relative whole cell activities on Avicel are also shown.

Table 3.1: Summary of the *m, E,* E^3, T_{50} and the whole cell activity on Avicel relative to P1, RA^{P1}, for the chimeras of Figures 3.1 and 3.2. The values of the relative whole cell activities correspond to the average across the first and second assay. The value presented for HJ^+ corresponds to the value obtained from the first assay since the HIS_6-tagged version was accidentally used in the second assay. The values for all the other cellulases are also based on the non-tagged enzymes.

enzyme	*m*	*E*	E^3	RA^{P1}	T_{50}
12-22232132	48	18	16	0.95±0.11	68.0
18-13231111	27	4	2	0.79±0.09	63.3
41-12133333	35	13	12	0.67±0.09	64.0
89-11113332	41	5	1	0.63±0.14	70.0
90-12113132	48	15	10	0.63±0.14	70.5
HJ^+-12222332	47	14	11	1.15±0.06	71.0
P1-11111111	0	0	0	1	64.8
P2-22222222	0	0	0	--	59.0
P3-33333333	0	0	0	0.47±0.00	64.0

Chimeras HJ^+ and 12 were chosen because their whole cell activity levels are comparable to P1 and they have the highest values of *m, E,* and E^3. Chimera 90 was chosen because it has the next highest values of *m, E,* and E^3. Chimera 89 was not intended to be included but a labeling error caused its inclusion.

3.3.2 Characterization of the mutational neighborhood of the selected chimeras using random mutagenesis.

The genes of P1, P3, and chimeras 12, 89, 90, and HJ^+ were amplified using error-prone PCR. The frequency of mutations can be controlled by varying the concentration of $MnCl_2$. For each enzyme, error-prone PCR was performed using concentrations of $MnCl_2$ of 150, 200, 250, and 300 μM as described in the *experimental methods* (EM). The

collection of mutants obtained for each enzyme and concentration of $MnCl_2$ are referred to as a library.

In order to determine the library with an appropriate mutation rate, 88 mutants from each of the libraries containing 150, 200, and 250 μM $MnCl_2$ were screened in 96-well plates using the high-throughput whole cell Avicel screen as described in the EM (a total of 6x3 = 18 plates). The library with [$MnCl_2$] = 300 μM was omitted from this initial characterization because it was assumed that this high concentration was unlikely to yield a sufficient fraction of functional clones. Clones were defined to be functional if their absorbance reading at 520 nm was 0.05 AU above the blank wells. This value corresponds to the minimum increase in absorbance relative to that of the blank required for this difference to be visible to the eye. For comparison this corresponds to about 15% of the whole cell activity of P1. The fraction of functional clones in the libraries with concentrations of $MnCl_2$ of 150, 200, and 250 μM was found to be 36 ± 3%, 29 ± 5%, and 15 ± 3%, respectively. Thus the library with [$MnCl_2$] = 150 μM was chosen for further characterization. A total of 600 clones were screened from each of the six libraries and clones exhibiting an increase of 20% or greater relative to their parents were chosen for a re-screen. A total of 0, 3, 9, 3, 12, and 7 clones were chosen from the P1, P3, 12, 89, 90, and HJ^+ libraries to be re-screened. Clones exhibiting an average of 20% or better improvement in the re-screen are summarized in Table 3.2.

Table 3.2: Summary of the *m, E,* E^3, T_{50} , the whole cell activity on Avicel relative to P1 (RA^{P1}) of the chimeras used to generate the random mutagenesis libraries, the number of improved mutants found in each library, and the improvement of each mutant relative to the enzyme it was derived from.

Enzyme	T_{50}	RA^{P1}	*m/ E/* E^3	*# improved clones*	*% improvement*
12-22232132	68	1.0	48/18/16	3	21%
					20%
					20%
89-11113332	71	0.6	41/5/1	2	28%
					22%
90-12113132	70	0.6	58/15/10	0	--
HJ^+-12222332	71	1.2	47/14/11	1	24%
P1-11111111	65	1	0/0/0	0	--
P3-33333333	64	0.5	0/0/0	0	--

The results in Table 3.2 show that all clones exhibiting improvement on Avicel are derived from chimeras. These improvements, however, can be attributed to either an increase in expression, stability, specific activity, or a combination of these three properties and further characterizations are necessary to verify any improvement in specific activity.

Only 600 clones were screened from each of the above libraries because the transformation efficiencies were poor. To obtain results of greater statistical significance, it was necessary to characterize a larger fraction of the neighborhood of each enzyme. To this end a new set of libraries was generated with concentrations of $MnCl_2$ of 50, 100, 150, 200, 250, 300, and 350 μM. Additionally, a mutated version of P2, P2C311S, was included because it was found that the C311S mutation increased expression by nearly 10 fold (a total of 7x7 = 49 new libraries were created). It was important to include this parent because chimeras 12 and HJ^+ are more closely related to P2 than they are to P1 and P3. A fair comparison of the evolvabilities of these two chimeras relative to that of the parental enzymes would have to involve P2 or its closely related mutant P2C311S.

Instead of screening a large number of clones from all seven libraries, it was decided to screen 3,000 clones from the most promising chimeric library, and then recombine, purify, and characterize the specific activity of the best hits, proceeding to screening other libraries if positive results were found. The library from chimera 12 was chosen because 1) its T_{50} is closer to that of the native enzymes, 2) its specific activity is comparable to that of its closest parent, P2C311S (Figure 3.5), 3) it has the highest values of *m, E,* and E^3, and 4) it gave rise to the greatest number improved clones in the initial characterization of 600 members of its neighborhood. To determine the library with the appropriate amount of mutations, a single 96-well plate from each of the seven libraries derived from chimera 12 containing different concentrations of $MnCl_2$ was assayed using the high-throughput Avicel screen. The fraction of functional clones was 81%, 65%, 64%, 68%, 37%, 36%, and 8% in the libraries containing 50, 100, 150, 200, 250, 300, and 350 μM $MnCl_2$, respectively. Because of the large gap in functional clones between the library containing 200 μM $MnCl_2$ (68% functional) and the library containing 250 μM $MnCl_2$ (37% functional) it was decided to screen 3,000 clones from each. A total of 20 clones exhibiting an improvement of 20% or better were selected for the re-screen. The improvements relative to chimeras 12, P1, P2C311S, and P3 of the top five clones in the rescreen are shown in Figure 3.3. Their non-synonymous mutations and relative improvement are summarized in Table 3.3. The locations of the mutations are shown in Figure 3.4.

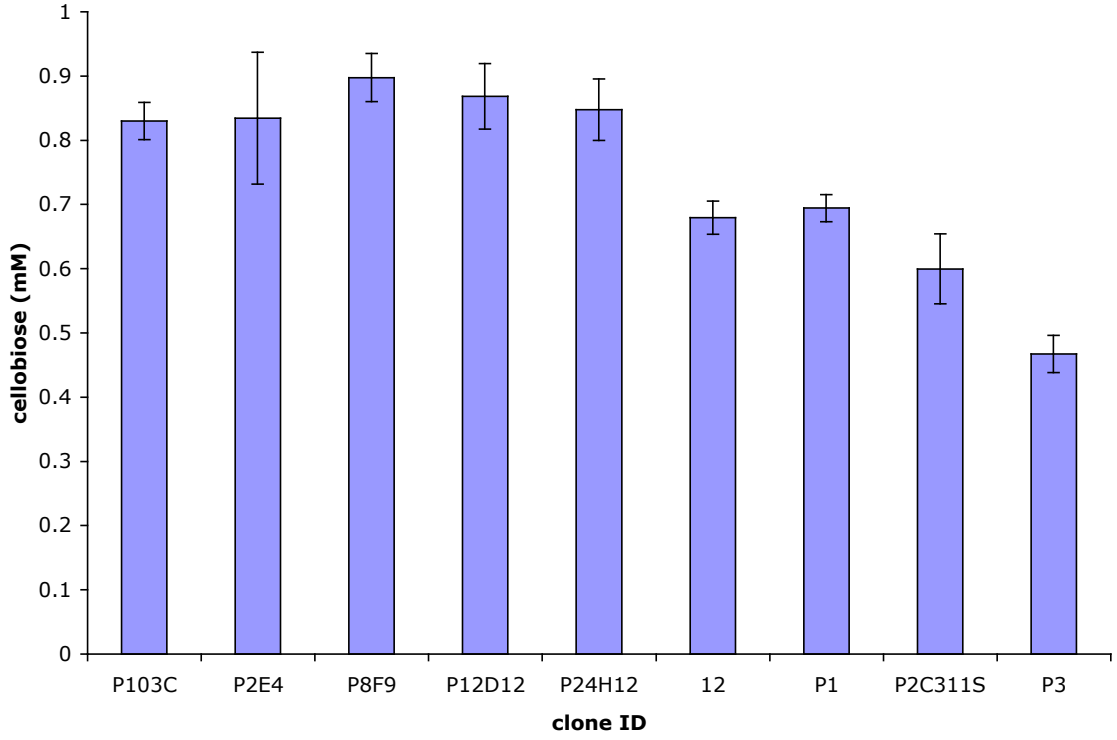


Figure 3.3: Product formed by the best five mutants found in the libraries derived from chimera 12 with [$MnCl_2$] = 200 and 250 μM. The values are based on the re-screen data (whole cell activity). Error bars represent the standard deviation across eight independent measurements.

Table 3.3: Summary of the mutations in the five clones selected from the re-screen and the relative improvement they provide relative to chimera 12, $RA^{C\text{-}12}$, based on the whole cell screen on Avicel. Underlined mutations occur in the linker region and mutations in bold appear in more than one selected clone. Numbering is based on the 1OCN structure [103].

Clone	$RA^{C\text{-}12}$	*Mutations*
P103C	22%	D282E
P2E4	23%	P283L, S410P
P8F9	32%	**N290D**, V398A
P12D12	28%	<u>Y93N</u>, **N290D**, N445Y
P24H12	25%	<u>T73A</u>, V440A

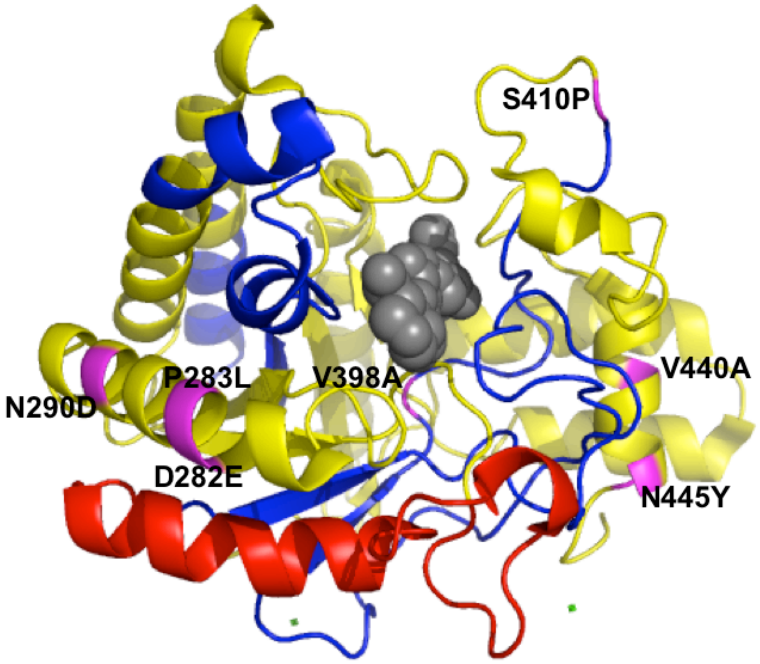


Figure 3.4: Locations of the best five mutations from the re-screen. The structure is based on the 1OCN structure [103]. The substrate, a cellobio-derived isofagomine glycosidase inhibitor, is shown in grey. The portions of the structure shown in red, yellow, and blue represent those derived from P1, P2, and P3 respectively. Mutations are shown in magenta.

Figure 3.4 shows that most mutations occur on the surface of the enzyme, with the exception of mutation V398A that occurs in the active site and is within 5.6 Å of the substrate. This residue is conserved in the three parents. This residue is in contact with seven other residues, but these are perfectly conserved in the three parents. It is also involved in eight third order interactions (three residues that are within 4.5 Å of each other) that are also perfectly conserved in the three parents. Thus, the V398A mutation occurs in a very native environment and can only make non-additive contributions to fitness if it is involved in long-range interactions with mutated residues. The same is true for all other mutated residues with the exception of residue D282. Residue D282 is involved in a third order interaction with residues Q286 and R345 in the background of chimera 12. These residues are QER, DQH, and QEQ in P1, P2, and P3, respectively.

The mutation D282E thus leads to a combination of amino acids, EQR, not accessible to any of the parental cellulases. However, D282E is a surface mutation and it is only via long-range interactions with the active site that this newly formed third order interaction can affect catalytic activity. Because the improvements in whole cell activities are not very high, the identified mutations were recombined with the purpose of selecting, purifying, and determining the specific activities of clones with improved whole cell activities of at least 50%.

3.3.3 Recombination of the mutations in the best five mutants.

The mutations in Table 3.3 were recombined using overlap extension PCR as described in the EM to create all possible $2^9 = 512$ combinations of amino acids. Over 2,000 clones were screened (four-fold over sampling) and the 18 clones exhibiting an improvement of 30% or better were re-screened. The top six clones with relative improvements of 57 (2C), 59 (3D), 57 (6D), 56 (7D), 54 (7G), and 54% (8E) with respect to the whole cell activity of chimera 12 were chosen for further characterization. The mutations found in these clones are summarized in Table 3.4.

Table 3.4: Summary of the mutations identified in the six best clones from the recombination library. A "*" means that the mutation is present in that clone.

	N290D	*D282E*	*S410P*	*T73A*	*V398A*	*V440A*	*P283L*	*Y93N*	*N445Y*
2C	*	*	*	*		*		*	
3D	*	*	*		*		*		
6D	*	*		*					
7D	*	*	*	*	*	*		*	
7G	*	*	*	*	*				*
8E	*	*	*	*	*		*		

Table 3.4 shows that two of the nine mutations, N290D and D282E, appear in all the selected clones. Mutations S410P and T73A occur in five of the six selected clones.

The active site mutation V398A appears in four of the six clones. Mutations P283L, V440A, and Y93N appear in two of the six clones and mutation N445Y appear in one of the six selected clones.

To determine the specific activity of the selected clones on Avicel they were tagged with six histidines (HIS_6) on their C-terminus, purified, and their specific activity on Avicel measured as described in the EM. The accuracy of the HIS_6 constructs was verified by sequencing. The concentrations were determined by measuring the absorbance at 280 nm and using the molar extinction coefficient of 92,425 $M^{-1}cm^{-1}$ [105] as described in the EM. The purification profiles, the SDS PAGE gels of the purified proteins, and the absorbance readings at 280 nm plotted as a function of volume of purified protein are shown in Figures 3.S1-3.S3 in the SM. To determine the errors in the measurement of the specific activity related to the process of purification, buffer exchange, and determination of concentration, the entire procedure was repeated independently twice for clone 6D, chimera 12, and P2C311S. The specific activity was measured using 300 nM protein, 50 mg/ml Avicel, and 80 mM NaCl for 2 hours at 50ºC as described in the EM. The specific activities of the clones are compared to chimera 12 and to parent P2C311S (the parent exhibiting greatest homology to chimera 12). The results are shown in Figure 3.5.

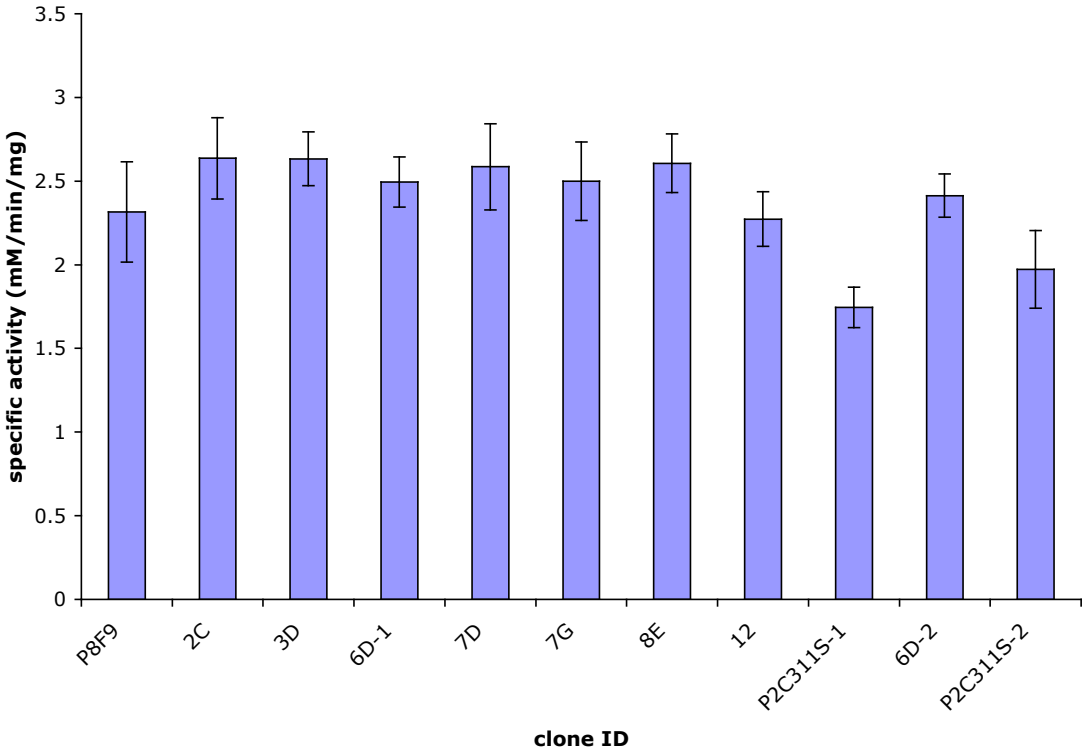


Figure 3.5: Specific activity (mM cellobiose/min/mg protein) of the top five clones from the recombination library, the top clone, P8F9, from the random mutagenesis library, chimera 12, and parent P2C311S. Two bars are shown for mutant 6D and parent P2C311S corresponding to the two independently purified batches. The specific activity was measured using 300 nM protein, 50 mg/ml Avicel, and 80 mM NaCl for 2 hours at 50ºC as described in the EM. The error bars represent the standard deviation across six independent measurements.

Figure 3.5 shows that the improvements in specific activity relative to chimera 12 are very modest and range from 5-10%. The improvements based on the whole cell activity assay ranged from 54-59%, suggesting that the bulk of the improvement can be attributed to an increase in expression. Figure 3.S1 in the SM shows the purification profiles of chimera 12, parent P2C311S, clone 6D, and clone 7D. These proteins were grown and purified in parallel under identical conditions so that the relative size of the peaks in the purification profile are an indication of their expression. Supplemental

Figure 3.S1 suggests that clones 6D and 7D are more highly expressed than their parent, chimera 12.

The improvements in specific activity are rather low but could be reproduced on different days and under different assay conditions (data not shown). The independently purified batches of mutant 6D show great consistency in specific activity: 2.5 ± 0.1 and 2.4 ± 0.1 mM/min/mg. The consistency of the two P2C311S batches is not as good, 1.7 ± 0.1 and 2.0 ± 0.2 mM/min/mg, but still acceptable. To determine whether the increments in specific activity could be attributed to an increase in thermal stability the fraction of enzyme remaining functional after various incubation times at 50°C was determined as described in the EM. Briefly, identical protein samples containing 1.8 μM protein were incubated at 50°C for 0, 3, 6, and 9 hours. The samples were then screened on Avicel at 30°C for 2 hours. The high protein concentration was necessary to obtain a reliable signal at the low assay temperature of 30°C. This experiment was performed on chimera 12, parent P2C311S, and clone 6D. The results are shown in Figure 3.6.

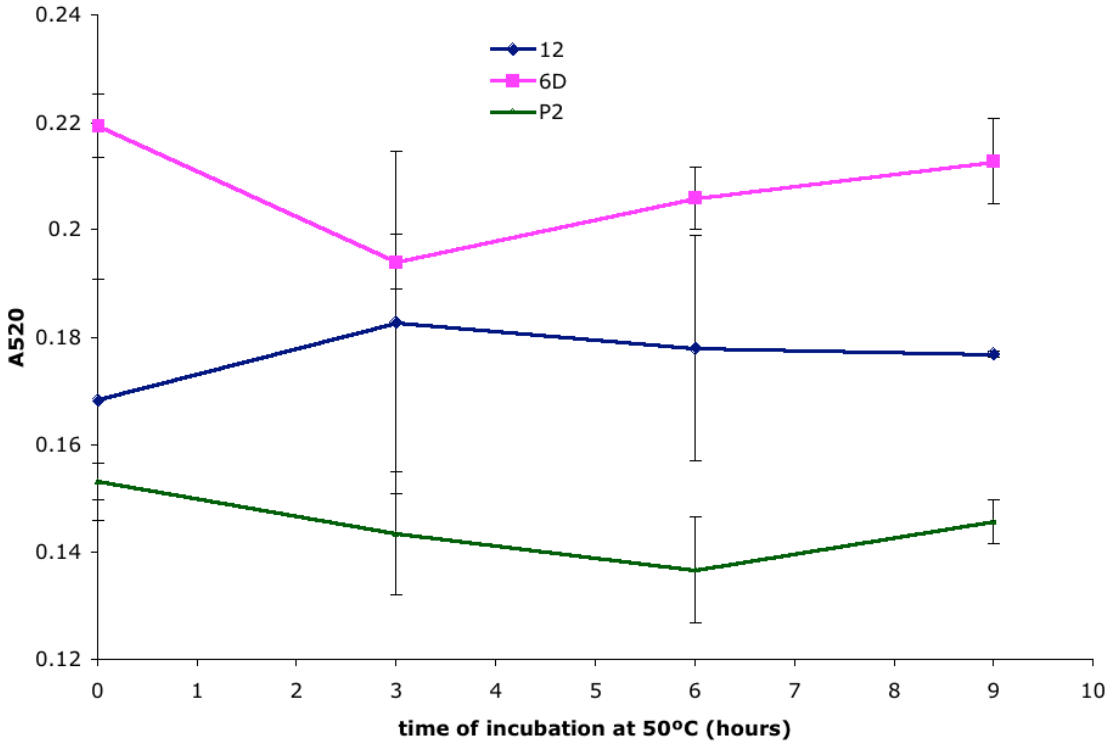


Figure 3.6: Relative Avicel activity remaining at 30°C after different incubations times at 50°C. Identical protein samples containing 1.8 μM proteins were incubated at 50°C for 0, 3, 6, and 9 hours and then screened on Avicel for 2 hours at 30°C.

Figure 3.6 shows that after as long as nine hours of incubation at 50°C the proteins still retain 100% of their activity. This excludes the possibility that the small increments in specific activity are the results of stabilizing mutations. To verify whether the increments in specific activity lead to a detectable increase in initial rate of reaction the initial rates of reaction on Avicel were determined as described in the EM for chimera 12, parent P2C311S, and clone 6D. The Park-Johnson assay [106] described in the EM was used instead of the Nelson-Somogyi [107,108] method because it is more suited to measuring the low concentrations of cellobiose that are reached during the first three minutes of the reaction. Each protein was tested twice, using the two independently purified batches of chimera 12, parent P2C311S, and clone 6D. The results are shown in Figure 3.7. The rate of reaction was measured during the first three minutes after the

addition of the protein to a pre-heated Avicel sample. Measurements were taken every 30 s. Details are described in the EM.

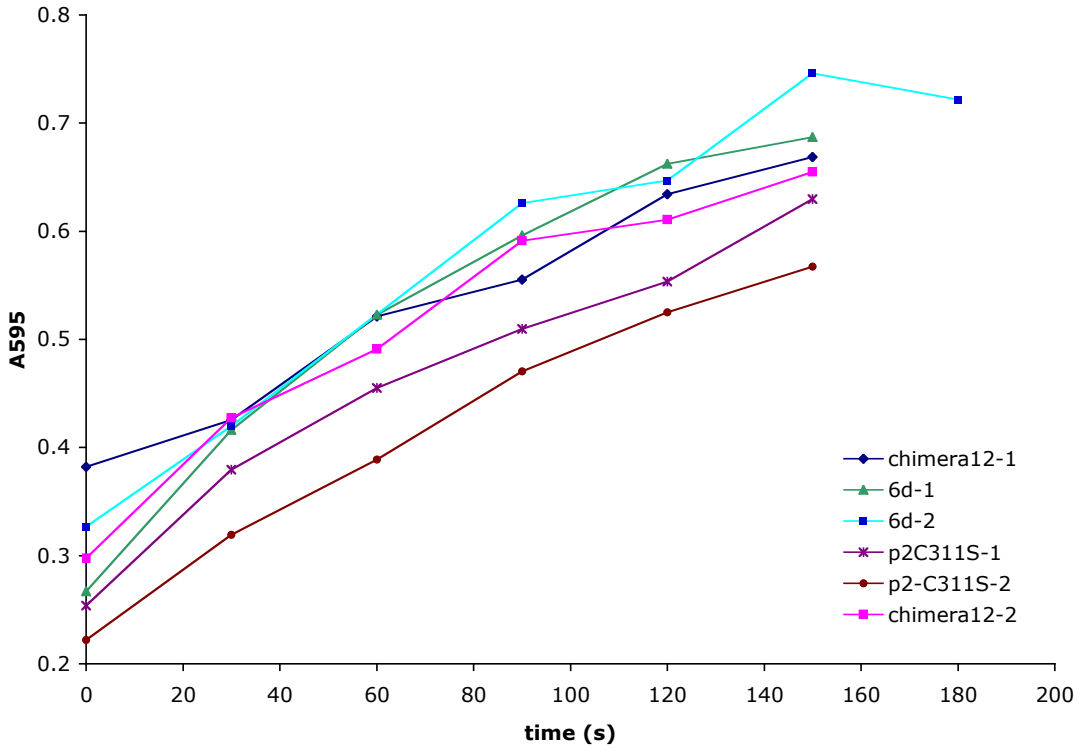


Figure 3.7: Initital rates of reactions for chimera 12, parent P2C311S, and clone 6D. The profile of clone 6D-2 extends to 180 s because the time point at 60 s was accidentally missed. Reactions were carried out using 300 nM protein, 134 Mm NaCl, and 10 mg/ml of Avicel at 50°C in 50 mM sodium acetate buffer (pH = 5.0).

Figure 3.7 shows that the small increment in specific activity observed in clone 6D cannot be detected in the initial rate of reaction. Although this was not repeated for the remaining mutants, their similarity in sequence (Table 3.4) and in specific activity to clone 6D strongly suggests that their initial rate of reaction profile will not exhibit significant differences to those shown in Figure 3.7.

3.4 Discussion

The experiments described in this chapter aimed to generate experimental evidence that chimeric enzymes have access to more and better beneficial mutations than their wild-type counterparts. This hypothesis was tested in the context of a previously characterized library of chimeric Cel6As [53]. The chimeras of this library differ, on average, by 50 amino acids relative to their closest parent. However, whether their mutational neighborhood is effectively unexplored by evolutionary processes is unclear because their building blocks are derived from native enzymes. To reduce bias arising from differences in fitness between the chimeras and their parents, the chimeras used in this study have activities on Avicel and stabilities comparable to those of their parents.

The mutational neighborhood of four chimeric and two native Cel6As was searched for mutants with improved whole cell activities on Avicel. All clones exhibiting improvements were derived from chimeric parents and none could be found in the parental libraries. Following this initial characterization, 6,000 clones in the neighborhood of chimera 12 were screened on Avicel. The best five mutants exhibited improvements in whole cell activity ranging from 22-32%. Their nine mutations were recombined using overlap extension PCR and 2,000 clones from this library (roughly 4-fold over sampling) were screened on Avicel. The best six clones, exhibiting whole cell activity improvements ranging from 54-59%, were purified and their specific activity on Avicel characterized. The improvements in specific activity were very modest, ranging from 5-10%, implying that the bulk of the ~55% whole cell activity improvement was the result of an increase in cellulase expression as also suggested by the purification profiles of Supplemental Figure 3.S1. It was verified that these improvements were not the result of stabilizing mutations. However, it was not possible to detect a faster rate of initial reaction. This was not surprising as increases in specific activity of only 5-10% over the course of a two hour reaction are unlikely to be detectable in the first three minutes of reaction. Furthermore, the improvements in specific activities may be an artifact arising from errors in the measurement of active cellulase concentration. As shown in the SDS PAGE gels of the purified proteins, there is an unidentified band at 30 kDa that contributes to the measured concentration of the purified cellulases. While this band is rather faint, and appears to be relatively uniform across the protein samples, the specific

activity improvement is so low that its mere existence is extremely sensitive to experimental errors. Finally, supposing that the improvements in specific activity are real, they do not represent progress on mutagenesis studies performed on native cellulases because increments in specific activity of 5-10% or more have already been reported.

Provided native enzymes are locally optimized, non-native enzymes will always have, on average, access to a greater number of beneficial mutations than native enzymes. The expected number of beneficial mutations accessible to locally optimized native enzymes is zero. For any non-native enzyme this number is strictly greater than zero, independently of fitness. Ultimately, however, to find a single beneficial mutation in the neighborhood of a chimera, this number must be greater than or equal to one. Therefore, besides an unlucky choice of chimera, the failure to identify sufficiently strong beneficial mutations (i.e., > 50% improvement in specific activity) in the neighborhood of chimera 12 can be attributed to 1) a low frequency of beneficial mutations within the reach of chimeras from this library or 2) a low frequency of beneficial mutations in the entirety of sequence space. Additionally, the high-throughput Avicel screen may be inadequate for selecting mutants with improved specific activity. The first two scenarios and the limitations of the Avicel screen are discussed in more detail below.

3.4.1 Beneficial Mutations within the Reach of the Chimeras from the SCHEMA Cel6A Library are too Rare to be Found

In order for chimeric Cel6As to have access to more and better beneficial mutations than their wild-type counterparts, there must exist mutations that are beneficial in the context of the former but not the latter. The last two chapters have emphasized the necessity to create new, non-native functionally important third or higher order interactions in order for this to occur. Thus, supposing that sequences encoding Cel6As that have higher activity than wild-type Cel6As exist in sequence space and are frequent enough to be found, the probability that they can be found using recombination depends on 1) the significance and frequency of functionally important high order interactions in the Cel6A scaffold, and 2) the likelihood that such interactions are broken by the crossovers of recombination to form, new, non-native interactions that make, on average, a neutral

contribution to fitness. Neutrality is important because, in general, forming new interactions will have a deleterious effect on protein function [77] and, in addition to trivializing the hypothesis, it is unlikely that any benefit can be derived from a chimera that is more evolvable but significantly less fit than a native enzyme.

Protein landscapes have often been described as "smooth" or "roughly additive." These conclusions are mostly derived from experimental observations that the contribution to fitness of a double mutant is often equal to the sum of the contributions of the individual mutations [71]. These results, however, usually break down when mutations are adjacent to one another in three dimensional space [71,99]. This reflects the fact that, unless long-range interactions are frequent and significant, high order contributions to fitness can only be detected when mutations are recruited into a non-native environment. Therefore, high order interactions could be extremely common and yet have gone unnoticed by double mutant cycles analyses in which the effects of just a few mutations at a time are evaluated.

Additivity has also been observed in many directed evolution studies in which beneficial mutations were accumulated one at a time and found to be independent of one another. However, Weinreich et al. [69], showed that only 18 of the possible 120 paths linking two ß-lactamases differing by five mutations and five orders of magnitude in antibiotic resistance conferred to bacteria were possible by adaptive evolution. Thus 102 out of the 120 paths linking the two enzymes and exhibiting non-additivity would never be found by a standard directed evolution study. Similar results were observed in [70]. Furthermore, additivity has been primarily verified with respect to thermostability [101,109]. When contributions to catalytic activity are being considered rather than contributions to stability (as is the case in the present study) non-additive effects become much more frequent [68-70,99]. Da Silva and co-workers [68], for example, showed that non-additivity was common, and often involved third or higher order interactions in a functionally important region of an HIV glycoprotein. Therefore, high order interactions do exist in sequence space and are possibly more frequent than expected.

In order for chimeras to have access to new beneficial mutations, the crossovers of recombination must break functionally important interactions and replace them with new ones that make neutral contributions to fitness. In the chimeras of the SCHEMA

Cel6A library, however, the active site is largely conserved making it impossible to break functionally important interactions within the active site. Numerous studies have reported catalytically important mutations that occur remotely from the active site [72-75]. In one study directed evolution was used to change the substrate specificity of aspartate aminotransferase. The mutant enzyme had a 10^6 fold increase in the k_{cat}/K_M on the non-native substrate valine. Only one of the 17 acquired mutations was in contact with the substrate. The three dimensional structure of the mutant enzyme bound to a valine analog showed that the remote mutations caused structural alterations in the active site and surroundings [72]. Thus, despite the conserved active site in the Cel6A library, mutations acquired on the surface of the chimeras may cause structural rearrangements that give them access to catalytically beneficial mutations that are not accessible to their parents. However, since the mutations acquired on the surface of the chimeras are derived from homologous enzymes, they may be unlikely to significantly alter their structural features. Additionally it is important to keep in mind that while the chimeras in the library are heavily mutated, they are only mildly disrupted by design. The SCHEMA algorithm minimizes local disruption by selecting crossovers that maximize the conservation of amino acids between the interfaces of the recombination fragments to minimize the number of non-native contacts formed upon recombination. One consequence of this is that the fragments make additive contributions to thermostability [17]. Another consequence may be that structural rearrangements that can alter the catalytic effects of mutations are highly unlikely.

Furthermore, chimeras that have roughly the same activity as their parents, such as those in the present study, are likely to contain less disruption than the average chimera in the library because, in general, non-native contacts are deleterious. Non-native contacts that make neutral contributions to fitness are less common. In fact, the six chimeras that were chosen because of their similarity in fitness to the native Cel6As have, on average, lower *m* and *E*, than the average *m* and *E* of the 38 chimeras they were selected from as shown in Figure 3.S4 in the SM. Thus, while it is desirable to break interactions to gain access to novel combinations of amino acids, too much disruption leads to poorly active chimeras, limiting the number of possible broken interactions. All of these limitations, combined with the reality that beneficial mutations are rare, may

make it extremely unlikely for SCHEMA to gain access to new beneficial mutaitons. The fact that no significantly beneficial mutations were found in the neighborhood of chimera 12 supports this conclusion.

Alternatively to homologous recombination, novel enzymes with unexplored mutational neighborhood may be generated using *de novo* design. In 1997 Mayo and co-workers pioneered the first fully automated design and experimental validation of a novel sequence for an entire protein. Using computational methods, they completely redesigned the sequence of a 28-residue zinc finger using an algorithm that used as input only the backbone fold and had no knowledge of the naturally occurring sequence. Their final designed sequence was shown to properly fold to the target structure despite very low sequence identity (21%) with the naturally occurring sequence [110]. Baker and co-workers extended this work to the complete redesign of nine globular proteins [111]. Again, their modeling algorithm had no knowledge of the natural sequences and used only the backbone structure to design the novel sequences. On average 65% of the residues in the designed sequences differed from wild-type over all protein residues and 50% differed from wild-type in the core. Yet eight out of the nine designs encoded properly folded proteins. In one case, the designed sequence encoded a protein that was a striking ~7 kcal/mol more stable than the corresponding wild-type. This technique could be used to design highly mutated cellulases based on the Cel6A backbone fold. Enzyme size would be the major limitation. The CD of cellulases is approximately 360 residues long while the designs by Baker and co-workers spanned at most 100 residues.

In 2008, Baker and co-workers [112] successfully designed eight enzymes with two different catalytic motifs that could catalyze the Kemp elimination reaction. If it is possible to design a novel enzymatic function, it may be possible to preserve an existing native one while simultaneously diversifying the active site. In the case of the Cel6As, this could be done by preserving the key catalytic residues, while mutating the active site in such a way that does not create steric hindrance with the substrate. If this is done with sufficient care, it may be possible to preserve most of the native activity while strongly diversifying the sequence of the designed enzyme- both on the surface and in the core. This approach may bypass the limitations of chimeragenesis because the designed enzymes would not have large portions of their structure identical to those of native

enzymes. If this were to succeed for several independent designs, while it would not guarantee that the activity of the non-native enzymes could be improved beyond that of their native counterparts, it may shed light on the distribution of activities in the cellulase landscape.

3.4.2 Fitness Peaks Taller than the Native Peak are Rare in Sequence Space

The hydrolysis and utilization of cellulose is widely distributed among many genera in the domain *Bacteria* and in the fungal groups within the domain *Eucarya*. Recently, it has been discovered that certain animal species, including crayfish and termites, produce their own cellulases [98]. Currently, there are hundreds of known genes encoding enzymes with cellulolytic activity. Often they exhibit fairly different folds. The abundance of cellulose, the strong selective pressure conferred to any organism capable of utilizing it efficiently, and the rich reportoire of identified cellulases could suggest that these enzymes have been highly optimized by natural evolution to degrade cellulose. However, cellulases appear to be much less active on their native substrate than their related glycoside hydrolases [98,113] are on their native substrates. Klyosov [113] calculates k_{cat} values of 0.5 to 0.6 s^{-1} for *T. reesei* cellulases, 58 s^{-1} for amylase, and up to 100 to 1,000 s^{-1} for other hydrolases. Another example is the specific activity of a *T. reesei* cellulase on crystalline cellulose (filter paper) which has been found to be 100 fold lower than that of amylase on starch [114].

Cellulase activity on soluble or pre-treated cellulose substrates is generally higher. For example Bernardez et al. [115] compared initial hydrolysis rates of a *Clostridium thermocellum* cellulase system on Avicel and dilute-acid-pretreated mixed hardwood. Pretreated wood was hydrolyzed up to 10-fold faster than Avicel. Similarily, Zhang and coworkers [116] found that the initial hydrolysis of PASC (phosphoric acid swollen cellulose) is more than 100 fold higher than that of Avicel. In fact, available data suggests that the specific activity of exo-acting saccharolytic enzymes on comparable substrates is similar. For example the specific activity of the *T. reesei* Cel6A on cellohexaose is highly comparable to that of a *Aspergillus awamori* glucoamylase on maltohexaose [116]. These arguments may imply that the recalcitrant nature of microcrystalline cellulose poses a

physical limitation to further improving the specific activity of cellulases (i.e., native cellulases are globally optimized) or that, alternatively, native cellulases are not globally optimized but better cellulases are so rare and distant in sequence (and perhaps structure) from native cellulases that evolution just has not found them yet and perhaps will never find them.

In both of these scenarios, the first being that native cellulases are globally optimized and the second being that more efficient cellulases are extremely rare, any engineering approach will fail at discovering cellulases with higher catalytic activities.

3.4.3 Limitations of the High-Throughput Avicel Screen

Besides the obvious limitation that a whole cell screen may miss mutants that have increased specific activity but not increased whole cell activity, the high-throughput Avicel screen may suffer from other limitations. In particular it is well known that the rate of cellulose hydrolysis declines sharply as the reactions proceeds. Several explanations to this have been proposed including enzyme deactivation, product inhibition, decreases in substrate reactivity (presumably because the more accessible glycosidic bonds are hydrolyzed first), and the formation of unproductive and irreversible substrate-enzyme complexes [98,114]. It is possible that the differences reflecting changes in specific activity are more visible early on in the reaction as opposed to two hours into it when these inhibitory effects may have become significant. In other words, if there is fixed amount of substrate that can be hydrolyzed before the reaction rate slows down dramatically (as would be the case in most of the scenarios proposed to lead to a decrease in reaction rate), and the native enzymes take two hours to hydrolyze that fixed amount of substrate, even a significantly more efficient mutant, that could do it, in say, only five minutes, would not be selected by this screen. Unfortunately the most obvious solutions to this problem are not very practical (screening for shorter amounts of time or using significantly more substrate).

However, the fact that all the parents, chimeras, and mutants that have been tested so far have almost identical specific activities on Avicel but not on PASC [53] suggests

that the current protocol of the Avicel screen may be inadequate for detecting differences in activity and that it may be worthwhile to investigate this further.

3.5 Experimental Procedures

3.5.1 Chimeras Construction and Generation of Random Mutagenesis Libraries Using Error-Prone PCR

Details of chimera construction have been reported previously [53]. *S. cerevisiae* cells bearing the Cel6A plasmids were obtained from Dr. Heinzelman and Indira Wu from the California Institute of Technology. All plasmid DNA used for cloning purposes was extracted from *S. cerevisiae* cells using the ZYMOPREP yeast miniprep kit and transformed into *E. coli*. Transformed cells were allowed to grow overnight and then the plasmid DNA was extracted using the Qiagen Miniprep Kit. The Cel6A genes were sequenced using primers cellSeqFor (gtcgggtccgacttgctgtgcttccgg) located in the linker region 224 base pairs upstream of the CD and cellSeqRev (gcaacacctggcaattccttacc) located 108 base pairs downstream of the gene to ensure the presence of the correct gene. To generate random mutagenesis libraries using error-prone PCR, the gene segments coding for the CD were amplified using the forward primer cellCloneFor (ccaacgactattactcccagtgtcttc) located in the linker region 180 base pairs upstream of the CD and the reverse primer cellCloneRev (gacatgggagatcgaattcaactcc) located 47 base pairs downstream of the gene. Error-prone PCR was carried out in 100 μl total volume containing 100 ng DNA, 0.2 μM of each primer, 10 μl of 10X Roche PCR buffer, 10 μl of 55 mM $MgCl_2$, 200 μM of dATP, 200 μM of dGTP, 500 μM of dTTP, 500 μM of dCTP, $MnCl_2$ (50-400 μM), 5 units of Roche *Taq* DNA polymerase. Libraries with $MnCl_2$ concentrations ranging from 50 to 400 μM were prepared for four chimeric and up to three native Cel6As. To reduce mutational bias across libraries with an equal concentration of $MnCl_2$ but different templates, PCR master mixes were prepared containing the primers, the PCR buffer, the nucleotides and the $MgCl_2$. This master mix was then divided into *x* tubes corresponding to the *x* concentrations of $MnCl_2$ and to each tube the appropriate amounts of $MnCl_2$ and PCR water were added. Each of these tubes was then split into *n* tubes corresponding to the *p* parental and the *n-p* chimeric Cel6As to

which template DNA and *Taq* DNA polymerase were added. The PCR program was 95°C for 30 s, and then 20 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 2 min, and then 72°C for 5 min. The mutated amplified DNA from each library was then gel purified and co-transformed with vector digested with the New England Biolabs (NEB) restriction enzymes Xho1 and Acc651 into *S. cerevisiae* using the homologous recombination protocol described in [117]. Transformed cells were then plated on SD-URA agar plates. The SD-URA mix was purchased from MP-biomedicals. Selected clones from this library were sequenced with forward and reverse primers, cellSeqFor and cellSeqRev, respectively.

3.5.2 Recombination of Best Mutants

The best nine mutations selected from the library of chimera 12 were recombined by amplifying six fragments spanning the nine mutations and then assembling them using overlap extension PCR. The forward and reverse primers for the six fragments were 1) cellCloneFor and 12REV_T73A (gtagtagtagaaccaggtggaggcgycgcggagctcga), 2) 12FOR_T73A (tcgagctccgcgrcgcctccacctggttctactactac) and 12REV_Y93N (tggattacctgaatwcgtagcggttcccg), 3) 12FOR_Y93N (cgggaaccgctacgwattcaggtaatcca) and 12REV_D282E_P283L_N290D (ttgtaaacatytgcaaacagctgcgcagcgrgwtcctgatttg), 4) 12FOR_D282E_P283L_N290D (caaatcaggawcycgctgcgcagctgtttgcaratgtttacaa) and 12REV_V398A_S410P (ggagcagaagratcactcgttccatctgactctcctccgggttttrcccaaacgaa), 5) 12FOR_V398A_S410P (ttcgtttgggyaaaacccggaggagagtcagatggaacgagtgatycttctgctcc) and 12REV_N445Y_V440A (ggattggcgtwagtcagtaattgtrcaaaataagcttggaacc), and 6) 12FOR_N445Y_V440A (ggttccaagcttattttgyacaattactgactwacgccaatcc) and CellCloneRev, respectively. The mutations carried by the primers are indicated in their names. The primers are degenerative and allow the inclusion of either the native or the mutated amino acid. The lengths of the six fragments are 139, 86, 615, 402, 160, and 126 base pairs. Six PCR reactions were carried out in parallel in 100 μl total volume containing 50 ng of chimera 12 DNA, 200 μM of each dNTP, 0.2 μM each of the forward and reverse primers listed above, 2 units of NEB Phusion DNA polymerase, and 20 μl of 5X Phusion HF buffer to amplify each fragment. The PCR program was 98°C for 30 s,

and then 30 cycles of 98°C for 10 s, 55°C for 20 s, and 72°C for 10 s, and then 72°C for 10 min for fragments 1, 2, 5, and 6. The PCR program for fragments 3 and 4 were identical but had an elongation time of 10 s. Amplified fragments were gel purified and then reassembled in three steps (five PCR reactions). In the first step fragments 1 and 2 and fragments 5 and 6 were assembled using primers cellCloneFor and 12REV_Y93N and 12FOR_V398A_S410P and cellCloneRev, respectively. The PCR reactions were carried out in 50 μl total volume containing 50 ng of each fragment, 200 μM of each dNTP, 0.2 μM each of the forward and reverse primers, 2 units of NEB Phusion DNA polymerase, and 10 μl of 5X Phusion HF buffer. The PCR program was the same as that used to generate the individual fragments with an elongation time of 10 s. The assembled fragments were gel purified and used for the next step of the assembly. In the second step the assembled fragments 1 and 2 were assembled with fragment 3 and the assembled fragments 5 and 6 were assembled with fragments 4 using the forward and reverse primers cellCloneFor and 12REV_D282E_P283L_N290D and 12FOR_D282E_P283L_N290D and cellCloneRev, respectively. The PCR conditions and program were the same as in the previous assembly step but the elongation times were 30 s. The assembled fragments were gel purified and used for the final assembly step in which assembled fragments 1, 2, and 3 were assembled with assembled fragments 4, 5, and 6 using forward and reverse primers cellCloneFor and cellCloneRev, respectively. The PCR conditions and program were identical to the previous assembly steps with an elongation time of 45 s. The final construct was gel purified and co-transformed with vector digested with Xho1 and Acc651 (NEB) into *S. cerevisiae* using the homologous recombination protocol described in [117]. The transformed cells were then plated on SD-URA agar plates. Selected clones from this library were sequenced with forward and reverse primers, cellSeqFor and cellSeqRev, respectively.

3.5.3 Addition of HIS_6 Tags to Best Mutants from the Recombination Library

PCR reactions were carried out to append a HIS_6 tag to the C-terminus of the best mutants from the recombination library. The forward primer is located 421 base pairs upstream of

the CD domain and includes the Nhe1 restriction site (gctgaagctgtcatcggttacttag) and the reverse primer has a HIS_6 over hang, the Acc651 restriction site, and a stop codon (ctgcaggtaccctaatggtggtgatggtgatgtagaaaactaggattggcgttagtcag). The PCR reactions were carried out in 50 μl total volume containing 50 ng of DNA, 0.2 μl of each the forward and reverse primer, 200 μM of each dNTP, 2 units of Phusion DNA polymerase and 10 μl of 20X HF Phusion buffer. The PCR program was 98°C for 30 s, and then 30 cycles of 98°C for 10 s, 55°C for 20 s, and 72°C for 60 s, and then 72°C for 10. Amplified DNA was digested using NEB dpn1, gel purified, and then digested using NEB restrictions enzymes Nhe1 and Acc651. The vector was prepared by digesting the plasmid bearing the gene of chimera 89 with the same restriction enzymes (Nhe1 and Acc651) and then by gel-purifying it. Vector and PCR inserts were ligated using NEB T4 ligase at 16°C for 16 hours. The ligation mixture was purified using the QIAGEN DNA purification kit and then transformed into *E. coli* cells and plated on LB plates and allowed to grow overnight. Individual colonies were then picked and grown overnight in 5 ml of LB media supplemented with ampicillin, and then the plasmid DNA was extracted using the QIAGEN miniprep kit and sequenced using primers cellSeqFor and cellSeqRev to ensure correct incorporation of the HIS_6 tag. Plasmids were then transformed into *S. cerevisiae* cells using the Zymo EZ frozen yeast transform kit and plated on SD-URA plates.

3.5.4 Protein Expression

Cells were grown for 72 hours at 30°C on SD-URA agar plates and then individual colonies were used to inoculate 96-well plates containing 50 μl of SD-URA media. Cells were allowed to grow overnight at 30°C and 250 RPM in a Kuhner shaker and were then expanded by adding 350 μl of YPD media (10 g yeast extract, 20 g peptone, and 20 g dextrose in 1 L of water). Cells were allowed to grow an additional 48 hours at 30°C and 250 RPM and were then centrifuged and the supernatant used for the high-throughput Avicel screen.

3.5.5 Protein Purification

Protein purification was achieved by growing cells for 72 hours at 30°C on SD-URA agar plates and then inoculating 5 ml of SD-URA media with a single colony. The cell culture was allowed to reach saturation overnight and then expanded by adding it to 50 ml of YPD media in Tunair flasks purchased from IBI Scientific. The cultures were then centrifuged, and the supernatant filtered using VWR 0.2 μm Nalgene filters. 500 μl of 100 mM phenylmethanesulfonylfluoride (PMSF), 500 μl of 2% NaN_3, and 50 μl of 10 M NaOH were then added to the sample to inhibit protease activity, preserve the sample, and improve binding to the column, respectively. The sample was then purified using a 1 ml HisTrap HP column precharged with nickel (GE Healthcare) and an AKTA purifier FPLC system (GE Healthcare). The binding buffer was composed of 20 mM Tris, 100 mM NaCl, and 10 mM imidazole at pH 8.0. The elution buffer was composed of 20 mM Tris, 100 mM NaCl, and 300 mM imidazole at pH 8.0. First, the column was equilibrated with five column volumes (cv) of binding buffer. Then, the sample was injected and washed with another seven cvs of binding buffer. Sample elution was achieved with a linear gradient. The proteins eluted at a concentration of about 100 mM imidazole. Buffer exchange was performed using SARTORIUS STEDIM 10,000 mwco VIVASPIN columns. The purified proteins were concentrated to about 500 μl in 50 mM sterile sodium acetate buffer at pH 5.0. An additional 5 μl of 100 mM PMSF and 5 μl of 2% NaN_3 were then added, and the purified protein was stored at 4°C. The presence of a band at about 55 kDa on an SDS PAGE gel verified the presence of the correct protein. The concentration of the purified protein was determined by adding different amounts of the purified protein to 1 ml of a solution composed of 6 M guanidine hydrochloride and 25 mM Na_2HPO_4 at pH 6.5 and measuring the absorption at 280 nm (A_{280}). Once a linear relationship was observed between the amount of added sample and the A_{280}, the concentration of the proteins was calculated using an extinction coefficient of 92,425 $M^{-1}cm^{-1}$ [105].

3.5.6 High-Throughput Avicel Activity Assays

An Avicel slurry containing 50 mg/ml of Avicel and 134 mM NaCl was stirred until it was visually homogenous. While still stirring, 60 μl of this slurry were added to PCR plates (TemPlate III 96-well, half skirted, 0.2 ml, thin wall, standard depth, and rimmed well, USA Scientific) with the aid of an 8-channel RAININ multichannel pipettor. 100 μl of protein supernatant were transferred from the 96-well culture plate to the PCR plate containing Avicel with the aid of a pipetting robot (Multimek 96 automated pipettor). The plates were then covered and placed at 4°C for 90 min to allow the enzymes to bind to the Avicel. The plates were then centrifuged to allow the Avicel and bound enzyme to settle at the bottom of the PCR plate wells. With the aid of the robot the supernatant was removed taking care not to disrupt the Avicel-enzyme pellet. With the aid of the pipetting robot, 180 μl of 50 mM sterile sodium acetate buffer pH 5.0 was then added to the pellet. The purpose of this step is to remove the sugars present from the growth media that would otherwise interfere with the reducing sugar assay. The centrifugation and wash steps were repeated four times. On the fourth time, the bound enzyme was re-suspended in 75 μl of 50 mM sterile sodium acetate buffer pH 5.0, the plates were sealed with Biorad microseal B film PCR sealers, and the reaction was initiated by placing the PCR plates in a Fisher Scientific Isotemp 220 water bath at 50°C. After two hours the plates were placed for ten min on ice to quench the reaction. The plates were then centrifuged and 50 μl of the supernatant were transferred to a new PCR plate to perform the Nelson-Somogy reducing sugar assay [107,108]. Somogyi reagent was prepared by mixing 4.8 ml of Somogyi reagent 1 with 1.2 ml of Somogyi reagent 2 per plate and then adding 50 μl of this solution to each well of the PCR plates with the aid of a 12-channel RAININ multichannel pipettor. Somogyi reagent 1 was prepared by dissolving 72 g of Na_2SO_4, 6 g of potassium sodium tartrate tetrahydrate (Rochelle salt), 12 g of Na_2CO_3, and 8 g of $NaHCO_3$ in 400 ml of ddH_20. The solution was then filtrated using a VWR 0.2 μm Nalgene filter for sterilization purposes. Somogyi reagent 2 was prepared by dissolving 18 g of Na_2SO_4, 2 g of $CuSO_4 \bullet 5H_20$ in 100 ml of ddH_20. The solution was then filtrated using a VWR 0.2 μm Nalgene filter for sterilization purposes. Somogyi reagents 1 and 2 were prepared in advance and stored at room temperature. The plates were then sealed with Biorad microseal B film PCR sealers and placed at 98°C in a Fisher Scientific

Isotemp 220 water bath for 15 min. The plates were then allowed to cool on ice for 15 min, and then 50 μl of Nelson reagent was added to each well with the aid of a RAININ multichannel pipettor. The Nelson reagent was prepared by dissolving 25 g of ammonium molybdate in 450 ml of ddH_20 in a glass bottle wrapped in aluminum foil to protect the solution from light. 21 ml of concentrated H_2SO_4 were then added to the solution. 3 g of Na_2H arsenate were dissolved in 25 ml of ddH_20 and then added to the ammonium molybdate-H_2SO_4 solution. The solution was incubated at 37°C for 24 hours and then filtrated for sterilization purposes. The Nelson reagent was stored in the dark at room temperature. The plates were then centrifuged to remove the air bubbles. The pipetting robot was then used to thoroughly mix the PCR wells by pipetting and dispensing several times. 100 μl of the reaction mixture were then transferred to 96-well assay plates and the absorbance at 520 nm was measured. The amount of released cellobiose was determined using a calibration curve constructed with a cellobiose standard.

3.5.7 Measurement of Avicel Specific Activity

Reactions were carried out in 100 μl total volume, 300 nM purified protein, 50 mg/ml Avicel, and 80 mM NaCl. Each protein was assayed six times in six wells of a PCR plate. The PCR plate was kept on ice as the components of the reaction were being added and was then sealed with Biorad microseal B film PCR sealers. The reaction was initiated by placing the PCR plate in a Fisher Scientific Isotemp 220 water bath at 50°C for two hours. After two hours the PCR plate was placed on ice for 10 min to quench the reaction. The plates were then centrifuged and 50 μl of the supernatant were then transferred to a new PCR plate and Avicel activity was measured using the Nelson-Somogyi method as described in section 3.5.6.

3.5.8 Determination of Initial Rate of Reaction on Avicel

The initial rate of reaction of purified proteins was determined using the Park-Johnson (PJ) assay [106]. There are three PJ reagents. PJ reagent A is made by dissolving 0.5 g of $K_3Fe(CN)_6$ and 34.84 g of K_2HPO_4 into 1 L of ddH_20, then adjusting the pH to 10.6 and

then filtrating the solution using a VWR 0.2 μm Nalgene filter for sterilization purposes. PJ reagent B is made by dissolving 2.65 g of Na_2CO_3 and 0.325 g of KCN into 500 ml of ddH_2O and then filtrating the solution for sterilization purposes. PJ reagent C is make by adding 2.5 g of $Fe(III)Cl_3$, 10 g of polyvinylpyrrolidone, and 56.4 ml of H_2SO_4 into 1 L of ddH_20 and then filtrating the solution for sterilization purposes. A slurry containing 10 mg/ml of Avicel and 134 mM NaCl was prepared and heated in an Eppendorf thermomixer set to 50°C for 1 hour. Enzyme was then added to the tube to reach a final concentration of 300 nM in a volume of 600 μl. The sample was immediately vortexed and 100 μl were removed and placed into another tube kept on ice. This corresponds to the first time point. This procedure was repeated 5 more times in intervals of 30 seconds to finish the 600 μl of solution. The tubes containing the 100 μl aliquots were then centrifuged and 50 μl of the supernatant was transferred to a PCR plate. Then 100 μl of PJ reagent A and 50 μl of PJ reagent B were added to the sample and the PCR plate was sealed with Biorad microseal B film PCR sealers. It was then incubated at 95°C in a PCR thermocycler for 15 minutes, and then 180 μl of this were transferred to an assay plate containing 90 μl of PJ reagent C. The absorbance at 595 nm was read immediately and the amount of sugar formed determined using a calibration curve made with a cellobiose standard and the PJ assay.

3.5.9 Stability Measurements

In order to determine whether increases in specific activity were due to increases in thermostability some of the purified enzymes were incubated for 0, 3, 6, and 9 hours at 50°C and then assayed for 2 hours at 30°C to determine whether enzyme denaturation was occurring under the assay conditions (50°C for 2 hours). A master mix of 550 μl containing 1.8 μM protein in 50 mM sodium acetate buffer pH 5.0 was prepared for each protein sample and then 100 μl of this solution were distributed to five separate 1.7 ml eppendorf tubes. Two of these tubes were placed on ice and the other three placed in a thermomixer at 50°C and stirred at 300 RPM for 3, 6, and 9 hours, respectively. Once a tube was removed from the thermomixer it was placed back on ice until all three tubes had completed their time at 50°C. At this point 60 μl of an Avicel slurry containing 50

mg/ml of Avicel and 134 mM NaCl was added to each of the tubes and they were then placed back in the thermomixer at 30°C at 300 RPM for 2 hours. After 2 hours the tubes were centrifuged and two 50 μl aliquots from each tube were placed in a PCR plate and the activity on Avicel measured as described in section 3.5.6. The amount of activity remaining relative to the sample that was kept on ice corresponds to the amount of protein that denatures at 50°C after the amount of time the tube was kept in the thermomixer at 50°C.

3.6 Acknowledgments

I would like to thank Dr. Pete Heinzelman for providing native and chimeric genes. I would like to thank Indira Wu for providing me with the experimental protocols she developed to express, screen, and purify the cellulases. In particular I would like to thank Indira Wu for invaluable insight and advice. I would like to thank Dr. Arnold for guidance and the NSF for financial support.

3.7 Supplementary Material

Table 3.S1: Summary of the sequences and T_{50}s of the 38 chimeras and two parental Cel6As used in the initial characterization performed to select the chimeras to mutate. The 8-digit chimera sequence specifies the parental origin of the blocks.

chimera	*sequence*	T_{50} *(°C)*	*chimera*	*sequence*	T_{50} *(°C)*
Chimera 3	11332333	65.3	Chimera 68	13322332	69.8
Chimera 12	22232132	68.0	Chimera 73	12111332	68.0
Chimera 14	33213332	66.0	Chimera 75	12311332	69.5
Chimera 15	23233133	61.0	Chimera 77	12131332	68.8
Chimera 18	13231111	63.3	Chimera 78	13131332	70.0
Chimera 20	12213111	63.3	Chimera 79	12331332	70.0
Chimera 23	31311112	61.0	Chimera 81	12112332	68.0
Chimera 35	22212231	62.0	Chimera 82	13112332	67.0
Chimera 41	12133333	64.0	Chimera 84	13312332	70.0
Chimera 42	13333232	67.3	Chimera 85	12132332	69.8
Chimera 47	23311333	66.0	Chimera 86	13132332	70.5
Chimera 48	33133132	65.0	Chimera 87	12332332	69.0
Chimera HJ^+	12222332	71.3	Chimera 88	13332332	69.8
Chimera 52	12112132	69.8	Chimera 89	11113332	70.0
Chimera 53	12111131	69.3	Chimera 90	12113132	70.5
Chimera 54	12132331	69.8	Chimera 91	13113132	70.0
Chimera 55	12131331	68.8	Chimera 92	11111132	70.8
Chimera 56	12332331	66.8	Chimera 93	11112132	70.3
Chimera 60	13332331	69.5	P1	11111111	64.8
Chimera 66	22311331	68.0	P3	33333333	64

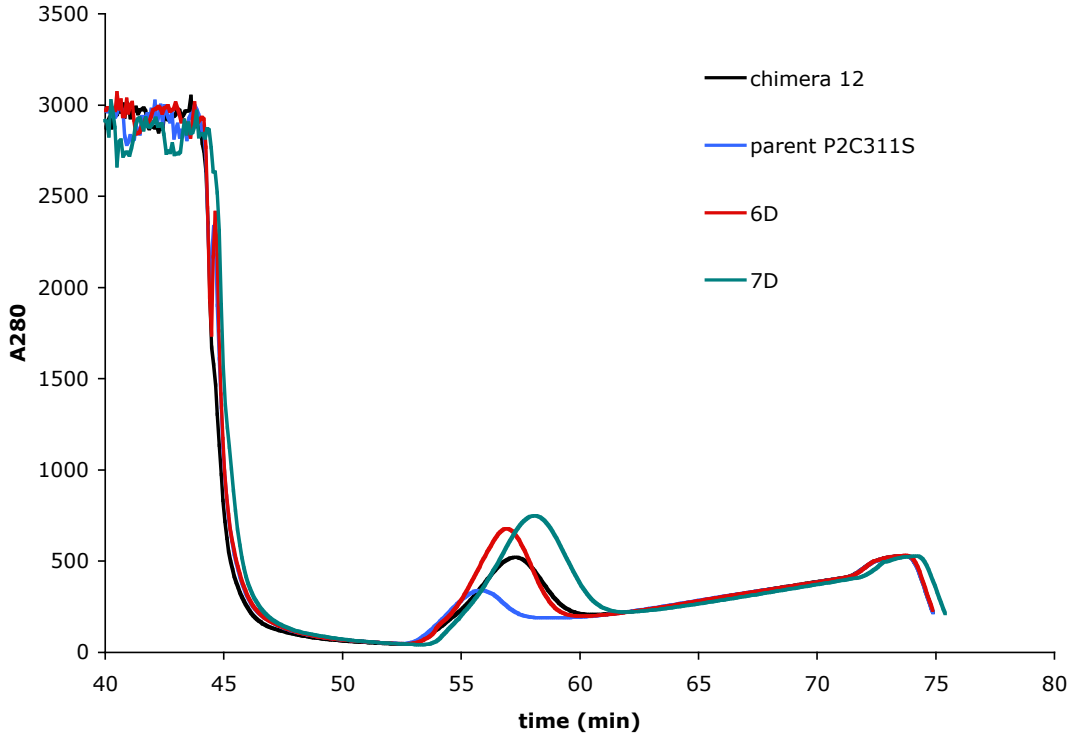


Figure 3.S1: Purification profile of chimera 12, parent P3C311S, and a selected set of enzyme mutants selected from the recombination library.

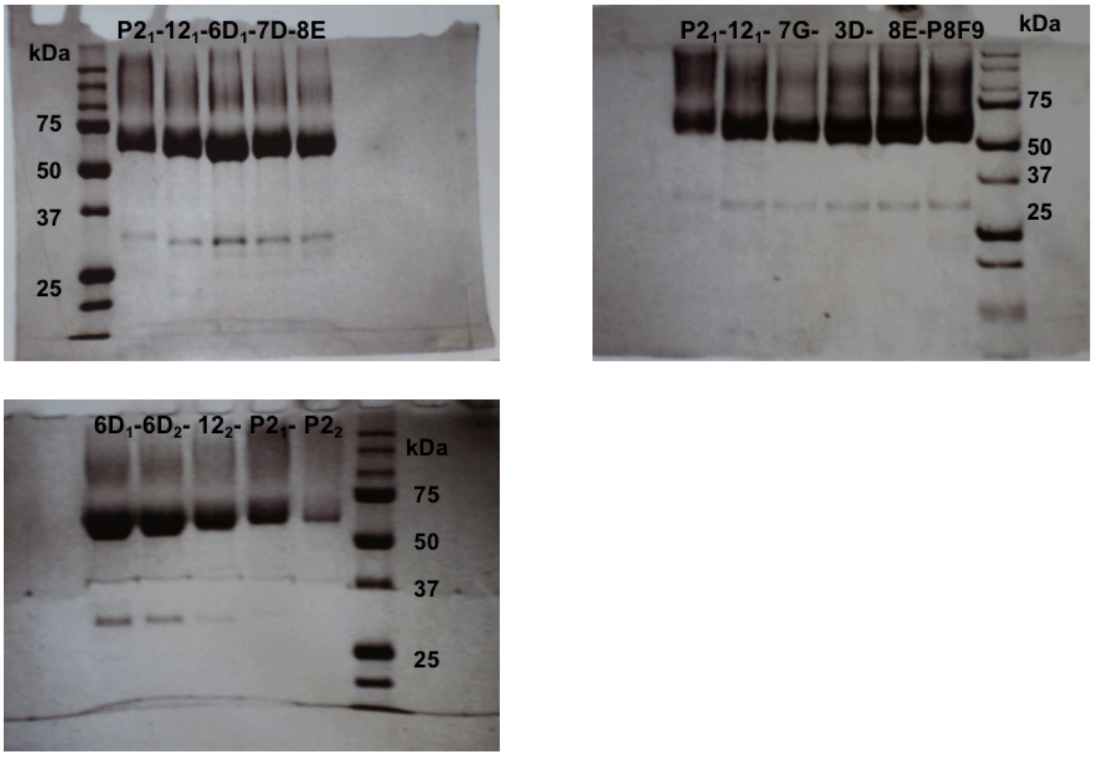


Figure 3.S2: SDS-PAGE gels of purified chimera 12, parent P2C311S (P2), the top mutants from the recombination library, 2C, 3D, 6D, 7D, 7G, and 8E, and the top mutant from the error-prone library, P8F9. Chimera 12, parent P2C311S (P2), and 6D were purified twice as indicated by the subscript.

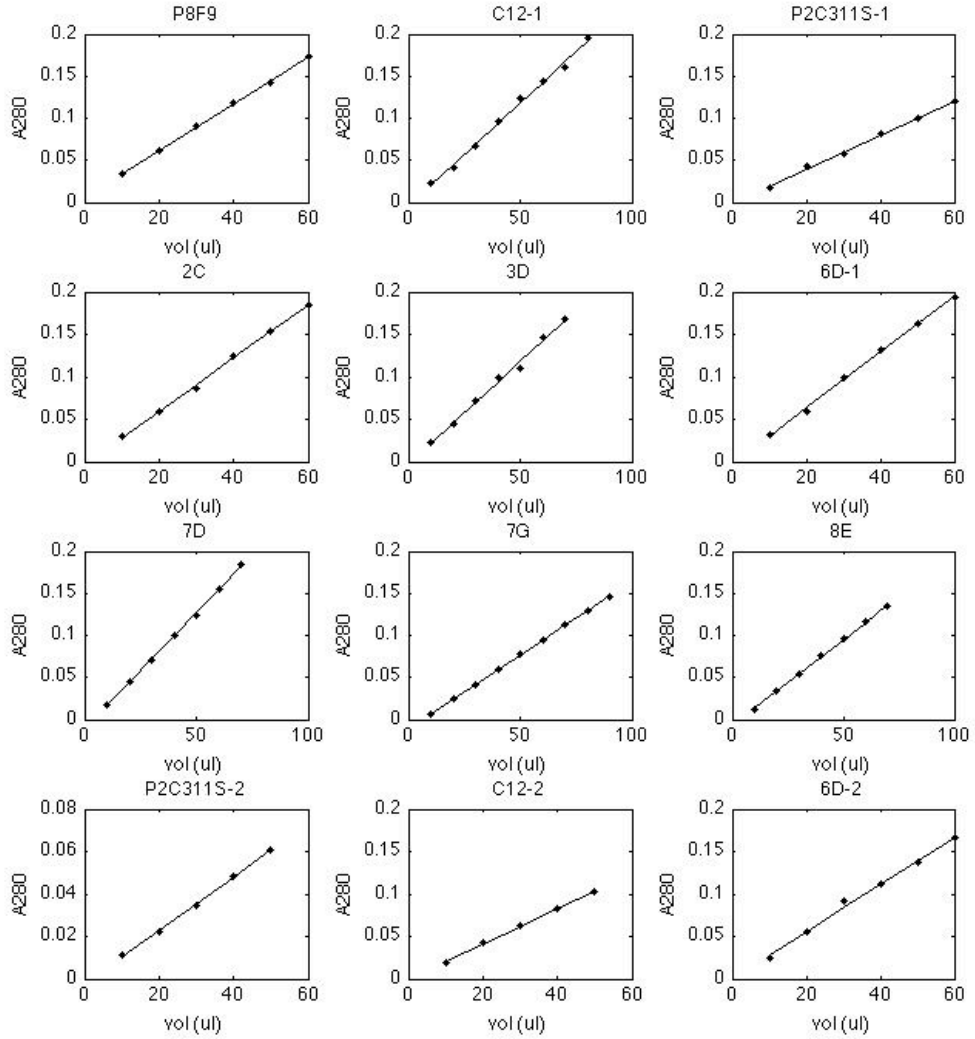


Figure 3.S3: Linear curves used to estimate the concentration of the purified proteins using the molar extinction coefficient of 92,425 $M^{-1}cm^{-1}$ [105] as described in the EM.

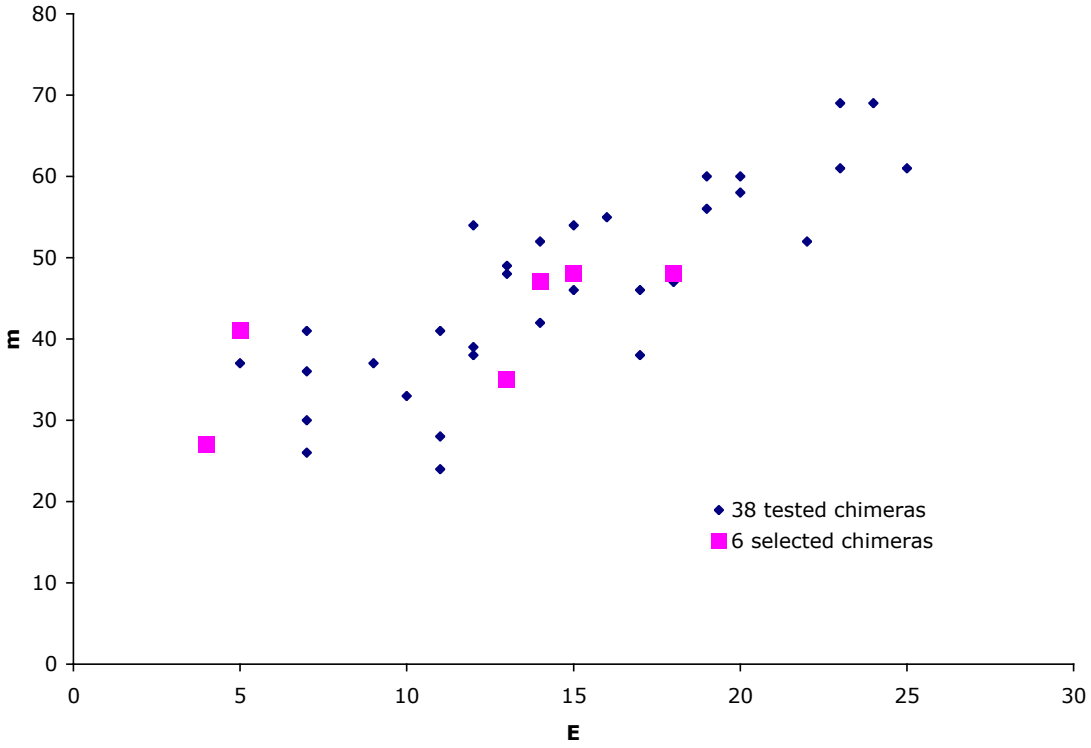


Figure 3.S4: Average *E* and *m* of the six chimeras selected for their similarity in whole cell activity on Avicel relative to P1 and P3.

Bibliography

1. Drummond DA, Silberg JJ, Meyer MM, Wilke CO, Arnold FH: **On the conservative nature of intragenic recombination**. *Proc. Natl. Acad. Sci. U S A* 2005, **102**:5380-5385.
2. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH: **Protein building blocks preserved by recombination**. *Nat. Struct. Biol.* 2002, **9**:553-558.
3. Otey CR, Landwehr M, Endelman JB, Hiraga K, Bloom JD, Arnold FH: **Structure-guided recombination creates an artificial family of cytochromes P450**. *PLoS Biol* 2006, **4**:e112.
4. Endelman JB, Silberg JJ, Wang ZG, Arnold FH: **Site-directed protein recombination as a shortest-path problem**. *Protein Eng. Des. Sel.* 2004, **17**:589-594.
5. Meyer MM, Hochrein L, Arnold FH: **Structure-guided SCHEMA recombination of distantly related beta-lactamases**. *Protein Eng. Des. Sel.* 2006, **19**:563-570.
6. Landwehr M, Carbone M, Otey CR, Li Y, Arnold FH: **Diversification of catalytic function in a synthetic family of chimeric cytochrome p450s**. *Chem. Biol.* 2007, **14**:269-278.
7. Taly V, Urban P, Truan G, Pompon D: **A combinatorial approach to substrate discrimination in the P450 CYP1A subfamily**. *Biochim. Biophys. Acta* 2006.
8. Raillard S, Krebber A, Chen Y, Ness JE, Bermudez E, Trinidad R, Fullem R, Davis C, Welch M, Seffernick J, et al.: **Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes**. *Chem. Biol.* 2001, **8**:891-898.
9. Ernstgard L, Warholm M, Johanson G: **Robustness of chlorzoxazone as an in vivo measure of cytochrome P450 2E1 activity**. *Br. J. Clin. Pharmacol.* 2004, **58**:190-200.
10. Hansson LO, Bolton-Grob R, Massoud T, Mannervik B: **Evolution of differential substrate specificities in Mu class glutathione transferases probed by DNA shuffling**. *J. Mol. Biol.* 1999, **287**:265-276.
11. Christians FC, Scapozza L, Crameri A, Folkers G, Stemmer WP: **Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling**. *Nat. Biotechnol.* 1999, **17**:259-264.
12. Griswold KE, Kawarasaki Y, Ghoneim N, Benkovic SJ, Iverson BL, Georgiou G: **Evolution of highly active enzymes by homology-independent recombination**. *Proc. Natl. Acad. Sci. USA* 2005, **102**:10082-10087.
13. Broo K, Larsson AK, Jemth P, Mannervik B: **An ensemble of theta class glutathione transferases with novel catalytic properties generated by stochastic recombination of fragments of two mammalian enzymes**. *J. Mol. Biol.* 2002, **318**:59-70.
14. Drummond DA, Silberg JJ, Meyer MM, Wilke CO, Arnold FH: **On the conservative nature of intragenic recombination**. *Proc. Natl. Acad. Sci. USA* 2005, **102**:5380-5385.

15. Otey CR, Landwehr M, Endelman JB, Hiraga K, Bloom JD, Arnold FH: **Structure-guided recombination creates an artificial family of cytochromes p450**. *PLoS Biol*. 2006, **4**:e112.
16. Otey CR, Silberg JJ, Voigt CA, Endelman JB, Bandara G, Arnold FH: **Functional evolution and structural conservation in chimeric cytochromes p450: calibrating a structure-guided approach**. *Chem. Biol*. 2004, **11**:309-318.
17. Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, Arnold FH: **A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments**. *Nat. Biotechnol*. 2007, **25**:1051-1056.
18. Brock BJ, Waterman MR: **The use of random chimeragenesis to study structure/function properties of rat and human P450c17**. *Arch. Biochem. Biophys*. 2000, **373**:401-408.
19. Ramarao MK, Straub P, Kemper B: **Identification by in vitro mutagenesis of the interaction of two segments of C2MstC1, a chimera of cytochromes P450 2C2 and P450 2C1**. *J. Biol. Chem*. 1995, **270**:1873-1880.
20. Lewis DFV: *Guide to Cytochromes P450 : Structure and Function*. London ; New York: Taylor & Francis; 2001.
21. Mansuy D: **The great diversity of reactions catalyzed by cytochromes P450**. *Comp. Biochem. Physiol. C-Pharmacol. Toxicol. Endocrinol*. 1998, **121**:5-14.
22. Ruettinger RT, Wen LP, Fulco AJ: **Coding nucleotide, 5' regulatory, and deduced amino-acid sequences of P-450BM-3, a single peptide cytochrome P-3´450-NADPH-P-450 reductase from *Bacillus megaterium***. *J. Biol. Chem*. 1989, **264**:10987-10995.
23. Ortiz de Montellano PR: *Cytochrome P450: Structure, Mechanism, and Biochemistry*. New York: Plenum Press; 1995.
24. Cirino PC, Arnold FH: **Regioselectivity and activity of cytochrome P450BM-3 and mutant F87A in reactions driven by hydrogen peroxide**. *Adv. Synth. Catal*. 2002, **344**:932-937.
25. Gustafsson MC, Roitel O, Marshall KR, Noble MA, Chapman SK, Pessegueiro A, Fulco AJ, Cheesman MR, von Wachenfeldt C, Munro AW: **Expression, purification, and characterization of *Bacillus subtilis* cytochromes P450 CYP102A2 and CYP102A3: Flavocytochrome homologues of P450 BM3 from *Bacillus megaterium***. *Biochemistry* 2004, **43**:5474-5487.
26. Otey CR, Bandara B, Lalonde J, Takahashi K, Arnold FH: **Preparation of human metabolites of propranolol using laboratory-evolved bacterial cytochromes P450**. *Biotechnol. Bioeng*. 2006, **93**:494-499.
27. Lee CR, Pieper JA, Frye RF, Hinderliter AL, Blaisdell JA, Goldstein JA: **Tolbutamide, flurbiprofen, and losartan as probes of CYP2C9 activity in humans**. *J. Clin. Pharmacol*. 2003, **43**:84-91.
28. Otey CR, Joern JM: **High-throughput screen for aromatic hydroxylation**. *Methods Mol. Biol*. 2003, **230**:141-148.
29. Schwaneberg U, Schmidt-Dannert C, Schmitt J, Schmid RD: **A continuous spectrophotometric assay for P450 BM-3, a fatty acid hydroxylating enzyme, and its mutant F87A**. *Anal. Biochem*. 1999, **269**:359-366.

30. McQueen J: **Some methods for classification and analysis of multivariate observations**. In *5th Berkeley Symposium on Mathematics, Statistics and Probability*: 1967:281-297.
31. Larsson AK, Emren LO, Bardsley WG, Mannervik B: **Directed enzyme evolution guided by multidimensional analysis of substrate-activity space**. *Protein Eng. Des. Sel.* 2004, **17**:49-55.
32. Kolodny R, Koehl P, Guibas L, Levitt M: **Small libraries of protein fragments model native protein structures accurately**. *J. Mol. Biol.* 2002, **323**:297-307.
33. Sykes MT, Levitt M: **Describing RNA structure by libraries of clustered nucleotide doublets**. *J. Mol. Biol.* 2005, **351**:26-38.
34. Shet MS, Fisher CW, Holmans PL, Estabrook RW: **Human cytochrome P450 3A4: Enzymatic properties of a purified recombinant fusion protein containing NADPH-P450 reductase**. *Proc. Natl. Acad. Sci. USA* 1993, **90**:11748-11752.
35. Shet MS, Fisher CW, Arlotto MP, Shackleton CH, Holmans PL, Martin-Wixtrom CA, Saeki Y, Estabrook RW: **Purification and enzymatic properties of a recombinant fusion protein expressed in *Escherichia coli* containing the domains of bovine P450 17A and rat NADPH-P450 reductase**. *Arch. Biochem. Biophys.* 1994, **311**:402-417.
36. Harlow GR, Halpert JR: **Mutagenesis study of Asp-290 in cytochrome P450 2B11 using a fusion protein with rat NADPH-cytochrome P450 reductase**. *Arch. Biochem. Biophys.* 1996, **326**:85-92.
37. Fairhead M, Giannini S, Gillam EM, Gilardi G: **Functional characterisation of an engineered multidomain human P450 2E1 by molecular Lego**. *J. Biol. Inorg. Chem.* 2005, **10**:842-853.
38. Fuziwara S, Sagami I, Rozhkova E, Craig D, Noble MA, Munro AW, Chapman SK, Shimizu T: **Catalytically functional flavocytochrome chimeras of P450 BM3 and nitric oxide synthase**. *J. Inorg. Biochem.* 2002, **91**:515-526.
39. Kubota M, Nodate M, Yasumoto-Hirose M, Uchiyama T, Kagami O, Shizuri Y, Misawa N: **Isolation and functional analysis of cytochrome P450 CYP153A genes from various environments**. *Biosci. Biotechnol. Biochem.* 2005, **69**:2421-2430.
40. Sevrioukova IF, Li H, Zhang H, Peterson JA, Poulos TL: **Structure of a cytochrome P450-redox partner electron-transfer complex**. *Proc. Natl. Acad. Sci. USA* 1999, **96**:1863-1868.
41. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs**. *Nuc. Acids Res.* 2003, **31**:3497-3500.
42. Klein ML, Fulco AJ: **Critical residues involved in FMN binding and catalytic activity in cytochrome P450BM-3**. *J. Biol. Chem.* 1993, **268**:7553-7561.
43. The PyMOL Molecular Graphics System - http://www.pymol.org.
44. Cytochrome P450 Homepage - http://drnelson.utmem.edu/CytochromeP450.html.
45. Higuchi R, Krummel B, Saiki RK: **A general-method of in vitro preparation and specific mutagenesis of DNA fragments - study of protein and DNA interactions**. *Nuc. Acids Res.* 1988, **16**:7351-7367.

46. Barnes HJ, Arlotto MP, Waterman MR: **Expression and enzymatic-activity of recombinant cytochrome- P450 17-alpha-hydroxylase in *Escherichia-coli***. *Proc. Natl. Acad. Sci. USA* 1991, **88**:5597-5601.
47. Schwaneberg U, Sprauer A, Schmidt-Dannert C, Schmid RD: **P450 monooxygenase in biotechnology I: Single-step, large-scale purification method for cytochrome P450 BM-3 by anion-exchange chromatography,** *J. Chromatogr. A*. 1999, **848**:149-159.
48. Otey CR: **High-throughput carbon monoxide binding assay for cytochromes P450**. In *Directed Enzyme Evolution: Screening and Selection Methods*. Edited by Arnold FH, Georgiou G: Humana Press; 2003:137-139.
49. Cirino PC, Arnold FH: **A self-sufficient peroxide-driven hydroxylation biocatalyst**. *Angew. Chem. Int. Ed. Engl*. 2003, **42**:3299-3301.
50. Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis**. *J. Comput. Appl. Math*. 1987, **20**:53-65.
51. Krupp F, Horn M: *Earth, The Sequel: The Race to Reinvent Energy and Stop Global Warming*. New York: W. W. Norton & Company; 2008.
52. Zhang YHP, Himmel ME, Mielenz JR: **Outlook for cellulase improvement: Screening and selection strategies**. *Biotechnol. Adv*. 2006, **24**:452-481.
53. Heinzelman P, Snow CD, Wu I, Nguyen C, Villalobos A, Govindarajan S, Minshull J, Arnold FH: **A family of thermostable fungal cellulases created by structure-guided recombination**. *Proc. Natl. Acad. Sci. U S A* 2009, **106**:5610-5615.
54. Taverna DM, Goldstein RA: **Why are proteins marginally stable?** *Proteins Struct. Funct. Genet*. 2002, **46**:105-109.
55. Bloom JD, Wilke CO, Arnold FH, Adami C: **Stability and the evolvability of function in a model protein**. *Biophys. J*. 2004, **86**:2758-2764.
56. Bornberg-Bauer E: **How are model protein structures distributed in sequence space?** *Biophys. J*. 1997, **73**:2393-2403.
57. Bornberg-Bauer E, Chan H: **Modeling evolutionary landscapes: Mutational stability, topology and superfunnels in sequence space.** *Proc. Natl. Acad. Sci. U S A* 1999, **96**:10689-10694.
58. Amin N, Liu AD, Ramer S, Aehle W, Meijer D, Metin M, Wong S, Gualfetti P, Schellenberger V: **Construction of stabilized proteins by combinatorial consensus mutagenesis**. *Protein Eng. Des. Sel*. 2004, **17**:787-793.
59. Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon APGM, Wyss M: **The consensus concept for thermostability engineering of proteins: further proof of concept**. *Protein Eng*. 2002, **15**:403-411.
60. Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L, van Loon APGM: **From DNA sequence to improved functionality: Using protein sequence comparisons to rapidly design a thermostable consensus phytase**. *Protein Eng*. 2000, **13**:49-57.
61. Lehmann M, Pasamontes L, Lassen SF, Wyss M: **The consensus concept for thermostability engineering of proteins**. *Biochim. Biophys. Acta* 2000, **1543**:408-415.
62. Bloom JD, Labthavikul ST, Otey CR, Arnold FH: **Protein stability promotes evolvability**. *Proc. Natl. Acad. Sci. U S A* 2006, **103**:5869-5874.

63. Chan H, Bornberg-Bauer E: **Perspectives on protein evolution from simple exact models**. *Appl. Bioinformatics* 2002, **1**:121-144.
64. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH: **Thermodynamic prediction of protein neutrality**. *Proc. Natl. Acad. Sci. U S A* 2005, **102**:606-611.
65. Miller DW, Dill KA: **Ligand binding to proteins: the binding landscape model**. *Protein Sci*. 1997, **6**:2166-2179.
66. Xia Y, Levitt M: **Simulating protein evolution in sequence and structure space**. *Curr. Opin. Struct. Biol*. 2004, **14**:202-207.
67. Miyazawa S, Jernigan RL: **Estimation of effective interresidue contact energies from protein crystal-structures - quasi-chemical approximation**. *Macromolecules* 1985, **18**:534-552.
68. da Silva S, Coetzer M, Nedellec R, Pastore C, Mosier D: **Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region**. *Genetics* 2010.
69. Weinreich D, Delaney N, DePristo M, Hartl D: **Darwinian evolution can follow only very few mutational paths to fitter proteins**. *Science* 2006, **312**:111-114.
70. Bridgham J, Ortlund E, Thornton J: **An epistatic ratchet constrains the direction of glucocorticoid receptor evolution**. *Nature* 2009, **461**:515-520.
71. Wells J: **Additivity of mutational effects in proteins**. *Biochemistry* 1990, **27**:8509-8517.
72. Oue S, Okamoto A, Yano T, Kagamiyama H: **Redesigning the substrate specificity of an enzyme by cumulative effects of the mutations of non-active site residues**. *J. Biol. Chem*. 1999, **274**:2344-2349.
73. Yano T, Oue S, Kagamiyama H: **Directed evolution of an aspartate aminotransferase with new substrate specificities.** *Proc. Natl. Acad. Sci. U S A* 1998, **95**:5511-5515.
74. Shimotohno A, Oue S, Yano T, Kuramitsu S, Kagamiyama H: **Demonstration of the importance and usefulness of manipulating non-active-site residues in protein design**. *J. Biochem*. 2001, **129**:943-948.
75. Reetz M, Puls M, Carballeira J, Vogel A, Jaeger K, Eggert T, Thiel W, Bocola M, Otte N: **Learning from directed evolution: further lessons from theoretical investigations into cooperative mutations in lipase enantioselectivity**. *Chembiochem*. 2007, **8**:106-112.
76. Massey FJ: **The Kolmogorov-Smirnov Test for Goodness of Fit**. *J. Amer. Statistical Assoc*. 1951, **46**:68-78.
77. Meyer MM, Silberg JJ, Voigt CA, Endelman JB, Mayo SL, Wang ZG, Arnold FH: **Library analysis of SCHEMA-guided protein recombination**. *Protein Sci*. 2003, **12**:1686-1693.
78. Arnold FH: **The race for new biofuels**. *Engineering & Science* 2008, **2**:12-19.
79. U. S. Department of Energy- http://www.energy.gov/
80. Wang T, Liu X, Yu Q, Zhang X, Qu Y, Gao P, Wang T: **Directed evolution for engineering pH profile of endoglucanase III from *Trichoderma reesei***. *Biomol. Eng*. 2005, **22**:89-94.

81. Qin YQ, Wei XM, Song X, Qu YB: **Engineering endoglucanase II from *Trichoderma reesei* to improve the catalytic efficiency at a higher pH optimum**. *J. Biotechnol.* 2008, **135**:190-195.
82. Becker D, Braet C, Brumer H, Claeyssens M, Divne C, Fagerstrom BR, Harris M, Jones TA, Kleywegt GJ, Koivula A, et al.: **Engineering of a glycosidase Family 7 cellobiohydrolase to more alkaline pH optimum: the pH behaviour of *Trichoderma reesei* CeI7A and its E223S/A224H/L225V/T226A/D262G mutant**. *Biochem. J.* 2001, **356**:19-30.
83. Cockburn D, Vandenende C, Clarke A: **Modulating the pH-activity profile of cellulase by substitution: replacing the general base catalyst aspartate with cysteinesulfinate in cellulase A from *Cellulomonas fimi*.** *Biochemistry* 2010, **49**:2042-2050.
84. Murashima K, Kosugi A, Doi RH: **Thermostabilization of cellulosomal endoglucanase EngB from *Clostridium cellulovorans* by in vitro DNA recombination with non-cellulosomal endoglucanase EngD**. *Mol. Microbiol.* 2002, **45**:617-626.
85. Voutilainen SP, Murray PG, Tuohy MG, Koivula A: **Expression of *Talaromyces emersonii* cellobiohydrolase Cel7A in *Saccharomyces cerevisiae* and rational mutagenesis to improve its thermostability and activity**. *Protein Eng. Des. Sel.* 2010, **23**:69-79.
86. Lebbink JHG, Kaper T, Bron P, van der Oost J, de Vos WM: **Improving low-temperature catalysis in the hyperthermostable *Pyrococcus furiosus* beta-glucosidase CelB by directed evolution**. *Biochemistry* 2000, **39**:3656-3665.
87. Gonzalez-Blasco G, Sanz-Aparicio J, Gonzalez B, Hermoso JA, Polaina J: **Directed evolution of beta-glucosidase A from *Paenibacillus polymyxa* to thermal resistance**. *J. Biol. Chem.* 2000, **275**:13708-13712.
88. Arrizubieta MJ, Polaina J: **Increased thermal resistance and modification of the catalytic properties of a beta-glucosidase by random mutagenesis and in vitro recombination**. *J. Biol. Chem.* 2000, **275**:28843-28848.
89. Nemeth A, Kamondi S, Szilagyi A, Magyar C, Kovari Z, Zavodszky P: **Increasing the thermal stability of cellulase C using rules learned from thermophilic proteins: a pilot study**. *Biophys. Chem.* 2002, **96**:229-241.
90. Heinzelman P, Snow CD, Smith MA, Yu XL, Kannan A, Boulware K, Villalobos A, Govindarajan S, Minshull J, Arnold FH: **SCHEMA Recombination of a Fungal Cellulase Uncovers a Single Mutation That Contributes Markedly to Stability**. *J. Biol. Chem.* 2009, **284**:26229-26233.
91. Voutilainen SP, Boer H, Alapuranen M, Janis J, Vehmaanpera J, Koivula A: **Improving the thermostability and activity of *Melanocarpus albomyces* cellobiohydrolase Cel7B**. *Appl. Microbiol. Biotechnol.* 2009, **83**:261-272.
92. Nakazawa H, Okada K, Onodera T, Ogasawara W, Okada H, Morikawa Y: **Directed evolution of endoglucanase III (Cel12A) from *Trichoderma reesei***. *Appl. Microbiol. Biotechnol.* 2009, **83**:649-657.
93. Kim YS, Jung HC, Pan JG: **Bacterial cell surface display of an enzyme library for selective screening of improved cellulase variants**. *Appl. Environ. Microbiol.* 2000, **66**:788-793.

94. McCarthy JK, Uzelac A, Davis DF, Eveleigh DE: **Improved catalytic efficiency and active site modification of 1,4-beta-D-glucan glucohydrolase A from *Thermotoga neapolitana* by directed evolution**. *J. Biol. Chem.* 2004, **279**:11495-11502.
95. Escovar-Kousen J, Wison D, Irwin D: **Integration of computer modeling and initial studies of site-directed mutagenesis to improve cellulase activity on Cel9A from *Thermobifica fusca*.** *Appl. Biochem. Biotechnol.* 2004, **113-116**:287-297.
96. Zhang S, Barr B, Wilson D: **Effects of noncatalytic residue mutations on substrate specificity and ligand binding of *Thermobifida fusca* endocellulase Cel6A**. *Eur. J. Biochem.* 2000, **267**.
97. Zhang S, Irwin D, Wison D: **Site-directed mutation of noncatalytic residues of *Thermobifida fusca* exocellulase Cel6B**. *Eur. J. Biochem.* 2000, **267**:3101-3115.
98. Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS: **Microbial cellulose utilization: Fundamentals and biotechnology (vol 66, pg 506, 2002)**. *Microbiol. Mol. Biol. Rev.* 2002, **66**:739-739.
99. Mildvan A, Weber D, Kuliopulos A: **Quantitative interpretations of double mutations of enzymes.** *Arch. Biochem. Biophys.* 1992, **294**:327-340.
100. Istomin A, Gromiha M, Vorov O, Jacobs D, Livesay D: **New insight into long-range nonadditivity within protein double-mutant cycles**. *Protein Struct. Funct. Bioinf.* 2008, **70**:915-924.
101. Serrano L, Day A, Fersht A: **Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability.** *J. Mol. Biol.* 1993, **233**:305-312.
102. Hidalgo P, MacKinnon R: **Revealing the architecture of a K+ channel pore through mutant cycles with a peptide inhibitor.** *Science* 1995, **268**:307-310.
103. Varrot A, Macdonald J, Stick RV, Pell G, Gilbert HJ, Davies GJ: **Distortion of a cellobio-derived isofagomine highlights the potential conformational itinerary of inverting beta-glucosidases**. *Chemical Communications* 2003:946-947.
104. Wood T, Bhat K: **Methods for measuring cellulase activities**. *Methods Enzymol.* 1988, **160**.
105. Swiss Institute of Bioinformatics: **ExPASy Proteomics Server- http://www.expasy.ch/tools/protparam.html**.
106. Park JT, Johnson MJ: **A Submicrodetermination of Glucose**. *J. Biol. Chem.* 1949, **181**:149-151.
107. Nelson N: **A photometric adaptation of the Somogyi method for the determination of glucose**. *J. Biol. Chem.* 1944, **153**:375-380.
108. Somogyi M: **Notes on Sugar Determination**. *J. Biol. Chem.* 1952, **195**:19-23.
109. Zhang X, Baase W, Shoichet B, Wilson K, Matthews B: **Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive.** *Protein Eng.* 1995, **8**:1017-1022.
110. Dahiyat BI, Mayo SL: **De novo protein design: Fully automated sequence selection**. *Science* 1997, **278**:82-87.

111. Dantas G, Kuhlman B, Callender D, Wong M, Baker D: **A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins**. *J. Mol. Biol.* 2003, **332**:449-460.
112. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, et al.: **Kemp elimination catalysts by computational enzyme design**. *Nature* 2008, **453**:190-U194.
113. Klyosov A: *Cellulases of the Third Generation. In: Biochemistry and genetics of cellulose degradation*. Edited by Aubert J, Beguin P, Millet J. London: Academic Press; 1988.
114. Mandels M: **Applications of Cellulases**. *Biochemical Society Transactions* 1985, **13**:414-416.
115. Bernardez TD, Lyford KA, Lynd LR: **Kinetics of the Extracellular Cellulases of Clostridium-Thermocellum Acting on Pretreated Mixed Hardwood and Avicel**. *Appl. Microbiol. Biotechnol.* 1994, **41**:620-625.
116. Zhang YHP, Lynd LR: **Toward an aggregated understanding of enzymatic hydrolysis of cellulose: Noncomplexed cellulase systems**. *Biotechnol. Bioeng.* 2004, **88**:797-824.
117. Chao G, Lau W, Hackel B, Sazinsky S, Lippow S, Wittrup K: **Isolation and engineering human antibodies using yeast surface display**. *Nat. Protoc.* 2006, **1**:755-768.