

Recombinatorial and Predictive Methods to Increase Cellulase Thermostability
and Structural Analysis of a Thermostable P450

Thesis by

Russell Scott Komor

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2012

(Defended December 14, 2011)

©2012

Russell Scott Komor

All Rights Reserved

Acknowledgments

Caltech has been a great place to study and work throughout my graduate career. I want to thank my thesis advisor, Frances Arnold, for giving me the opportunity to work in her group on challenging and cutting-edge projects with some of the top scientists in the world. I am grateful to the members of my thesis committee, David Tirrell, Stephen Mayo, and Harry Gray, for advice throughout this process. I am also grateful to the Department of Defense for conferring on me the National Defense Science and Engineering Graduate Fellowship and to the Rosen family for conferring on me the Benjamin M. Rosen Fellowship through Caltech.

The Arnold laboratory is filled with brilliant scientists who do not hesitate to teach younger members their skills. I am extremely grateful to Andrew Sawayama, Sabine Bastian, and Pete Heinzelman for teaching me the fundamentals of molecular biology and protein engineering. In my time here, I have also had the pleasure of working with Phil Romero, Christopher Snow, Eric Brustad, Mike Chen, Pavle Nikolovski, Jens Kaiser, Catherine Xie, Ahalya Prabakar, Avin Andrade, Dan Koch, and many others.

And finally, I want to thank my family for supporting me throughout all of my education. My parents, Peter and Carol, have always promoted an environment that encouraged learning and caused me to grow into the scientist I am today. My sister, Alexis, is also a PhD student at Caltech and has been a constant source of inspiration to me throughout my education.

Abstract

To address the world's need for improved biomass breakdown for the production of renewable fuel, we sought to improve cellulase thermostability and thereby enzyme lifetime, operating temperature, and specific activity. We created an eight block SCHEMA recombination library based on five fungal cellobiohydrolase class I (CBHI) enzymes. By characterizing this library, we identified several stabilizing sequence blocks and combined these to produce a set of well-expressed, thermostable CBHI chimeras. To further increase the stability of these chimeras, we used a combination of the chimera thermostability screening data, a consensus analysis of 40 naturally occurring CBHI sequences, and FoldX $\Delta\Delta G$ predictions to identify individual mutations for testing. Our final enzyme has a T_{50} 9.3 °C greater than that of the most stable parental CBHI, resulting in a 10 °C increase in optimal temperature and a 50% increase in total sugar production at the optimal temperature.

To produce an ideal parent for directed evolution for improved activity on varied compounds, we increased the thermostability of a P450_{BM3} enzyme with broad substrate specificity to produce enzyme 9-10ATS. Directed evolution libraries based on 9-10ATS produced variants with improved activity on a number of structurally diverse compounds. We determined the structure of 9-10ATS using x-ray crystallography and compared it to other P450_{BM3} structures. Examination of the structure shows clear structural basis for the thermostabilizing mutations and broad substrate specificity.

Table of Contents

Acknowledgements		iii
Abstract		iv
Table of Contents		v
Figures and Tables		vi
Chapters		
Chapter 1	<i>Efficient Screening of Fungal Cellobiohydrolase Class I Enzymes for Thermostabilizing Sequence Blocks by SCHEMA Structure-Guided Recombination</i>	1
Chapter 2	<i>A Combination of Predictive Methods to Identify Stabilizing Mutations in Cellobiohydrolase Class I Enzymes</i>	50
Chapter 3	<i>Construction and Structural Analysis of a Thermostable P450 Enzyme with Broad Substrate Specificity</i>	84
Chapter 4	<i>Materials and Methods</i>	129

Figures and Tables

Figure 1.1	RASPP curve for CBHI SCHEMA libraries	7
Figure 1.2	Predicted vs. measured T_{50} values for chimeric P450s	15
Figure 1.3	The structure of cellulose	17
Figure 1.4	Mechanism of cellulose hydrolysis by cellulases	19
Figure 1.5	Structure of the carbohydrate binding module	20
Figure 1.6	Structure of cellobiohydrolase class I	27
Table 1.1	Pairwise sequence alignment of CBHI parent enzymes	29
Table 1.2	CBHI parental enzyme properties	31
Figure 1.7	CBHI SCHEMA library diagram	32
Figure 1.8	Block effects on total secreted activity	34
Figure 1.9	Block effects on T_{50}	35
Table 1.3	Disulfide-paired CBHI chimera properties	37
Figure 1.10	Sub-block effects on total secreted activity	39
Figure 1.11	Sub-block effects on T_{50}	40
Table 1.4	Properties of thermostable chimeras	42
Table 1.5	Properties of thermostable chimeras containing sub-blocks	43
Figure 1.12	Multiple sequence alignment of CBHI parental enzymes	45
Table 2.1	Sub-block C sequence comparison	53
Table 2.2	Mutation properties	59
Table 2.3	Consensus and FoldX mutation properties	60
Figure 2.1	Predicted vs. measured values for FoldX and consensus mutations	61

Table 2.4	FoldX mutation properties	63
Table 2.5	Stabilizing mutation properties	64
Figure 2.2	Beta sheet containing three stabilizing mutations in CBHI	67
Figure 2.3	Hydrophobic pocket around residues 425 in CBHI	68
Figure 2.4	Mutation N92K in CBHI	69
Figure 2.5	Temperature profiles of thermostable CBHIs on 4-methylumbelliferyl lactopyranoside	71
Figure 2.6	Temperature profiles of thermostable CBHIs on microcrystalline cellulose	72
Figure 2.7	Multiple sequence alignment of 40 CBHIs	75
Figure 3.1	Distribution of mutations' effects on protein stability	90
Figure 3.2	Screening compounds' structures and properties	96
Figure 3.3	Conversions of 9-10A variants on screening compounds	97
Table 3.1	Sequences of thermostable 9-10A variants	100
Figure 3.4	Half-lives at 50 °C of thermostable 9-10A variants	100
Figure 3.5	Total activity of thermostable 9-10A variants	101
Figure 3.6	Structures of compounds used in library design and screening	103
Table 3.2	Sequence-activity relationship for alanine substitution	104
Table 3.3	Substrate scope and reaction selectivity	106
Table 3.4	Screening conditions for crystallization of 9-10ATS	108
Table 3.5	Data collection and refinement statistics for 9-10ATS structure	110
Figure 3.7	Pairwise, residue RMSD values for P450 _{BM3} crystal structures	112
Figure 3.8	P450 crystal structure B-factors by residue	116

Figure 3.9	Mutation V78A in P450 _{BM3}	117
Figure 3.10	Thermostabilizing mutations in P450 _{BM3}	119
Table 3.6	P450 _{BM3} variants used in compound screening	121
Figure 4.1	Plasmid construct used for expression of CBHI genes	131
Table 4.1	Primer list for cellulase constructs	133
Figure 4.2	Plasmid construct used for expression of P450 genes	140
Table 4.2	Primer list for P450 constructs	142
Figure 4.3	Formaldehyde detection scheme	146
Figure 4.4	Aromatic alcohol detection scheme	147
Figure 4.5	SDS-PAGE gel showing P450 purity	150

Chapter 1

Efficient Screening of Fungal Cellobiohydrolase

Class I Enzymes for Thermostabilizing Sequence

Blocks by SCHEMA Structure-Guided

Recombination

Abstract

Homologous recombination is ideal for creating diverse sets of chimeric proteins with varied secondary properties, such as stability, while maintaining parental fold and function. SCHEMA structure-guided recombination produces chimera libraries with minimal average disruption and thus maximal fraction of folded enzymes. A further consequence of SCHEMA recombination is that the recombination blocks display additivity toward properties such as stability, allowing simple regression models to accurately predict chimera properties. To address the world's need for improved biomass breakdown for the production of renewable fuel, we applied SCHEMA recombination to a cellulase enzyme to create more stable cellulases. Increased stability results in longer enzyme lifetimes, increased operating temperatures, and higher specific activity. Our chimera library of cellobiohydrolase class I (CBHI) enzymes is based on proteins from five thermophilic fungi, each broken into eight recombination blocks. To characterize the library, we created a set of 32 monomers, where blocks from four parents are substituted one at a time into the most stable and highly expressed parent. Substitution at block 7, the largest block, resulted in a total loss of expression. However, examination of smaller sub-blocks of block 7 revealed that some have stabilizing effects. From these results, we designed and constructed a set of thermostable chimeras with T_{50S} up to 4.6 °C higher than the most stable parent.

A. Homologous Recombination

Homologous recombination is a powerful tool for producing large sets of novel proteins with varying properties. Because the active site and other essential regions of proteins are conserved even among distantly related proteins, recombination causes variation mostly on the surface and other noncritical portions of the protein. This results in changes to secondary properties of the protein such as stability, substrate specificity, and expression without drastically affecting protein fold or activity. Homologous recombination is thus ideally suited for constructing diverse families of proteins with varied properties while preserving core function and fold.

By combining blocks of sequence from related proteins, recombination produces proteins that differ from each other and from the parental proteins at dozens to hundreds of positions. The reason that many simultaneous mutations are possible is that all of the mutations accessible by recombination are compatible with the protein fold in at least one of the wild type proteins. The accepted theory that most random mutations are destabilizing¹ therefore does not apply to recombination: these mutations are not random. Because of its conservative nature, recombination may at first seem ill suited to producing proteins with properties outside the range of the parent proteins, but this has proven not to be the case. Homologous recombination has been successfully used to create chimeras with stability greater² than and activity unseen^{3; 4; 5} in any of the parents.

B. Structure-Guided Recombination

There are many ways to choose recombination breakpoints, including randomly. However, breakpoints chosen randomly can lead to a low fraction of folded chimeras. By intelligently choosing recombination breakpoints, one can maximize the chance of producing a highly functional library. The multitude of three-dimensional structures available makes it likely that a structure of at least one of the parental proteins will exist. These structures can be used in conjunction with design algorithms to predict how recombination at different positions will affect the folded protein.

SCHEMA is a general method for using structural information to predict optimal crossover positions for a recombination library⁶. The form of SCHEMA most often used predicts and minimizes structural disruption by identifying important residue pairs and conserving their association during recombination. SCHEMA's only inputs are a three-dimensional structure of the protein fold and a multiple sequence alignment of all the parental amino acid sequences. Our current implementation of SCHEMA uses the 3D structure and alignment to create a contact matrix for the set of parents, defining a contact as a pair of amino acids with heavy atoms within 4.5 Å of each other⁷. The contact matrix is thus a sum total of all interacting residues from any parent, regardless of in how many parents a contact appears. The contact matrix can be created using any structure, but the choice of structure can affect SCHEMA design. Ideally, a crystal structure for one of the parents will be available, but homology models can also be constructed and used. These models are based on combining the information from crystal structures of many proteins related to the parental enzymes in the hopes of

creating a suitable approximation. With a contact matrix defined, SCHEMA calculates the number of disruptions (SCHEMA energy, E) upon recombination, defining a disruption as an amino acid pair not present in any of the parents at that contact point. Each chimera's SCHEMA energy is then the sum total of disruptions present. For a chimera containing residues i in fragment α , and residues j , in fragment β from a different parent, the SCHEMA energy is given by

$$E = \sum_{i \in \alpha} \sum_{j \in \beta} c_{ij} \Delta_{ij} \quad \text{Eq. 1.1}$$

where $c_{ij} = 1$ if residues i and j are contacting; otherwise $c_{ij} = 0$. And the SCHEMA delta function $\Delta_{ij} = 0$ if residues i and j are identical in any of the parents; otherwise $\Delta_{ij} = 1$. Meyer showed that SCHEMA energy correlates inversely with probability of folding⁶. Thus, minimizing E should maximize the fraction of folded proteins in a given recombination library. However, the library with minimal E for any recombination library is one with no crossovers, since the parental enzymes have zero E . The fundamental problem is then the trade-off between diversity and fraction of folded proteins. However, with SCHEMA E as a predictor of the folding percentage of the chimeras, one can choose crossover points to create a library with the optimum trade-off between these two properties.

C. RASPP Library Design

In order to solve this trade-off between diversity and percentage folded, our lab developed the library design algorithm RASPP (recombination as a shortest-path problem) to find the libraries with optimum properties without evaluating all possible libraries for a given set of constraints⁸. RASPP treats each potential crossover location as a node, and defines each possible library as a path connecting the start of the protein to its end while passing through the designated number of nodes. The “lengths” of these paths are then scored using SCHEMA (or another pairwise energy function) to determine statistics for each library as a whole. The relevant statistics are the average disruption, $\langle E \rangle$, and average number of mutations, $\langle m \rangle$, (defined as the number of residues differing from the most closely related parent) of every chimera in the library.

RASPP can be fine-tuned to take several inputs, with block number and length being the most commonly used. Eight blocks has been the standard of choice, as this eases construction of the library (two four-fragment ligation steps). However, the falling cost of gene synthesis allows for the creation of libraries without construction constraints. Constraints on block size vary depending on the length of the entire protein but can create an even distribution of block lengths. Libraries containing one large block can lead to difficulty in data analysis, and blocks under 40 bp are difficult to purify using standard kit chemistry (another constraint avoided by gene synthesis).

Using the specific inputs, RASPP generates a set of optimized libraries and calculates their statistics. It is important to review statistics on the block length and

mutation distributions, but the most important statistics, average SCHEMA energy and average mutation, should be plotted as shown in Figure 1.1 for a library of CBHI enzymes. RASPP curves are usually of this shape, with a gradual slope followed by a steep increase in $\langle E \rangle$ over a short increase in $\langle m \rangle$. The optimal libraries are the ones that appear just before this sharp turn, shown by the highlighted region in Figure 1.1, as they have a high mutation rate without paying a large penalty in SCHEMA energy.

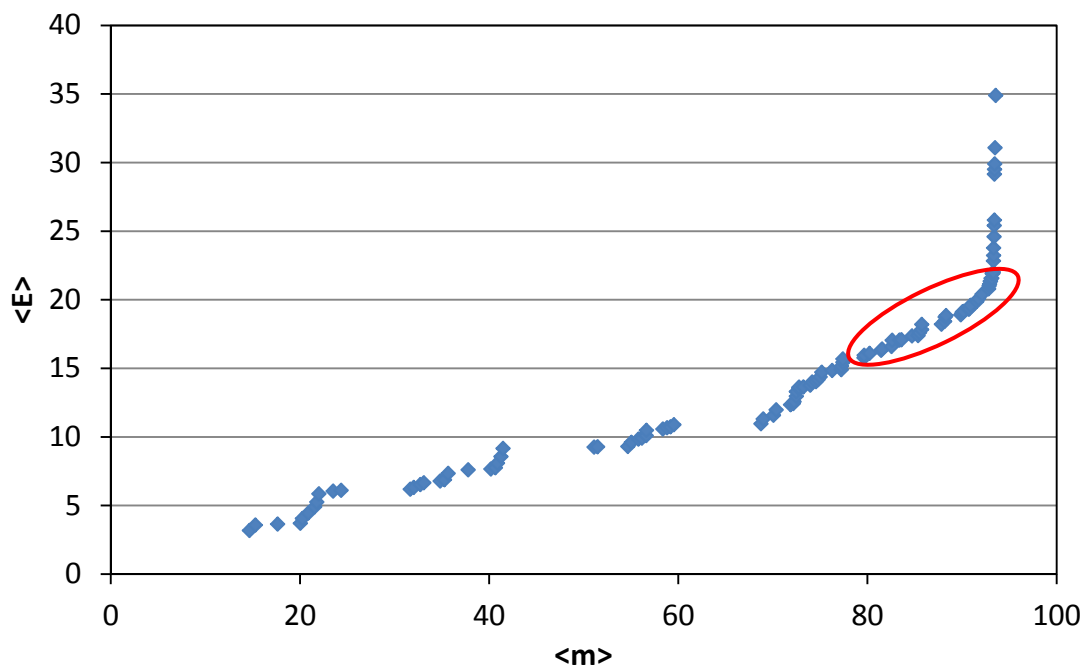


Figure 1.1. RASPP curve for eight block recombination libraries of five fungal CBHI enzymes sharing 61% to 81% sequence identity. The optimal libraries appear in the highlighted region.

D. Additivity

The true strength of SCHEMA lies in its predictive power. The cost in resources and time to build every member of a chimera library can be large (a five parent, eight block library contains 390,625 possible chimeras). Once built, it would take too long to

fully characterize the stability of every chimera. To circumvent this substantial amount of work, one can harness simple linear regression models to accurately fit SCHEMA library data sets. This ability stems from the fact that SCHEMA seeks to group contacts together inside blocks and thus minimizes interblock contacts. Ideally then, each block can be thought of as a completely independent element and have its own unique properties that are minimally affected by the identity of other blocks present. In other words, SCHEMA maximizes the additivity of the blocks for certain properties. SCHEMA's predictive power comes from determining block contributions from a limited number of measurements and using these values to predict the properties of any possible chimera.

The properties best predicted by this type of linear regression are those whose changes due to mutations are themselves additive. In general, stability is considered to be additive unless there is coupling between residues. Coupled residues are usually spatially close to each other, leading them to be considered contacting by SCHEMA and thus grouped together into the same block as much as possible. Stability is also one of the properties recombination can easily enhance, as it is affected by all portions of the protein, not just the active site residues. Stability should therefore be modeled well by SCHEMA libraries and this has proven to be the case on several occasions^{9;10}. There has been some success modeling expression levels in yeast (Arnold lab unpublished data), but this method has not produced accurate models for more complex properties that are thought to be highly nonadditive, such as catalytic activity.

E. Regression Modeling

The purpose of formulating a predictive model is to necessitate less data collection, but it is important to gather enough data to generate an accurate model. The quality of the model will of course increase with more data points, but the minimum number of measurements needed to construct a linear regression model is given by

$$(P - 1)B + 1 \qquad \text{Eq. 1.2}$$

where P is the number of parental enzymes and B is the number of recombination sequence blocks that each enzyme is separated into. This formula stems from the fact that one parent is used as a reference point and the block contributions of the other parents are calculated according to how they affect the properties of the reference parent. The actual linear regression models are easily created using any standard regression software such as MATLAB.

After formulating the model, it should be validated by determining the fit of the calculated versus measured values, before being used for prediction. More rigorous cross-validation methods are also useful here. Cross-validation models involve individually removing each data point, formulating a new model based on the remaining data, and then using that model to predict the removed point. The fit to this kind of plot will give a better look at the accuracy of the model. A bad cross-validated fit does not necessarily mean a bad model and could instead be due to the nature of the data on which the model is based. If there is only one data point for a certain block, a model

excluding that point will have no way to accurately predict it. One should exclude these types of data points from the cross-validated fit, but if there are many such data points this type of analysis loses meaning.

The model calculates the contributions of every block from each parent and therefore identifies the beneficial and neutral blocks. In combination with beneficial blocks, neutral blocks add diversity to the improved chimera set without penalties. It is typical to identify only a handful of beneficial blocks, but when combined with many neutral blocks, this yields a decent (20–30) sized set of improved chimeras, even taking into account that some of these chimeras may not express.

F. Library Construction

The set of chimeras chosen for analysis affects the model quality, with some sets being more information rich than others. It is possible to use optimal experimental design to select a set number of chimeras that give the most information and thus result in the most accurate model. There are several optimal experimental design criteria available, each with different applications¹¹.

The widespread commercial availability of DNA synthesis makes chimeragenesis a simple process. As earlier discussed, only a small sample set of chimeras need be characterized to accurately predict the properties of the full library, making gene synthesis a viable option even for research groups with limited resources.

If more chimeras are needed than can be afforded by gene synthesis, there are detailed methods for producing the whole chimera library using standard cloning

methods¹². This technique relies on Type IIB restriction enzymes inserted into sites between each block and can be used to generate every possible chimera. However, this method requires a significant cost in sequencing to verify proper construction and library statistics. It is important to note that while this method produces all possible chimeras, it gives no simple way to isolate a particular chimera of interest.

Specific chimeras can be synthesized or constructed by hand using splicing by overlap extension (SOE) PCR¹³ or ligating purified block fragments. SOE PCR is straightforward but tedious, as separate fragment and assembly steps are necessary for each junction between blocks from different parents. This can be several steps for the more complex chimeras and each unique junction requires a separate set of primers.

While it is a relatively simple and straightforward procedure to ligate together a chimera from its component blocks, it is much more difficult to individually obtain those blocks. One often imposes block size constraints when designing SCHEMA libraries in order to have a uniform block size distribution, making it nearly impossible to separate blocks based on size. While the cost of DNA synthesis is still too high to synthesize all the members of a full chimera library, it is not prohibitively expensive to individually synthesize blocks in plasmids flanked by the Type IIB restriction sites. Obtaining the individual blocks then consists of digesting the plasmids and purifying the correct pieces. Most DNA synthesis companies, such as DNA2.0 (Menlo Park), price constructs by the base pair but with a minimum price. If the blocks are larger than the minimum, the cost to synthesize the individual blocks will be nearly the same as the full length genes (the

difference comes from the number of restriction sites needed). In addition, companies may be willing to reduce the minimum synthesis cost for a large order.

G. Library Characterization

After construction and verification by sequencing if necessary, the members of a chimera library must be evaluated to identify the improved variants. Since an improved variant depends on whatever property is of interest, a reliable assay is necessary to evaluate this property. As mentioned, recombination has been successfully used to produce chimeras with improved activity¹⁴. In these cases, the authors identified improved variants using colorimetric reactions that are easily seen by the naked eye and can be quantified using a spectrophotometer¹⁵. These screens take advantage of molecules that form colored adducts with products of the desired reaction and are thus a measure of total activity. Total activity data can be combined with a protein concentration assay to calculate specific activity. However, this is usually difficult as most screening takes place in either lysate (bacteria) or secretion culture (yeast), and thus many other proteins and cell components that interfere with specific protein concentration determination are present. Some proteins have properties that can be used to quantify their concentration, such as the Soret band of P450 enzymes¹⁶, but more often total activity is used in place of specific activity. Whatever the screen, the reaction conditions, such as temperature, time, substrate concentration, etc., must be fine-tuned so that improved clones are easily identifiable. This is most easily done by

running under conditions where the parental enzyme performs only moderately well and improvements are readily visible.

Stability is a property that is commonly and effectively increased using recombination^{9; 10}. A straightforward measure of stability is activity at a higher temperature. The only modification needed to a preexisting effective screen for activity is then to run it at higher temperature. Changes in expression can also complicate this method, so the ratio of activity at a high temperature to activity at a lower temperature is often a better metric.

Another type of stability measurement involves the use of residual activity after incubation at an elevated temperature. Using this type of measurement, stability can be reported as a single value, such as a T_{50} , here defined as the temperature at which after a ten-minute incubation half of the protein sample is unfolded. One performs residual activity assays under conditions where the parent enzyme has strong activity without incubation. For the chimeras to be tested, one must adjust the amount of enzyme or expression culture so that the unheated samples give similar signal to unheated parent. This is to eliminate changes in stability due to protein concentration, which can be quite large. The samples are divided into aliquots which are incubated at different temperatures for 10 minutes and then cooled on ice. After all the samples have been incubated, they are reacted with substrate under the previously determined conditions. One normalizes the activity to the unheated samples' values and calculates the T_{50} by a linear fit through the portion of the curve where the protein is unfolding. Temperature

ranges should be adjusted so that the parent protein fully unfolds and the chimeras pass through their T_{50S} .

When dealing with residual activity, it is important to verify that the protein under study unfolds irreversibly, otherwise these types of measurements lose meaning. Proteins with multiple disulfide bonds are especially suspect as these can reform once the temperature is lowered. To prevent this, incubations are run under reducing conditions, such as in the presence of dithiothreitol (DTT). When DTT is present under denaturing conditions, it reduces the now exposed cysteine residues, preventing them from forming disulfide bonds even once the denaturing conditions are removed.

H. Previous Applications of SCHEMA

Our lab has designed SCHEMA libraries for a variety of enzymes, including β -lactamases¹⁷, cytochromes P450¹⁸, as well as bacterial¹⁹ and fungal cellulases^{2;9}. These libraries are diverse sets of folded and active enzymes based on parents that share as little as 30 % to as high as 80 % sequence identity.

An excellent example of SCHEMA's implementation is its application to increase stability of cytochrome P450 enzymes¹⁰. Li et al. took an eight block, three parent SCHEMA library of bacterial P450s and measured the stabilities of 184 random chimeras. They fit a simple linear regression model to these measurements and used it to accurately predict the stabilities of all possible chimeras, including several that were more stable than any of the parents (Figure 1.2).

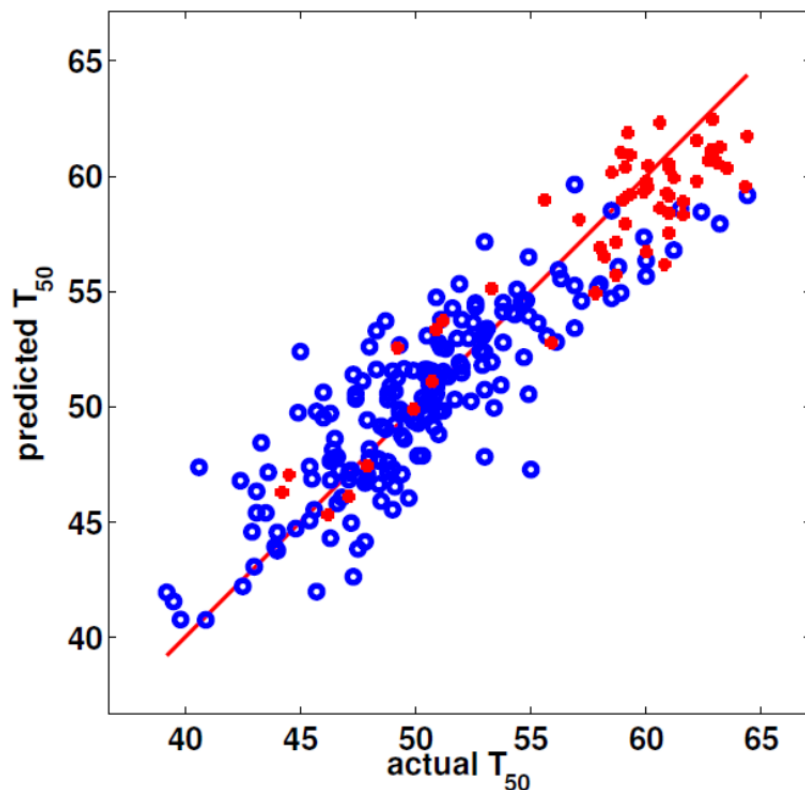


Figure 1.2. A linear model based on the T_{50} s of 184 P450 chimeras (blue) was used to accurately predict a set of thermostable P450 chimeras (red)¹⁰

This project set out to produce a diverse set of P450s with thermostabilities surpassing those of the parental enzymes. SCHEMA design was able to accomplish this and more: further characterization revealed that the chimera library has a wide variation in other properties, including substrate specificity¹⁴. These chimeras are used both in our lab and by the company Codexis (Redwood City, CA, USA), which offers them for applications such as drug lead diversification and metabolite production.

I. Renewable Energy Sources

Recently, our lab has begun work engineering cellulases to help address the national need for renewable fuels. Currently, only 3% of energy for transportation is

renewable²⁰. The majority comes from fossil fuels, of which over two-thirds are imported. This causes our heavy reliance on foreign sources of oil and has made future energy independence a large part of national security. The Energy Independence and Security Act (EISA) of 2007 mandated that the U.S. increase its production of renewable fuels to 36 billion gallons a year by 2022, nearly half of which are expected to come from cellulosic feedstock²¹. Besides decreasing our reliance on foreign fuel sources, renewable fuels from biomass sources can achieve zero net carbon dioxide emissions²². However, much of our current renewable fuels are in the form of corn ethanol. This is considered a transitional technology in that it requires large amounts of farmland and fresh water and is very energy intensive. To meet EISA goals, conversion technology must use cellulose as found in trees and grasses.

However, cellulose found in plant matter is notoriously difficult to harvest and break down due to a number of factors. The overall glucose yield from cellulose is dependent on the source of biomass, pretreatment steps, and cellulase cocktail. In plant walls, cellulose is associated with lignin and hemicellulose, both of which must be removed before the cellulose can be processed. Lignin is a highly cross-linked racemic macromolecule with huge molecular masses (in excess of 10,000 Da)²³. Lignin is highly aromatic and hydrophobic, making it difficult to process chemically or enzymatically. In plants, lignin forms a seal around the cellulose and exhibits limited covalent association with the hemicellulose²⁴. As such, lignin must be removed during pretreatment and is often burned for energy. Like cellulose, hemicellulose is a sugar polymer, but unlike cellulose, hemicellulose is a heteropolymer composed of many different sugar

monomers including xylose, mannose, galactose, rhamnose, and arabinose²⁵. During pretreatment, hemicellulose is solubilized and separated from the cellulose so it can be broken down by enzymes such as xylanases (xylose is the sugar present in the largest amount).

Biomass pretreatment steps fall into two categories: chemical or hydrothermal. The objectives of both categories are to liberate the cellulose from lignin and hemicellulose as well as to increase the accessible surface area, disrupt the cellulose crystallinity, and increase the pore volume²⁶. Chemical pretreatments use dilute acids, bases, or organic solvents at moderate temperature and pressures. These chemical steps can be used simultaneously or additionally to hydrothermal steps at high temperature and pressure for increased effect at the cost of larger energy inputs.

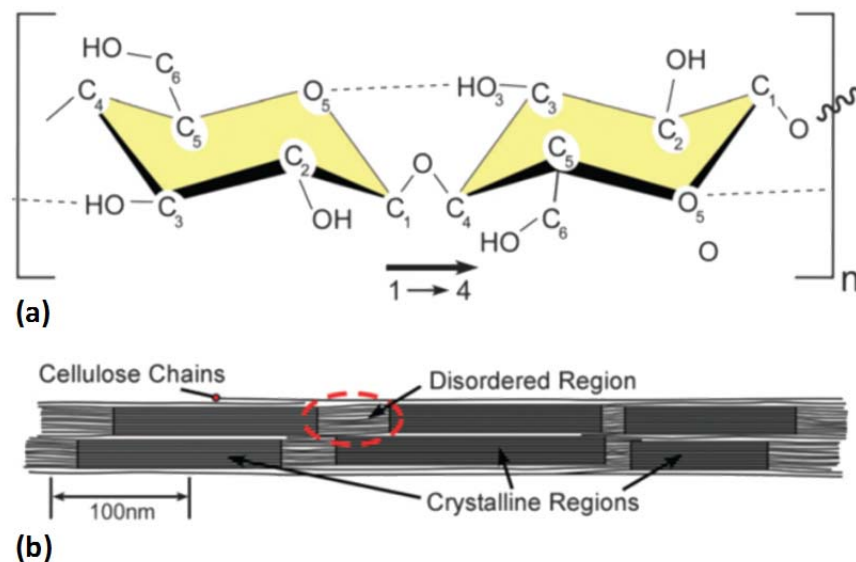


Figure 1.3. (a) The polymer structure of cellulose showing the direction of the β -1,4 bond. (b) Cellulose microfibril containing amorphous and crystalline regions²⁷.

Once the cellulose has been pretreated, it must be broken down into simple sugars that can then be used as the carbon source in fermentation. As seen in Figure 1.3, cellulose is a polymer composed of repeating units of two D-anhydroglucopyranose joined by β -1,4 linkages (anhydrocellobiose) with a degree of polymerization ranging from 100 to 20,000²⁸. The cellulose chains form two distinct regions, disordered amorphous regions and highly ordered crystalline regions. Both regions have a high degree of intrachain hydrogen bonding, whereas the crystalline region is characterized by interchain hydrogen bonding as well. The vast hydrogen bonding network makes it more difficult for enzymes to reach and hydrolyze the crystalline regions, whereas the amorphous region is readily accessible and usually broken down first. The amorphous region also contains a higher percentage of cellulose polymer-free ends, making it more soluble and presenting the exo-acting cellulases with a higher substrate concentration than in the ordered regions.

J. Cellulases

The actual mechanism of cellulose breakdown varies between different organisms, but typically involves at least three classes of cellulase. Figure 1.4 shows a schematic of the typical cellulases produced by fungi. Bacteria produce cellulases with similar function but which are often bound together on scaffold proteins to form cellulosomes. The rest of this work will refer to fungal cellulase systems. The exo-acting cellulases, called cellobiohydrolases (CBH), release smaller sugar chains, mostly the disaccharide cellobiose shown in Figure 1.3(a). However, the CBHs work processively

down from the free ends of the cellulose polymer (CBHI and CBHII work from the reducing and nonreducing ends, respectively) and have little or no activity on the interior of the cellulose chain. This is a consequence of the structure of CBHs, whose active sites form tunnels where cellulose chains enter one side and smaller sugar chains exit the other after hydrolysis²⁹.

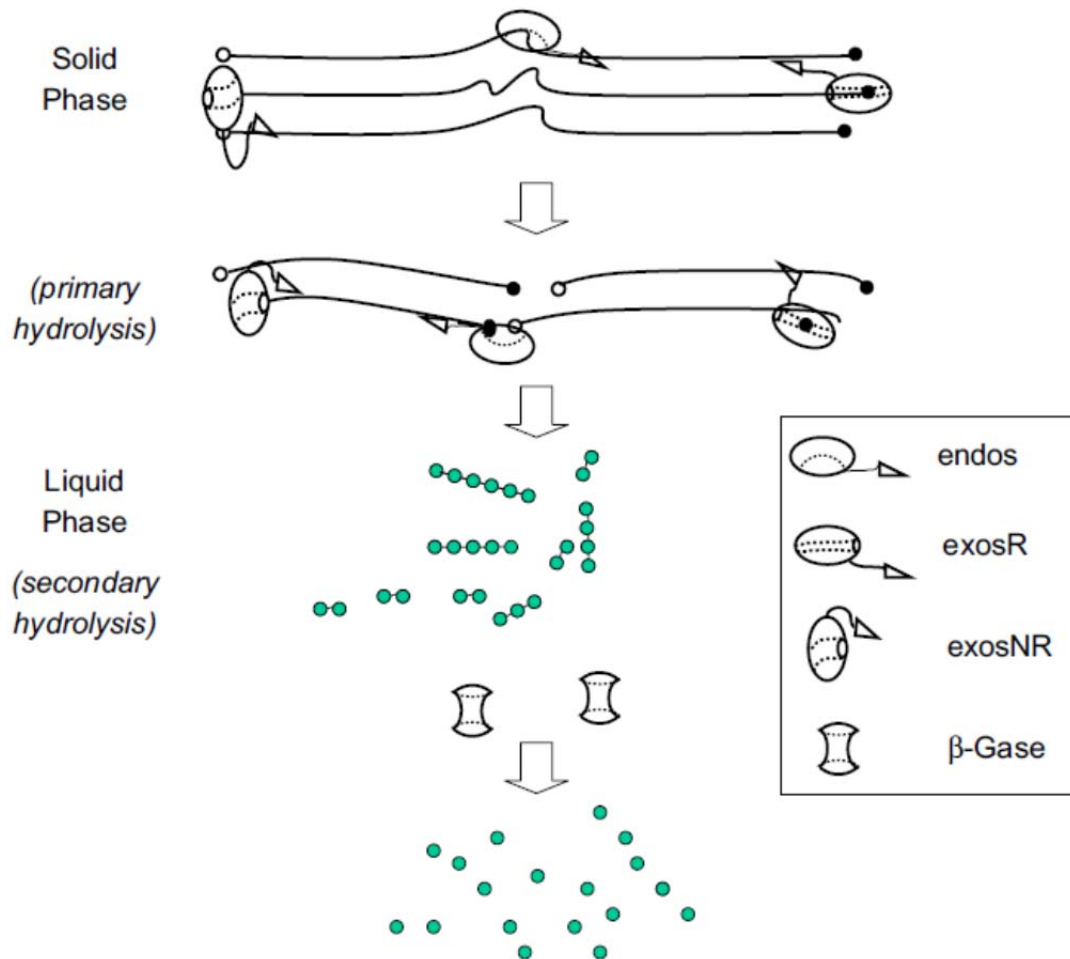


Figure 1.4. Cellulose hydrolysis via different classes of cellulases working synergistically³⁰.

Endoglucanases work to produce free ends from the interior of the cellulose chain by cleaving at random positions in the interior of the chain. Both endoglucanases and CBHs bind to the cellulose through a carbohydrate binding module (CBM) attached to their catalytic domains via a short linker peptide that is highly glycosylated. The CBMs for all three enzyme types are highly similar, with two disulfide bonds and three aromatic side chains (tyrosine or tryptophan) that interact with the six-membered rings of the glucose molecules in the cellulose as well as several other residues that interact with other parts of the cellulose chain (Figure 1.5). The catalytic domains alone have very little affinity for cellulose, causing speculation that the CBM's only purpose is to bring the catalytic domains into contact with the cellulose. Accordingly, hydrolysis rates are greatly increased in the presence of a CBM³¹.

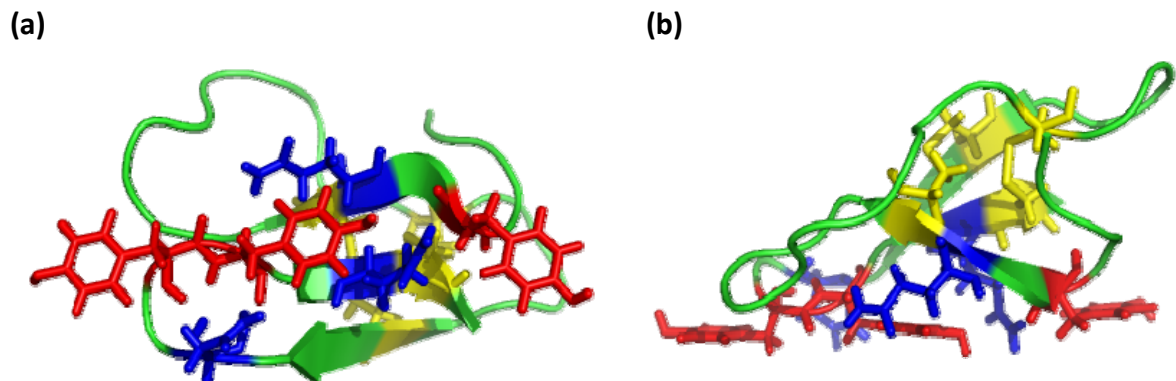


Figure 1.5. (a) Bottom and (b) side view of the CBM of *H. jecorina* CBHI with disulfide bonds shown in yellow, aromatic residues shown in red, and other residues that interact with cellulose in blue.

The activity of the endo- and exo-glucanases causes variation in the properties of the cellulose such as chain end number, accessibility, and topography over time. This

results in rapid changes in hydrolysis rate, of which this primary hydrolysis is the limiting step. The smaller liberated sugar chains are highly soluble and acted on by β -glucosidase enzymes to produce individual glucose units. These glucose molecules, along with any xylose or other sugars released from hemicellulose in separate reactions, are used as the carbon source for fermentation reactions that produce fuel. Fermentation is typically carried out in a separate bioreactor by fermentative microbes (usually yeast) that grow under milder conditions than that of the cellulose hydrolysis reactions. There are schemes where the pretreated biomass is hydrolyzed and fermented in the same reactor. In simultaneous saccharification and fermentation, the cellulase enzymes and fermentative machinery are not produced by the same organism, but they are both present in the same bioreactor³². In consolidated bioprocessing, the same organism breaks down pretreated biomass and ferments sugar to biomass³³. These schemes are limited by the fact that cellulase reaction rates are quite low at the temperatures under which industrial fermentative microbes grow.

K. Cellulase Limitations

Even under optimal operating conditions, the breakdown of pretreated biomass into simple sugars is the economically limiting step in the production of biofuels, and enzyme cost can account for nearly a fourth of the total production cost³⁴. The high cost of enzymes is due to the fact that large amounts of enzyme are needed, which is in turn due to the relatively slow catalytic rates and short operating lifetimes of these cellulase enzymes³⁵. Attempts to increase the hydrolysis rates of cellulases have not been

successful, mostly likely due to the fact that cellulose hydrolysis is a native ability of enzymes that have been evolving toward this function for millions of years. Their cellulose hydrolysis activity is therefore likely near a local maximum of fitness with little room for improvement, at least under condition close to those encountered in nature.

To circumvent this difficulty in increasing the catalytic rate of cellulases, scientists have identified other properties to improve, namely thermostability. Increasing the thermostability of cellulases allows them to operate for longer periods of time and thus hydrolyze more cellulose per enzyme. This reduces enzyme loading and production cost. The rate constants for the cellulose hydrolysis step increase with temperature according to the Arrhenius equation

$$k(T) = k_0 e^{-E/RT} \quad \text{Eq. 1.3}$$

where $k(T)$ is the rate constant at temperature T , k_0 is the rate constant at a reference temperature, and E is the activation energy of the reaction. The activation energy for cellulase reactions has been estimated to be 5540 cal/mol³⁶, which results in a 1.9-fold increase in specific activity at 70 °C compared to 45 °C. Increasing the operating temperature will increase the rate of chemical reaction, but total sugar production is dependent on several other factors as well. In any reaction process, substrate or catalyst diffusion can be limiting, causing the reaction to proceed slower than the predicted rates. To determine whether reaction or diffusion is limiting, one calculates the Thiele

modulus, a dimensionless number composed of the square root of the characteristic reaction rate divided by the characteristic diffusion rate

$$\phi = \sqrt{\frac{kC}{DC}} = L\sqrt{\frac{k}{D}} \quad \text{Eq. 1. 4}$$

where D is the diffusion coefficient of the mobile species in solvent and L is the characteristic length of the system (here taken as the approximate length of the enzyme substrate channel: 100 Å). The Thiele modulus indicates which process is rate limiting³⁷. A high Thiele modulus (>0.5) means that the reaction proceeds at a rate where diffusion cannot supply substrate fast enough, causing the local concentration of substrate around the catalyst to drop. A low Thiele modulus then means that diffusion is sufficient to maintain near-bulk substrate concentration around the catalyst. For most aqueous reactions with a soluble substrate, the Thiele modulus is small. For example, the Thiele modulus for a CBH class I (CBHI) on soluble sugars is 0.002–0.003, depending on the temperature (using kinetic parameters from Kadam et al.³⁶ and diffusion coefficients for hexoses from Ribeiro et al.³⁸).

However, in the case of CBHs reacting on cellulose, the “diffusion” term must include contributions from diffusion of the cellulase in aqueous solvent, binding of the cellulase to solid cellulose, and diffusion of the bound cellulase along the surface of the cellulose chains. Diffusion coefficients of enzymes in aqueous solution vary from

1×10^{-10} to 1×10^{-11} m²/s depending on the size of the protein. This is one to two orders of magnitude lower than that of small molecules, such as sugars, in aqueous solution. Binding of cellulases to cellulose is complicated by the heterogeneity of the substrate and the presence of two domains. It is best described by a CBM adsorption step followed by a catalytic domain/substrate complexation step³⁹. There have been many models over the years seeking to describe cellulase adsorption, most using variations of the Langmuir isotherm. While the Langmuir isotherm often fits experimental data, it is based on assumptions that are not representative of the cellulase system, most notably that the substrate is a flat plane with equivalent sites⁴⁰. As such, there are inconsistencies between experimental data and predictions⁴¹. The adsorption and complexation steps are assumed to be fast in comparison with the other steps, allowing them to be considered at equilibrium⁴⁰. The limiting component of the diffusion term is the diffusion of the enzyme along the surface of the cellulose, with measured diffusion coefficients of $2\text{--}4 \times 10^{-15}$ m²/s at 25 °C⁴², several orders of magnitude smaller than that of the free enzyme. Using this value, the Thiele modulus is 1.0, suggesting that diffusion is limiting. In general diffusion is more weakly activated than reaction³⁷, and so as temperature increases, the overall reaction will become even more diffusion limited. Additionally, as the temperature rises, the equilibrium of bound vs. free enzyme will shift to a lower concentration of bound cellulase. This will further hinder the rate of sugar production.

Increasing the reaction temperature will still result in higher specific activity even with a diffusion limited reaction, as diffusion coefficients follow Arrhenius behavior as

well. Increasing cellulase thermostability will thus allow for a higher reaction temperature and higher specific activity in addition to longer enzyme lifetimes.

L. Cellobiohydrolase I

CBHI is the enzyme shown in Figure 1.4 that acts to liberate cellobiose processively from the reducing end of the cellulose chain. It has a tunnel-shaped active site through which the cellulose polymer enters and cellobiose exits. The active site is composed of three conserved residues with acidic side chains (two glutamic acids and one aspartic acid, shown in yellow in Figure 1.6) inside the substrate channel. CBHI is of interest because it is the principal component of industrial cellulase mixtures and accounts for ~60 wt% of the cellulases secreted by the prevalent commercial cellulase production host, the filamentous fungus *Hypocrea jecorina* (*Trichoderma reesei*)⁴³. As such, CBHIs have been the subject of multiple enzyme engineering efforts aimed primarily at improving CBHI thermostability. Both rational disulfide bond engineering⁴⁴ and high throughput screening of CBHI random mutant libraries⁴⁵ have been employed to create stable CBHI variants. Disulfide bond engineering is limited to CBHIs for which a crystal structure is available. While there are crystal structures of CBHIs from several different fungi in the protein database, there are few available from thermophilic fungi. High throughput screening is limited to CBHIs that are expressed by a suitable heterologous host at sufficient levels.

Difficulty in engineering fungal CBHIs has stemmed from the fact that they are notoriously difficult to express in a heterologous host. In *E. coli*, the enzyme either

misfolds or aggregates into inclusion bodies from which active enzyme cannot be extracted⁴⁶. In yeast, problems arise from post-translation events that occur during secretion, such as glycosylation. Glycosylation patterns and amounts vary widely between organisms and even strains⁴⁷, and their effect on expression is poorly understood. CBHI enzymes have 2–5 N-glycosylation sites (the sequence of amino acids: Asn-Xaa-Ser/Thr where Xaa is not Pro)⁴⁸ on the catalytic domain, as shown in blue in Figure 1.6. In addition, most CBHIs have a CBM attached via a linker domain that is highly O-glycosylated (the linker and CBM domains are not shown in Figure 1.6 due to the high flexibility of the linker interfering with crystal packing and thus structure determination). Early studies in expressing CBHIs in yeast either produced misfolded or inactive enzyme^{49; 50}. Later studies produced enzyme with greatly lowered specific activity due to hyperglycosylation⁵¹ and finally, active enzyme at low amount (less than 50 mg/L) with careful choice of yeast strains⁵².

Another factor possibly affecting expression is the existence of 8–10 disulfide bonds in the catalytic domain and another two in the CBM. Disulfide bonds increase the stability of the protein but form after translation. As such, the ability of cysteine residues to pair is dependent on the cellular environment and host secretion machinery, which vary greatly between yeast and filamentous fungi. Unpaired cysteine residues can negatively affect the protein's stability by interfering with the disulfide bonds⁵³.

recombination. As the property of interest is stability, recombination is ideal for producing diverse chimeras that maintain cellulose hydrolysis activity but have varied stabilities. Using SCHEMA recombination, we can assay a limited number of chimeras and use linear regression to predict the properties of the rest without large culture volume being a problem. Furthermore, the conservative nature of the mutations from recombination coupled with a library designed to minimize disruption will maximize the fraction of folded proteins and give the best chance of avoiding expression problems.

M. CBHI Wild-Type Enzymes

The following work is published in *Protein Engineering Design & Selection* (2010) **23**, p. 871-880 by Heinzelman, P., Komor, R., Kanaan, A., Romero, P., Yu, X. L. , Mohler, S., Snow, C. and Arnold, F.⁵⁴.

The first step in creation of a recombination library is selection of parental genes. As the goal is to produce CBHIs with high stability, we sought out CBHIs from thermophilic fungi so as to have a high starting stability. CBHIs from *Acremonium thermophilum*, *Thermoascus aurantiacus*, *Chaetomium thermophilum*, and *Talaromyces emersonii* were thus chosen. The first three enzymes had the additional benefit of being readily expressed in *H. jecorina* industrial production strains⁵⁵ and the enzyme from *T. emersonii* had a high reported thermostability⁵⁶. We selected the CBHI from *H. jecorina* as the final parent as it is the most industrially relevant enzyme. The sequence identity of the parental catalytic domains ranges from 61% to 81% as shown in Table 1.1.

Table 1.1. ClustalW pairwise sequence alignment of the CBHI parental enzymes for the SCHEMA library⁵⁷.

SeqA	Name	Length	SeqB	Name	Length	Score
1	<i>C. thermophilum</i>	434	2	<i>T. aurantiacus</i>	440	69
1	<i>C. thermophilum</i>	434	3	<i>H. jecorina</i>	441	61
1	<i>C. thermophilum</i>	434	4	<i>A. thermophilum</i>	433	71
1	<i>C. thermophilum</i>	434	5	<i>T. emersonii</i>	437	64
2	<i>T. aurantiacus</i>	440	3	<i>H. jecorina</i>	441	65
2	<i>T. aurantiacus</i>	440	4	<i>A. thermophilum</i>	433	74
2	<i>T. aurantiacus</i>	440	5	<i>T. emersonii</i>	437	81
3	<i>H. jecorina</i>	441	4	<i>A. thermophilum</i>	433	64
3	<i>H. jecorina</i>	441	5	<i>T. emersonii</i>	437	66
4	<i>A. thermophilum</i>	433	5	<i>T. emersonii</i>	437	71

The native sequence of the CBHIs from *T. emersonii* and *T. aurantiacus* contain only nine disulfide bonds while those of the other three parents contain ten. To prevent the generation of unpaired cysteine residues upon recombination, we mutated residues G4 and A72 to cysteines in both the *T. emersonii* and *T. aurantiacus* CBHIs so that each parent CBHI catalytic domain contained ten disulfide bonds. The *T. emersonii* and *T. aurantiacus* natural CBHIs also do not contain CBMs, so the CBM from the *H. jecorina* CBHI was appended with its C-terminal linker to the two catalytic domains, mimicking a construction previously used for heterologous expression of the *T. aurantiacus* CBHI⁵⁵. The *A. thermophilum*, *C. thermophilum*, and *H. jecorina* CBHIs featured their respective wild-type linkers and CBMs. As a result the chimeras have the linker and CBM domain corresponding to their final block's parental identity. A multiple sequence alignment of

the five parental genes as modified is provided in Figure 1.12 in Supplemental Information.

The parental constructs were expressed in yeast and characterized; their total secreted activity levels and thermostability values are listed in Table 1.2. Of the five parental enzymes, three expressed at significant amounts, as measured by total yeast secreted activity on 4-methylumbelliferyl lactopyranoside (MUL) (section D of Materials and Methods). The *T. emersonii* CBHI had a much higher level of total secreted activity than the other parents. The *T. aurantiacus* CBHI had a total secreted activity of just above the 1.6×10^5 mol MUL/l/s cutoff required for accurate thermostability measurements. The *H. jecorina* and *A. thermophilum* CBHIs' total secreted activity was below this threshold and so they were classified as not secreted (NS).

Thermostability measurements in the form of 10 min T_{50} values were determined as described in section E of Materials and Methods. Besides being the parent with the highest total secreted activity, the *T. emersonii* CBHI was also the most stable, followed by the *T. aurantiacus* and then *C. thermophilum* CBHIs. Thermostability values of the *H. jecorina* and *A. thermophilum* CBHIs could not be determined due to their low total secreted activity levels. For the *T. emersonii* CBHI, a measured half-life (section F of Materials and Methods) of <3 min at 70 °C and a melting temperature determined by circular dichroism (section G of Materials and Methods) of ~65 °C are consistent with the measured T_{50} value. As such, we used T_{50} values as our measure of stability.

Table 1.2. Stability and total secreted activity values of the five modified parental CBHIs.

Numbering	CBHI Source Organism	T ₅₀ (°C)	Total Yeast Secreted Activity (mol MUL/L/s × 10 ⁻⁹)
P1	<i>C. thermophilum</i>	59.9 ± 0.5	7.5
P2	<i>T. aurantiacus</i>	62.2 ± 0.4	1.9
P3	<i>H. jecorina</i>	Not Secreted (NS)	0.6
P4	<i>A. thermophilum</i>	Not Secreted (NS)	1.1
P5	<i>T. emersonii</i>	62.9 ± 0.3	23.0

N. CBHI SCHEMA Library Design

SCHEMA and RASPP calculations were done using the parental sequence alignment shown in Figure 1.12 in Supplemental Information and a crystal structure of the CBHI from *T. emersonii* (PDB ID 1Q9H). RASPP calculations generated the curve shown in Figure 1.1 and calculations using the crystal structure of the CBHI from *H. jecorina* (PDB ID 1CEL) gave similar results. A library with block boundaries after residues 32, 76, 107, 155, 201, 248, and 367 (residue numbering is from the *T. emersonii* CBHI) was selected as shown in Figure 1.7. The library is composed of eight blocks from each of five parents for a total of $5^8 = 390,625$ possible chimeras. The average SCHEMA disruption of the chimeras in the library is 20.3 and the average number of mutations is 66, providing a desirable balance between a large number of mutations and a low number of broken contacts.

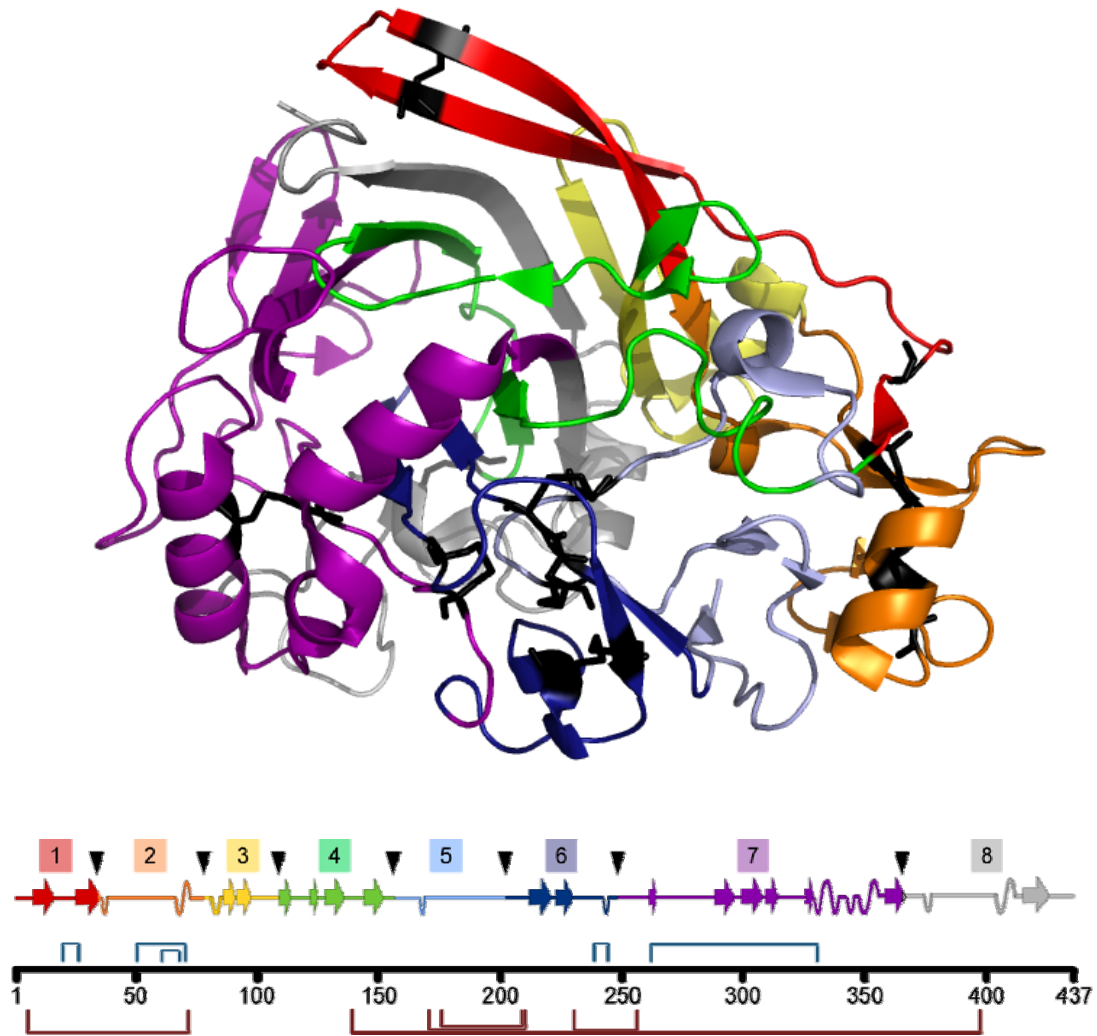


Figure 1.7. CBHI catalytic domain structure and recombination block divisions with secondary structure diagram. Disulfide bonds are denoted by black sticks in the three-dimensional structure. On the block diagram, interblock disulfide bonds are denoted by maroon lines, intrablock disulfide bonds by light blue lines and block divisions by black arrows. Residue numbering is from the *T. emersonii* CBHI.

O. Monomeras

Fungal CBHIs are poorly expressed in *S. cerevisiae*, as evidenced by the total secreted activity results of the parental enzymes. As such, we did not start our screening by constructing a large library of chimeras composed of combinations from multiple

parents at multiple blocks. Instead, we opted to move blocks from four of the parents one at a time into the background of the most highly secreted CBHI, that from *T. emersonii*. This construction strategy creates 32 “monomeras,” chimeras containing seven blocks from the *T. emersonii* CBHI and one block from another of the four parents. Relative to the *T. emersonii* CBHI background, the other four parents contain a total of 346 mutations. The monomera sample set has an average disruption of 5.9 and average mutation level of 15.6. These are considerably lower than the average values of the entire chimera family, and the monomeras are therefore expected to have a high likelihood of retaining fold and cellulase function.

All 32 monomeras were constructed as described in section A of Materials and Methods, and 28 (88%) of them were secreted in functional form from *S. cerevisiae*. For each of the monomeras and parents, the amount of secreted protein relative to the *T. emersonii* CBHI is shown in Figure 1.8. The four monomeras that did not express all contained substitutions at block 7, the largest block. Substitutions at block 4 also resulted in monomeras with significantly decreased total secreted activity, but at high enough levels for thermostability measurements. On the other hand, nine of the monomeras (three with substitutions at block 2, two at block 3, three at block 5, and one at block 8) had higher total secreted activity than that of the *T. emersonii* CBHI parent.

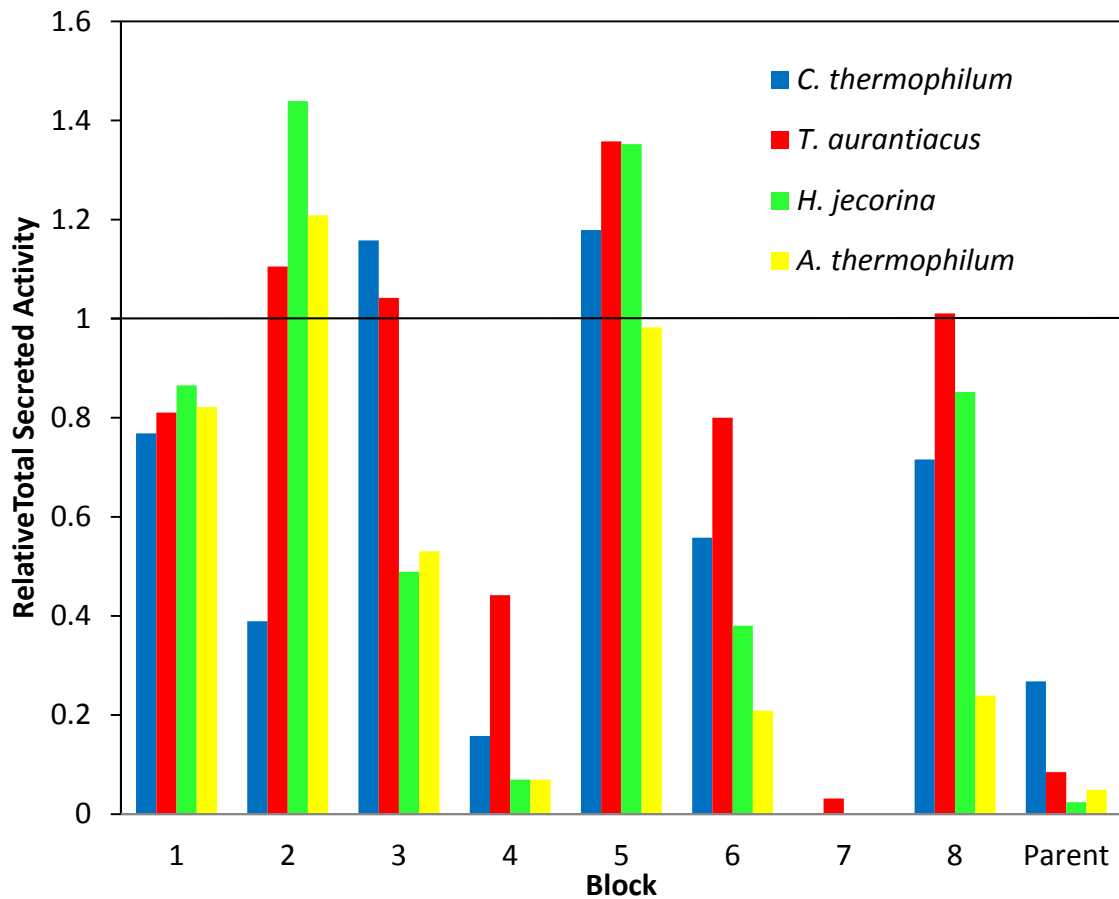


Figure 1.8. Relative total yeast secretion activity, of the 32 monomeras and four parental CBHIs compared to the *T. emersonii* CBHI. Monomeras contain single-block substitutions from the four other parents into the *T. emersonii* CBHI. Error in these measurements was under 10%.

We determined T_{50} values of the 28 monomeras with high enough total secreted activity as described in section E of Materials and Methods, and their effects on stability relative to the stability of the *T. emersonii* CBHI are shown in Figure 1.9. Four of the blocks (two from the *T. aurantiacus* CBHI and two from the *C. thermophilum* CBHI) have a significant stabilizing effect, resulting in an increase in T_{50} of ~ 0.7 to ~ 1.6 °C. Additionally, nine other blocks have small effects on stability, meaning they can be used to increase chimera sequence diversity without having a large negative impact on

stability. Overall, a total of 18 blocks were identified as useable for constructing diverse and thermostable CBHI chimeras.

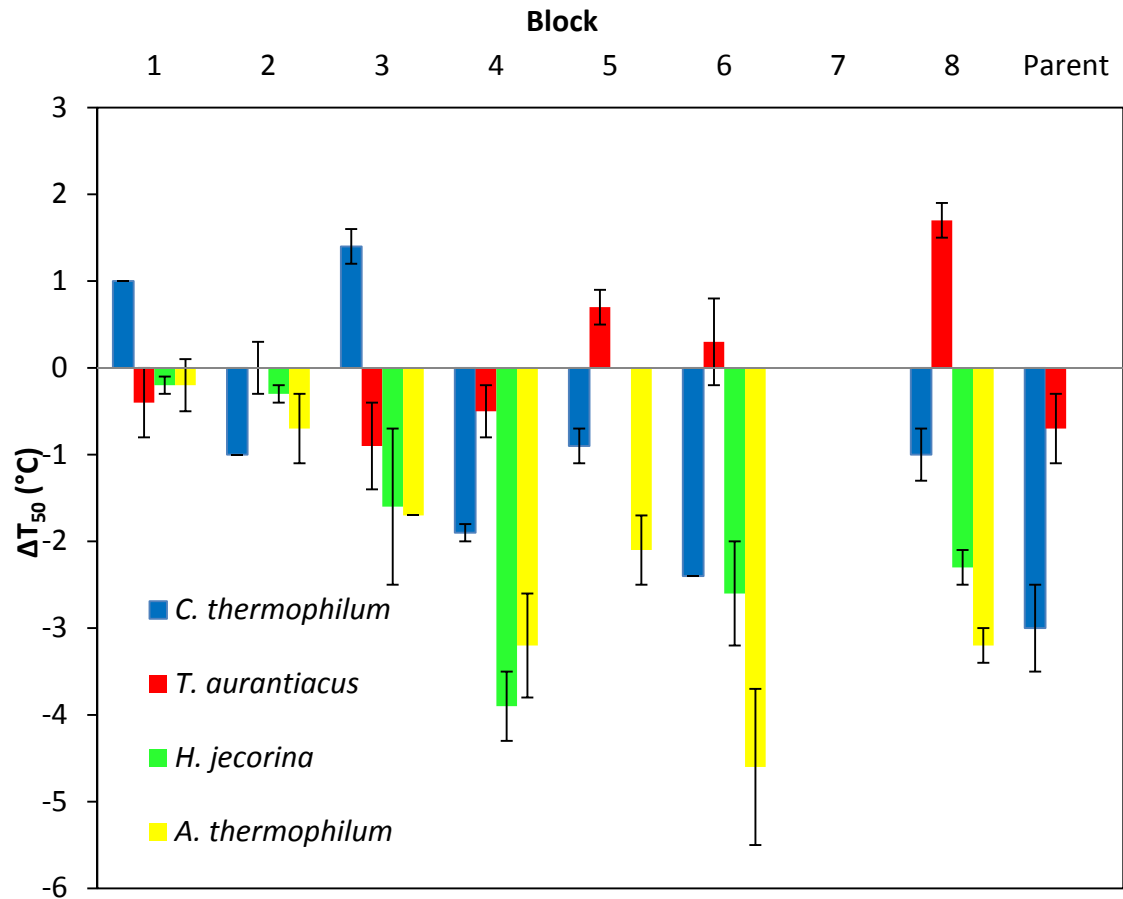


Figure 1.9. The effect on T_{50} of single block substitutions into the *T. emersonii* CBHI background (monomeras).

P. Disulfide Bonds

One of the concerns when designing the SCHEMA library was the presence of ten disulfide bonds, five of which cross over block boundaries as seen in Figure 1.7. All of the parents have cysteine residues present at these positions so SCHEMA treats all of the contacts involving them as being conserved upon recombination. However, it is

unclear whether recombination preserves the appropriate position and orientation of the cysteine residues for disulfide formation. If not, the presence of unpaired cysteine residues could dramatically affect the protein's stability and expression. The data from the monomera screening showed that substitutions at block 4 and block 7 resulted in detrimental effects on secretion and stability. Both block 4 and block 7 contain cysteine residues that are part of interblock disulfide bonds (Cys135 of block 4 forms a disulfide bond with Cys401 of block 8, and Cys253 of block 7 forms a disulfide bond with Cys227 of block 6).

To test whether these interblock disulfide bonds were precluding high total secreted activity of chimeras with substitutions at blocks 4 or 7, we made chimeras with the 4-8 and 6-7 block pairs to conserve the parental source of the disulfides. Their total secreted activity and stabilities are reported in Table 1.3. Chimeras with simultaneous substitutions at block 6 and 7 from the same parent did not express, suggesting that it is not the presence of cysteine residues from different parents that prohibits expression of chimeras with block 7 substitutions. Further supporting this hypothesis, chimeras with simultaneous substitutions at blocks 4 and 8 resulted in chimeras with total secreted activity that fall between those of the monomeras containing the respective single-block substitutions. The T_{50} values are also close to what would be expected using the monomera data and assuming additivity.

This was the first SCHEMA library based on parents that contain a large number of disulfide bonds. Our analysis shows that SCHEMA is robust enough to generate a

library with a large fraction of active members even when the protein is cross-linked by disulfide bonds.

Table 1.3. Total secreted activity and stability measurements for disulfide-paired CBHI chimeras and underlying monomeras. See Table 1.2 for parent numbering system.

CBHI Sequence	Total Yeast Secreted Activity (mol MUL/L/s x 10 ⁻⁹)	T ₅₀ (°C)
55515555	2.6	61.0 ± 0.1
55555551	11.8	61.9 ± 0.2
55515551	6.3	58.2 ± 0.3
55525555	6.7	62.4 ± 0.2
55555552	21.8	64.6 ± 0.2
55525552	11.0	63.1 ± 0.1
55555155	8.6	60.5 ± 0.0
55555515	0.3	NS
55555115	0.1	NS
55555255	14.1	63.2 ± 0.5
55555525	1.0	NS
55555225	0.2	NS

Q. Block 7 Subblocks

The monomera dataset gives block property contributions for all the blocks save for block 7, as substitutions into the *T. emersonii* CBHI background at this position abrogated expression. This leaves a significant gap in our predictive model, as block 7 is the largest block, containing 116 residues and 119 of the 346 total unique mutations in the 32 monomera sample set. In order to garner more information on the stability effects of block 7, we subdivided it into six sub-blocks. We selected block boundaries (after residues 287, 303, 327, 339, and 352) from the unused libraries identified by RASPP that equally distributed the residues of interest in block 7.

We constructed monomeras of the sub-blocks into the *T. emersonii* CBHI as was done for the full blocks. The effects on total secreted activity are shown in Figure 1.10 and give surprising results. Even though substitution of the entire block 7 abolishes expression, a number of sub-blocks actually increase total secreted activity (sub-blocks D and E from the *C. thermophilum* CBHI, sub-blocks C and D from the *T. aurantiacus* CBHI, sub-block C from the *A. thermophilum* CBHI). Substitutions at sub-block A, C, and F are generally not tolerated and result in significantly reduced total secreted activity. Substitutions from the *T. aurantiacus* and *A. thermophilum* CBHI parents resulted in a larger number of active sub-block monomeras as well as sub-block monomeras with higher total secreted activity. This is most likely due to the fact that they have a much higher sequence identity to *T. emersonii* CBHI at block 7 than the *C. thermophilum* and *H. jecorina* CBHI (74% and 70% vs. 57% and 53%, respectively).

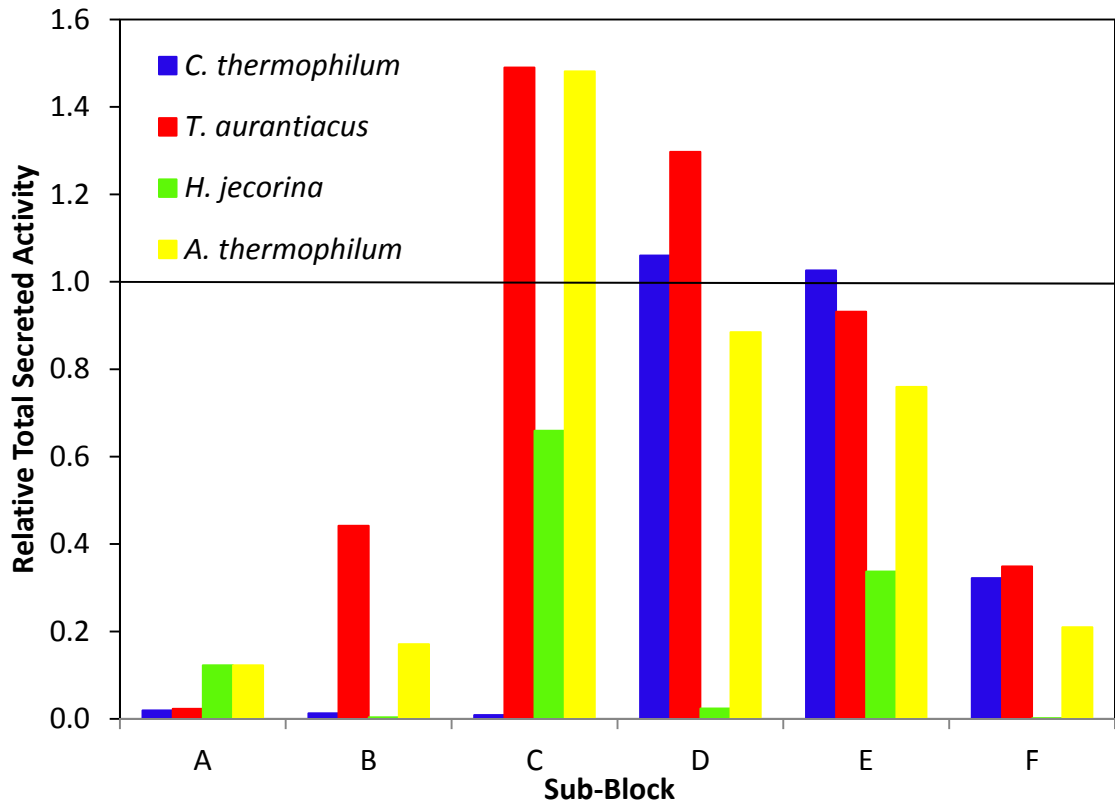


Figure 1.10. Relative total secreted activity, of the 24 sub-block monomeras compared to the *T. emersonii* CBHI. Monomeras contain single sub-block substitutions from the four other parents into the *T. emersonii* CBHI. Error in these measurements was under 10%.

The sub-blocks' effects on stability are shown in Figure 1.11. Two of the sub-blocks (sub-block C from both the *T. aurantiacus* and *A. thermophilum* CBHI) are stabilizing and are used in combination with the stabilizing and neutral blocks identified previously to make up the set of usable blocks. Readily apparent from Figure 1.11 is the fact that several of the sub-blocks have a large destabilizing effect. Stability and total secreted activity are linked, and these highly destabilizing blocks could account for the abrogation of expression caused by the substitution of the entire block 7. Also of note is that the *H. jecorina* and *A. thermophilum* CBHIs both have multiple highly destabilizing

sub-blocks, which could partially account for the inability to express either of these parents in *S. cerevisiae*.

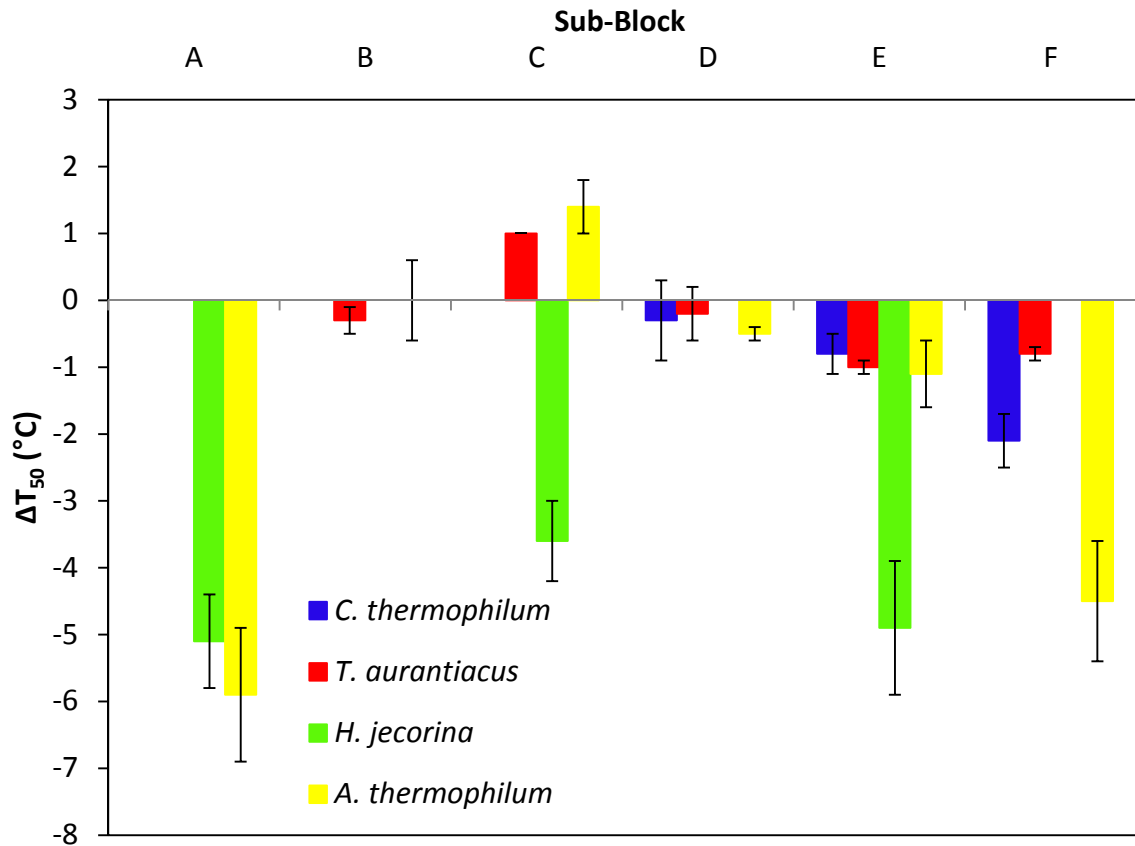


Figure 1.11. The effect on T_{50} of single sub-block substitutions into the *T. emersonii* CBHI background.

R. A Diverse Set of Thermostable Chimeras

The monomera screening identified 18 stabilizing or neutral blocks, which were then selected to build a set of diverse and thermostable chimeras. Due to negative effects of substitution on total secreted activity, all of the chimeras contain *T. emersonii* CBHI sequences at blocks 4 and 7 (see Table 1.2 for parent numbering system). Block 3 from the *C. thermophilum* CBHI and block 8 from the *T. aurantiacus* CBHI both have

large positive effects on stability and so are present in every chimera. The other four block positions (blocks 1, 2, 5, and 6) contain combinations of stabilizing and neutral blocks. All five of the parental CBHIs have blocks represented in the thermostable set.

Table 1.4 shows the stability and total secreted activity values of the thermostable chimera set. All but one of the chimeras have higher total secreted activity than the parent with the second highest total secreted activity, and many have higher total secreted activity than the parent with the highest total secreted activity. All of the chimeras are more stable than the most stable CBHI parent, that from *T. emersonii*, with increases in T_{50} ranging from 1.1 to 3.4 °C.

Using block T_{50} contributions determined from the monomera set (calculated for each block by subtracting the T_{50} of the corresponding monomera from that of the *T. emersonii* CBHI) it is possible to calculate predicted T_{50} values for each of the chimeras, as shown in Table 1.4. When compared to the measured T_{50} values, there is considerable difference for most of the thermostable chimeras (shown in the difference column in Table 1.4). This is most likely due to the fact that the monomeras are very close in sequence to the *T. emersonii* CBHI parent, whereas the chimeras can vary significantly. As such, the monomera-based block property model is heavily biased toward the *T. emersonii* CBHI sequence, and loses accuracy as sequence diverges. The monomera-based predictions consistently overestimate the T_{50} of the chimeras. Overestimates could also stem from increased divergence from the *T. emersonii* CBHI sequence. As the sequence diverges, structural clashes could accumulate, decreasing

stability. Even with the overestimation, the monomera-based model consistently predicts thermostable sequences. As such, the model is best used for predicting whether a block is stabilizing, neutral, or destabilizing and less so for predicting exact contributions to T_{50} values. Even so, it is clear that using a set of 32 measurements we were able to accurately identify a diverse set of 16 thermostable chimeras from the 390,625 possible chimeras.

Table 1.4. Sequences and properties of thermostable chimeras. See Table 1.2 for parent numbering system. The predicted T_{50} values are the sum of block T_{50} contributions calculated from the monomera screening data. The T_{50} s of parents 1–4 could not be calculated without block 7 T_{50} data. The difference values are the difference between predicted and measured T_{50} values.

CBHI Sequence	Total Secreted Activity (mol MUL/L/s x 10^{-9})	Predicted T_{50} (°C)	T_{50} (°C)	Difference (°C)
11111111	7.5	N.A.	59.9 ± 0.5	N.A.
22222222	1.9	N.A.	62.2 ± 0.4	N.A.
33333333	0.6	N.A.	NS	N.A.
44444444	1.1	N.A.	NS	N.A.
55555555	23.0	N.A.	62.9 ± 0.3	N.A.
34152252	22.6	66.1	64.0 ± 0.1	2.1
55153552	33.2	66.0	64.3 ± 0.0	1.7
32153252	10.2	66.1	64.3 ± 0.2	1.8
55155552	21.9	66.0	64.4 ± 0.7	1.6
22153252	12.8	65.9	64.4 ± 0.2	1.5
52152552	34.3	66.7	64.5 ± 0.0	2.2
12153252	6.4	67.3	64.7 ± 0.2	2.6
45153252	25.3	66.1	64.8 ± 0.2	1.3
12153552	10.9	67.0	64.9 ± 0.3	2.1
25152252	22.2	66.6	65.0 ± 0.1	1.6
13152552	34.7	67.4	65.0 ± 0.0	2.4
12152252	10.2	68.0	65.3 ± 0.1	2.7
55153252	20.0	66.3	65.3 ± 0.2	1.0
55552252	28.5	65.6	65.6 ± 0.7	0.0
55152552	29.4	66.7	65.7 ± 0.1	1.0
55152252	19.6	67.0	66.3 ± 1.0	0.7

We next incorporated the stabilizing sub-blocks (sub-blocks C from the *T. aurantiacus* and *A. thermophilum* CBHIs) into a number of thermostable chimeras. Sub-blocks increased the T_{50} of each chimera by at least 1 °C and in many cases more than 2 °C. In general, sub-block C from the *A. thermophilum* CBHI resulted in a slightly greater increase in T_{50} than that from the *T. aurantiacus* CBHI. The most stable chimera has a T_{50} of 67.5 °C compared to 62.9 °C of the most stable parent, an increase of 4.6 °C.

Table 1.5. Sequences and properties of thermostable chimeras containing sub-blocks. See Table 1.2 for parent numbering system. Block 5* contains sub-blocks A, B, D, E, and F from the *T. emersonii* CBHI and sub-block C from the listed source.

CBHI Sequence	Sub-block Source	Total Secreted Activity (mol MUL/L/s x 10 ⁻⁹)	T_{50} (°C)
5515355*2	<i>T. aurantiacus</i>	42.2	65.7 ± 0.2
1215325*2	<i>T. aurantiacus</i>	10.6	66.0 ± 0.0
5515225*2	<i>T. aurantiacus</i>	31.1	66.6 ± 0.2
1215225*2	<i>T. aurantiacus</i>	10.8	66.6 ± 0.7
5215255*2	<i>A. thermophilum</i>	37.5	66.7 ± 0.0
2515225*2	<i>T. aurantiacus</i>	28.7	66.8 ± 0.1
5515255*2	<i>A. thermophilum</i>	45.7	67.0 ± 0.1
5515225*2	<i>A. thermophilum</i>	27.2	67.0 ± 0.2
5555225*2	<i>A. thermophilum</i>	33.2	67.2 ± 0.2
1215225*2	<i>A. thermophilum</i>	15.2	67.4 ± 0.4
1515255*2	<i>A. thermophilum</i>	20.7	67.5 ± 0.0

S. Conclusions

Using a structure-guided SCHEMA recombination library based on five CBHIs from thermophilic fungi, we characterized 28 of the 32 possible monomers and used the resulting block stability contribution values to produce a set of thermostable chimeras, the most stable of which had a T_{50} 3.2 °C greater than that of the best parent. The measured stabilities of the thermostable chimeras were all lower than predicted

based on the monomera stability data. SCHEMA libraries of other enzymes did not show this lack of additivity, but their models were based on a set of random chimeras with an even distribution of blocks from each parent, and as such, were not biased toward any one parent. Although we ruled out disulfide bonds between cysteine residues from different parents as interfering with expression and stability, substitution at two blocks that contain interblock disulfides resulted in the loss of expression. Further examination of one of the blocks that could not be substituted identified sub-blocks from two parents that further increased the T_{50} by more than 1 °C. The identification of smaller, stabilizing sequence elements within larger, destabilizing ones suggests that there are more gains in stability that could be gleaned from further inspection of the five parental CBHs. The most stable chimera constructed has a T_{50} of 67.5 °C, 4.6 °C higher than that of the most stable parent. While gains in stability are our immediate goal, we ultimately wish to show that this increase in stability translates into increased sugar production at higher temperatures.

<i>Talaromyces emersonii</i>	PTVCASGTTTCQVLN P YYSQCL	498
<i>Hypocrea jecorina</i>	PTVCASGTTTCQVLN P YYSQCL	497
<i>Acremonium thermophilum</i>	PTVCQSPY T CKYSNDWYSQCL	506
<i>Chaetomium thermophilum</i>	CTNCVAGTTCTQLN P WYSQCL	514
	* * : ** * :*****	

Figure 1.12. ClustalW2 multiple sequence alignment of the CBHI parental enzymes for the SCHEMA library⁵⁷. The consensus symbols have the following meaning: an * (asterisk) indicates positions which have a single, fully conserved residue, a : (colon) indicates conservation between groups of strongly similar properties, and a . (period) indicates conservation between groups of weakly similar properties. The colors have the following meaning: red indicates small and hydrophobic, blue indicates acidic, magenta indicates basic, and green indicates everything else.

U. References

1. Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C. & Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 606-611.
2. Heinzelman, P., Snow, C. D., Wu, I., Nguyen, C., Villalobos, A., Govindarajan, S., Minshull, J. & Arnold, F. H. (2009). A family of thermostable fungal cellulases created by structure-guided recombination. *Proc Natl Acad Sci U S A* **106**, 5610-5.
3. Zheng, W., Griswold, K. E. & Bailey-Kellogg, C. (2010). Protein Fragment Swapping: A Method for Asymmetric, Selective Site-Directed Recombination. *Journal of Computational Biology* **17**, 459-475.
4. Ness, J. E., Kim, S., Gottman, A., Pak, R., Kriebber, A., Borchert, T. V., Govindarajan, S., Mundorff, E. C. & Minshull, J. (2002). Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nature Biotechnology* **20**, 1251-1255.
5. d'Abbadie, M., Hofreiter, M., Vaisman, A., Loakes, D., Gasparutto, D., Cadet, J., Woodgate, R., Paabo, S. & Holliger, P. (2007). Molecular breeding of polymerases for amplification of ancient DNA. *Nature Biotechnology* **25**, 939-943.
6. Meyer, M. M., Silberg, J. J., Voigt, C. A., Endelman, J. B., Mayo, S. L., Wang, Z. G. & Arnold, F. H. (2003). Library analysis of SCHEMA-guided protein recombination. *Protein Science* **12**, 1686-1693.
7. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nature Structural Biology* **9**, 553-558.
8. Endelman, J. B., Silberg, J. J., Wang, Z. G. & Arnold, F. H. (2004). Site-directed protein recombination as a shortest-path problem. *Protein Engineering Design & Selection* **17**, 589-594.
9. Heinzelman, P., Snow, C. D., Smith, M. A., Yu, X., Kannan, A., Boulware, K., Villalobos, A., Govindarajan, S., Minshull, J. & Arnold, F. H. (2009). SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *J Biol Chem* **284**, 26229-33.
10. Li, Y. G., Drummond, D. A., Sawayama, A. M., Snow, C. D., Bloom, J. D. & Arnold, F. H. (2007). A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments (vol 25, pg 1051, 2007). *Nature Biotechnology* **25**, 1488-1488.

11. Atkinson, A. C. & Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford Statistical Science Series, 8, Oxford University Press, Oxford.
12. Farrow, M. F. & Arnold, F. H. (2010). Combinatorial Recombination of Gene Fragments to Construct a Library of Chimeras. In *Current Protocols in Protein Science*.
13. Higuchi, R., Krummel, B. & Saiki, R. K. (1988). A General-Method of Invitro Preparation and Specific Mutagenesis of DNA Fragments - Study of Protein and DNA Interactions. *Nucleic Acids Research* **16**, 7351-7367.
14. Landwehr, M., Carbone, M., Otey, C. R., Li, Y. G. & Arnold, F. H. (2007). Diversification of catalytic function in a synthetic family of chimeric cytochrome P450s. *Chemistry & Biology* **14**, 269-278.
15. Arnold, F. H. & Georgiou, G. (2003). *Directed Enzyme Evolution: Screening and Selection Methods*. Methods in Molecular Biology, 230, Humana Press, Totowa, New Jersey.
16. Omura, T. & Sato, R. (1964). Carbon Monoxide-Binding Pigment of Liver Microsomes .I. Evidence for Its Hemoprotein Nature. *Journal of Biological Chemistry* **239**, 2370-&.
17. Meyer, M. M., Hochrein, L. & Arnold, F. H. (2006). Structure-guided SCHEMA recombination of distantly related beta-lactamases. *Protein Engineering Design & Selection* **19**, 563-570.
18. Otey, C. R., Landwehr, M., Endelman, J. B., Hiraga, K., Bloom, J. D. & Arnold, F. H. (2006). Structure-guided recombination creates an artificial family of cytochromes P450. *Plos Biology* **4**, 789-798.
19. Mingardon, F., Bagert, J. D., Maisonnier, C., Trudeau, D. L. & Arnold, F. H. (2011). Comparison of Family 9 Cellulases from Mesophilic and Thermophilic Bacteria. *Applied and Environmental Microbiology* **77**, 1436-1442.
20. (2010). Annual Energy Review 2009. Administration, U. S. E. I.
21. (2010). USDA Biofuels Strategic Production Report. USDA.
22. Hoffert, M. I., Caldeira, K., Benford, G., Criswell, D. R., Green, C., Herzog, H., Jain, A. K., Keshgi, H. S., Lackner, K. S., Lewis, J. S., Lightfoot, H. D., Manheimer, W., Mankins, J. C., Mael, M. E., Perkins, L. J., Schlesinger, M. E., Volk, T. & Wigley, T. M. L. (2002). Advanced technology paths to global climate stability: Energy for a greenhouse planet. *Science* **298**, 981-987.
23. Horley, S. (2001). Lignin and its Properties: Glossary of Lignin Nomenclature. *Dialogue/Newsletters* **9**.
24. Demain, A. L., Newcomb, M. & Wu, J. H. D. (2005). Cellulase, clostridia, and ethanol. *Microbiology and Molecular Biology Reviews* **69**, 124-+.
25. Wedig, C. L., Jaster, E. H. & Moore, K. J. (1987). Hemicellulose Monosaccharide Composition and Invitro Disappearance of Orchard Grass and Alfalfa Hay. *Journal of Agricultural and Food Chemistry* **35**, 214-218.
26. Weil, J., Westgate, P., Kohlmann, K. & Ladisch, M. R. (1994). Cellulose Pretreatments of Lignocellulosic Substrates. *Enzyme and Microbial Technology* **16**, 1002-1004.
27. Moon, R. J., Martini, A., Nairn, J., Simonsen, J. & Youngblood, J. (2011). Cellulose nanomaterials review: structure, properties and nanocomposites. *Chemical Society Reviews* **40**, 3941-3994.
28. OSullivan, A. C. (1997). Cellulose: the structure slowly unravels. *Cellulose* **4**, 173-207.
29. Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J. K., Teeri, T. T. & Jones, T. A. (1994). The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. *Science* **265**, 524-8.

30. Zhang, Y. H. P., Himmel, M. E. & Mielenz, J. R. (2006). Outlook for cellulase improvement: Screening and selection strategies. *Biotechnology Advances* **24**, 452-481.
31. Kim, T. W., Chokhawala, H. A., Nadler, D., Blanch, H. W. & Clark, D. S. (2010). Binding Modules Alter the Activity of Chimeric Cellulases: Effects of Biomass Pretreatment and Enzyme Source. *Biotechnology and Bioengineering* **107**, 601-611.
32. Yu, E. K. C. & Saddler, J. N. (1985). Biomass Conversion to Butanediol by Simultaneous Saccharification and Fermentation. *Trends in Biotechnology* **3**, 100-104.
33. Himmel, M. E., Xu, Q. & Singh, A. (2009). Perspectives and new directions for the production of bioethanol using consolidated bioprocessing of lignocellulose. *Current Opinion in Biotechnology* **20**, 364-371.
34. Gregg, D. J., Boussaid, A. & Saddler, J. N. (1998). Techno-economic evaluations of a generic wood-to-ethanol process: Effect of increased cellulose yields and enzyme recycle. *Bioresource Technology* **63**, 7-12.
35. Nguyen, Q. A. & Saddler, J. N. (1991). An Integrated Model for the Technical and Economic-Evaluation of an Enzymatic Biomass Conversion Process. *Bioresource Technology* **35**, 275-282.
36. Kadam, K. L., Rydholm, E. C. & McMillan, J. D. (2004). Development and validation of a kinetic model for enzymatic saccharification of lignocellulosic biomass. *Biotechnol Prog* **20**, 698-705.
37. Davis, M. E. & Davis, R. J. (2003). *Fundamentals of Chemical Reaction Engineering*, McGraw-Hill, New York, NY.
38. Ribeiro, A. C. F., Ortona, O., Simoes, S. M. N., Santos, C. I. A. V., Prazeres, P. M. R. A., Valente, A. J. M., Lobo, V. M. M. & Burrows, H. D. (2006). Binary mutual diffusion coefficients of aqueous solutions of sucrose, lactose, glucose, and fructose in the temperature range from (298.15 to 328.15) K. *Journal of Chemical and Engineering Data* **51**, 1836-1840.
39. Igarashi, K., Koivula, A., Wada, M., Kimura, S., Penttila, M. & Samejima, M. (2009). High speed atomic force microscopy visualizes processive movement of *Trichoderma reesei* cellobiohydrolase I on crystalline cellulose. *J Biol Chem* **284**, 36186-90.
40. Levine, S. E., Fox, J. M., Blanch, H. W. & Clark, D. S. (2010). A Mechanistic Model of the Enzymatic Hydrolysis of Cellulose. *Biotechnology and Bioengineering* **107**, 37-51.
41. Medve, J., Stahlberg, J. & Tjerneld, F. (1997). Isotherms for adsorption of cellobiohydrolase I and II from *Trichoderma reesei* on microcrystalline cellulose. *Appl Biochem Biotechnol* **66**, 39-56.
42. Jervis, E. J., Haynes, C. A. & Kilburn, D. G. (1997). Surface diffusion of cellulases and their isolated binding domains on cellulose. *Journal of Biological Chemistry* **272**, 24016-24023.
43. Baker, S. E., Le Crom, S., Schackwitz, W., Pennacchio, L., Magnuson, J. K., Culley, D. E., Collett, J. R., Martin, J., Druzhinina, I. S., Mathis, H., Monot, F., Seiboth, B., Cherry, B., Rey, M., Berka, R., Kubicek, C. P. & Margeot, A. (2009). Tracking the roots of cellulase hyperproduction by the fungus *Trichoderma reesei* using massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 16151-16156.
44. Koivula, A., Voutilainen, S. P., Murray, P. G. & Tuohy, M. G. (2010). Expression of *Talaromyces emersonii* cellobiohydrolase Cel7A in *Saccharomyces cerevisiae* and rational mutagenesis to improve its thermostability and activity. *Protein Engineering Design & Selection* **23**, 69-79.

45. Voutilainen, S. P., Boer, H., Alapuranen, M., Janis, J., Vehmaanpera, J. & Koivula, A. (2009). Improving the thermostability and activity of *Melanocarpus albomyces* cellobiohydrolase Cel7B. *Appl Microbiol Biotechnol* **83**, 261-72.
46. Laymon, R. A., Adney, W. S., Mohagheghi, A., Himmel, M. E. & Thomas, S. R. (1996). Cloning and expression of full-length *Trichoderma reesei* cellobiohydrolase I cDNAs in *Escherichia coli*. *Applied Biochemistry and Biotechnology* **57-8**, 389-397.
47. Adney, W. S., Jeoh, T., Michener, W., Himmel, M. E. & Decker, S. R. (2008). Implications of cellobiohydrolase glycosylation for use in biomass conversion. *Biotechnology for Biofuels* **1**.
48. Eriksson, T., Stals, I., Collen, A., Tjerneld, F., Claeysens, M., Stalbrand, H. & Brumer, H. (2004). Heterogeneity of homologously expressed *Hypocrea jecorina* (*Trichoderma reesei*) Cel7B catalytic module. *European Journal of Biochemistry* **271**, 1266-1276.
49. Reinikainen, T., Ruohonen, L., Nevanen, T., Laaksonen, L., Kraulis, P., Jones, T. A., Knowles, J. K. C. & Teeri, T. T. (1992). Investigation of the Function of Mutated Cellulose-Binding Domains of *Trichoderma-Reesei* Cellobiohydrolase-I. *Proteins-Structure Function and Genetics* **14**, 475-482.
50. Penttila, M. E., Andre, L., Lehtovaara, P., Bailey, M., Teeri, T. T. & Knowles, J. K. C. (1988). Efficient Secretion of 2 Fungal Cellobiohydrolases by *Saccharomyces-Cerevisiae*. *Gene* **63**, 103-112.
51. Godbole, S., Decker, S. R., Nieves, R. A., Adney, W. S., Vinzant, T. B., Baker, J. O., Thomas, S. R. & Himmel, M. E. (1999). Cloning and expression of *Trichoderma reesei* cellobiohydrolase I in *Pichia pastoris*. *Biotechnology Progress* **15**, 828-833.
52. Koivula, A., Voutilainen, S. P., Boer, H., Linder, M. B., Puranen, T., Rouvinen, J. & Vehmaanpera, J. (2007). Heterologous expression of *Melanocarpus albomyces* cellobiohydrolase Cel7B, and random mutagenesis to improve its thermostability. *Enzyme and Microbial Technology* **41**, 234-243.
53. Perry, L. J. & Wetzel, R. (1986). Unpaired Cysteine-54 Interferes with the Ability of an Engineered Disulfide to Stabilize T4-Lysozyme. *Biochemistry* **25**, 733-739.
54. Heinzelman, P., Komor, R., Kanaan, A., Romero, P., Yu, X. L., Mohler, S., Snow, C. & Arnold, F. (2010). Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Engineering Design & Selection* **23**, 871-880.
55. Voutilainen, S. P., Puranen, T., Siika-Aho, M., Lappalainen, A., Alapuranen, M., Kallio, J., Hooman, S., Viikari, L., Vehmaanpera, J. & Koivula, A. (2008). Cloning, expression, and characterization of novel thermostable family 7 cellobiohydrolases. *Biotechnol Bioeng* **101**, 515-28.
56. Voutilainen, S. P., Murray, P. G., Tuohy, M. G. & Koivula, A. (2009). Expression of *Talaromyces emersonii* cellobiohydrolase Cel7A in *Saccharomyces cerevisiae* and rational mutagenesis to improve its thermostability and activity. *Protein Eng Des Sel*.
57. Higgins, D. G., Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D. & Gibson, T. J. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Chapter 2

A Combination of Predictive Methods to Identify Stabilizing Mutations in Cellobiohydrolase Class I

Enzymes

Abstract

To further increase the stability of the thermostable chimeras created by SCHEMA structure-guided recombination, we used a combination of data analysis and predictive methods. Using the chimera thermostability screening data and comparing amino acid sequences, we identified groups of residues predicted to be responsible for large changes in stability. We used two additional metrics to further filter these mutations: consensus sequence alignments and FoldX $\Delta\Delta G$ predictions. Using an alignment of 40 CBHI sequences we calculated the frequencies of each amino acid at every position and chose mutations that resulted in an amino acid with high frequency. Using the FoldX force field, we simulated the effect on ΔG of folding of each mutation in a number of CBHI structures and chose those mutations that were predicted to be stabilizing in multiple structures. Choosing mutations based on a combination of all three methods, we increased the T_{50} of the most thermostable chimera by an additional 4.7 °C. The overall increase in stability resulted in a 10 °C increase in optimal temperature and a 50% increase in total sugar production at the optimal temperature compared to that of the most stable parent CBHI.

A. Sequence Regression

Up to this point we have treated recombination blocks as single entities and ignored the fact that they are amino acids sequences. The block property contributions coupled with a closer look at the block sequences can yield more information about the stability effects of smaller residue groups. This is done by comparing the sequences of stabilizing blocks to those of neutral or destabilizing blocks and selecting the amino acids that appear only in the stabilizing blocks. For example, the sequences of sub-block C for each parental CBHI are shown in Table 2.1. Relative to the *T. emersonii* CBHI, the *A. thermophilum* and *T. aurantiacus* CBHI sub-blocks result in an increase in T_{50} , while the *H. jecorina* CBHI sub-block results in a decrease and the *C. thermophilum* CBHI causes the sub-block monomera not to express. To identify the residues from the *A. thermophilum* and *T. aurantiacus* CBHI responsible for the increase in stability, we first can eliminate residues that have the same amino acids at that location in the *T. emersonii* and *H. jecorina* CBHIs. The remaining residues, colored purple in Table 2.1, are those most likely responsible for the stabilizing effect of the *A. thermophilum* and *T. aurantiacus* CBHI sub-block C.

Table 2.1. Sequences of block 7 sub-block C from each parental CBHI. Their effects on stability relative to the *T. emersonii* CBHI reference is given by a + for stabilizing, a - for destabilizing, or a NS for causing the sub-block monomera to not express. The colored residues are those that differ from the reference sequence. Those in purple are predicted to be stabilizing.

Effect on Stability	CBHI Source	Sequence
Ref.	<i>T. emersonii</i>	K R F Y I Q N S N V I P Q P N S D I S G
+	<i>A. thermophilum</i>	K R F Y V Q N G K V I P N S E S K I A G
+	<i>T. aurantiacus</i>	K R F Y V Q N G K V I P Q S E S T I S G
-	<i>H. jecorina</i>	N R Y Y V Q N G V T F Q Q P N A E L G S
NS	<i>C. thermophilum</i>	K R F Y V Q D G K I I A N A E S K I P G

The assumption made in selecting the amino acids only present in the stabilizing sequences is that on average most mutations have little effect on stability. In other words, when multiple residues together are neutral, it is much more likely that they are all individually neutral and not that several stabilizing and destabilizing mutations are masking each other. This can be stated mathematically as the mutations' effects on stability form a distribution around 0^{1; 2; 3}.

Using this method, we can identify residue groups predicted to have a significant effect on stability and then test individual mutations for incorporation into the set of thermostable chimeras. Even groups of residues with negative effects on stability can be of use, as reverting them leads to an increase in stability. Applying this method to our monomera and sub-block monomera data we identified nearly 50 residue groups

predicted to have a significant effect on stability. However, these groups contained anywhere from 2 to 40 residues each: still far too many mutations to test individually. Therefore, we sought to use a combination of this method and other metrics to choose individual mutations for testing.

B. Consensus Amino Acids from Multiple Sequence Alignments

One approach to identifying stabilizing amino acids is through the use of multiple sequence alignments of evolutionarily related (homologous) proteins. Care must be taken to ensure that the chosen proteins are not biased by evolutionary relationships, such as being derived from a common ancestor, which can result in a lack of statistical independence⁴. However, if enough sequences are used, the set of sequences can approximate a canonical ensemble, and the most probable distribution of amino acids at a specific position is given by Boltzmann's law. The actual frequency of an amino acid at a given position then measures its deviation from randomness and the effect on stability of a mutation is estimated by

$$\Delta\Delta G_{fold} = -RT \ln \frac{f_{mut}}{f_{WT}} \quad \text{Eq. 2.1}$$

where f_{mut} is the frequency of the new amino acid at that residue and f_{WT} is the frequency of the wild type amino acid at that position⁵. This increase in stability from a more frequent amino acid arises from selection for protein stability. Although above a critical stability threshold mutations have no effect on phenotype, there is still selective pressure to remain above that critical threshold and retain biological function. Among a

group of homologous proteins, amino acids selected for increased stability occur at frequencies above the average at a given position. However, this theory is only applicable to portions of the protein where stability is the main property under selection. Residues important for catalysis or substrate binding have other selective pressures that can be greater than that for stability, but even at these positions highly destabilizing mutations are often not allowed due to the marginal stability of natural proteins. Therefore, the consensus approach can be applied throughout the entire protein sequence.

To assemble enough sequences to reach an approximation of a canonical ensemble, we searched the protein database and selected 40 CBHI sequences (including the five CBHIs used as parents in the SCHEMA library) with at least 54% sequence identity (see Figure 2.7 in Supplemental Information). From this alignment we determined the frequencies of each amino acid at every residue position in the protein. These frequencies were used as an additional metric in selecting mutations from the sequence regression data for testing. Amino acids with a frequency of more than 0.20 at a given position were considered for inclusion as potentially stabilizing.

C. FoldX Force Field Energy Calculations

The third metric we used to select mutations for trial was the FoldX algorithm. FoldX gives a fast and quantitative estimation of mutations' effect on stability. It uses a full atomic description of the structure of the proteins and has been tested on a large

set of point mutants (1088) spanning most of the structural environments found in proteins^{6;7}. The free energy of folding (ΔG) of a target protein is calculated using

$$\Delta G = W_{vdw} \cdot \Delta G_{vdw} + W_{solvH} \cdot \Delta G_{solvH} + W_{solvP} \cdot \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + W_{mc} \cdot T \cdot \Delta S_{mc} + W_{sc} \cdot T \cdot \Delta S_{sc} \quad \text{Eq. 2.2}$$

where the ΔG_{vdw} term is the sum of the van der Waals contributions of all atoms with respect to the same interactions with the solvent. The ΔG_{solvH} and ΔG_{solvP} terms are the difference in solvation energy for apolar and polar groups respectively when these change from the unfolded to the folded state. The ΔG_{hbond} term is the free energy difference between the formation of an intramolecular hydrogen bond compared to intermolecular hydrogen-bond formation (with solvent). The ΔG_{wb} term is the extra stabilizing free energy provided by a water molecule making more than one hydrogen bond to the protein (water bridges that cannot be taken into account with nonexplicit solvent approximations)⁸. The ΔG_{el} term is the electrostatic contribution of charged groups, including the helix dipole. The $T \cdot \Delta S_{mc}$ term is the entropy cost of fixing the backbone in the folded state; this term is dependent on the intrinsic tendency of a particular amino acid to adopt certain dihedral angles⁹. Finally, the ΔS_{sc} term is the entropic cost of fixing a side chain in a particular conformation¹⁰.

As inputs, FoldX requires only a crystal structure and mutation and predicts the effect on ΔG according to eq. 2.5. We used FoldX to calculate the effect on stability of all of the mutations of interest. However, as the set of thermostable chimeras are

composed of blocks from five different CBHIs, we expect that their structures diverge significantly from the available natural structures. To circumvent errors that would arise from using an inappropriate structure, we instead calculated the effects on ΔG of each mutation in 39 different CBHI crystal structures available in the PDB. As no single structure perfectly matches the chimeras, using multiple structures allows us to choose mutations that are predicted to be stabilizing in the majority of structures.

D. Selecting Mutations

In choosing which potentially stabilizing mutations to test, we first chose to limit our search to mutations that could be incorporated into useful SCHEMA blocks and thus into the set of thermostable chimeras described in Chapter 1. We then applied all three of the criteria and evaluated each mutation based on all three values. To limit the number of mutations for testing to a tractable number, we chose mutations that satisfied at least two of our three criteria. The 13 chosen mutations and their criteria values are shown in Table 2.2, with the values that meet or exceed each cutoff shown in bold. The mutations cover a wide range in values from the regression data, from slightly destabilizing to highly stabilizing. Also of note is that the most stabilizing group of mutations (predicted to add 3.48 °C to T_{50}) contains 41 residues. This group comes from the highly stabilizing effects of block 7 (minus sub-block C) and cannot be broken down into smaller groups due to lack of measurements from so many of the sub-block monomers from the *C. thermophilum* CBHI not expressing. None of the mutations result in a rare amino acid at the position (frequency < 0.10), and several are present in

over half of the sequences sampled. All of the chosen mutations are predicted by FoldX to be significantly stabilizing in the majority of the sampled crystal structures.

The mutations were incorporated into thermostable chimeras, and their effects on T_{50} were measured (Table 2.2). Of the 13 mutations tested, five were found to be significantly stabilizing, five were neutral, two were significantly destabilizing, and one caused the chimera to no longer express. The five stabilizing mutations would combine for an increase in T_{50} of 4.5 °C if the mutations sum additively. All five of these mutations were combined into chimera CBHI TS0 (1515255*2 with sub-block C from the *A. thermophilum* CBHI), which is both the most stable chimera and contains the appropriate blocks for each of the mutations. The chimera with all five mutations, CBHI TS5, has a T_{50} of 70.0 ± 0.2 °C, an increase in 2.6 °C from the chimera alone. As with the chimeras vs. monomeras, we see a less than additive effect, resulting in a smaller than expected increase in stability.

Table 2.2. The chosen mutations and their values for each of the evaluating criteria in the given chimera. The “regression” column shows the predicted effect on T_{50} based on regression analysis of the monomera and sub-block monomera data. The “number in group” column shows the number of residues in the regression group predicted to have this effect. The “ f_{mut} ” column gives the fraction of CBHI sequences in the multiple sequence alignment that contain the new amino acid. The “ $\Delta\Delta G$ ” column gives the average value from the FoldX calculations on multiple CBHI structures with error given as the standard deviation between values from the different structures. The “ ΔT_{50} ” column gives the experimentally calculated effect on T_{50} of the mutation. NS indicates insufficient secretion for a T_{50} value. The values in bold satisfy the cutoff values for that particular criterion and those in red gave significantly stabilizing effects. The combination of the five stabilizing mutations in chimera 1515255*2 resulted in enzyme CBHI TS5)

Mutations	Chimera	Regression (°C)	Number in group	f_{mut}	$\Delta\Delta G$ (kcal/mol)	ΔT_{50} (°C)
S13P	1215225*2	-0.03	1	0.275	-1.43 ± 0.28	0.7
T41V	1315255*2	0.12	2	0.400	-1.55 ± 0.47	-0.3
Y60I	5515225*2	-0.06	1	0.450	-3.24 ± 6.33	0.1
Y60L	5515225*2	-0.12	2	0.400	-3.12 ± 5.90	0.8
V109L	5515225*2	-0.12	7	0.425	-1.35 ± 0.83	-0.9
T162K	5515355*2	0.10	2	0.775	-0.95 ± 0.33	-0.6
A199P	5515225*2	0.33	3	0.100	-0.91 ± .037	0.2
T252V	5515225*2	3.48	41	0.100	-1.01 ± 0.34	0.1
T268K	5515225*2	3.48	41	0.250	-0.88 ± 0.23	-0.1
S320P	5515225*2	-0.41	2	0.375	-2.14 ± 1.27	0.9
A378Y	5515225*2	0.34	4	0.175	-0.89 ± 0.49	0.6
Y425F	5515225*2	-0.34	4	0.525	-1.68 ± 1.08	1.5
N434G	5515225*2	-0.34	4	0.600	-2.59 ± 0.73	NS
S13P, Y60L, S320P, A378Y, Y425F (CBHI TS5)	1515255*2	N. A.	N. A.	N. A.	-9.26 ± 9.02	2.6

E. Evaluation of Individual Methods

Evaluating the results of our mutations does not clearly show that any one of the predictive methods outperforms the others. For example, some of the most stabilizing mutations predicted by FoldX and consensus had significantly stabilizing effects (Y60L

and Y425F, respectively), whereas others had destabilizing effects (T41V and T162K, respectively). Also, mutations barely above the cutoff for either method proved to be highly stabilizing (S13P and A379Y, respectively). To test whether any of these methods is efficient on its own, we took the top six mutations (that had not already been tested) predicted by FoldX or consensus alone and tested them in the thermostable chimeras. The results are shown in Table 2.3.

Table 2.3. The top six mutations predicted by consensus or FoldX alone. The “ f_{mut} ” column gives the fraction of CBHI sequences in the multiple sequence alignment that contain the new amino acid. The “ $\Delta\Delta G$ ” column gives the average value from the FoldX calculations on multiple CBHI structures with error given as the standard deviation between values from the different structures. The “ ΔT_{50} ” column gives the experimentally calculated effect on T_{50} of the mutation in a stable CBHI chimera. The values in bold satisfy the cutoff values for that particular criterion and those in red gave significantly stabilizing effects.

Mutation	Chimera	f_{mut}	$\Delta\Delta G$ (kcal/mol)	ΔT_{50} (°C)
S5T	1215225*2	0.725	0.06 ± 0.22	0.2
D52T	1315255*2	0.875	-0.07 ± 0.51	-0.9
S57D	1315255*2	0.725	-0.82 ± 1.12	-0.8
L109M	5515225*2	0.825	0.83 ± 1.02	-0.5
S125T	5515225*2	0.875	-0.30 ± 1.03	0.4
V215I	5515225*2	0.875	-0.01 ± 0.83	-0.3
N92K	5515225*2	0.125	-1.12 ± 0.59	2.1
N121G	5515225*2	0.425	-2.46 ± 0.94	-0.8
H206Y	5515225*2	0.150	-1.27 ± 0.69	0.7
S220K	5515225*2	0.075	-1.43 ± 0.50	0.0
D346V	5515225*2	0.100	-2.25 ± 1.25	1.0
T403D	5515225*2	0.025	-1.22 ± 0.55	0.2

From the results for these six mutations predicted by each method, it is clear that FoldX is better than consensus at predicting the stabilizing mutations, but still only results in correct predictions for CBHI half of the time. In addition, several of the

mutations chosen by consensus and one chosen by FoldX are actually destabilizing. To get a broader picture of how these methods performed, we plotted the predictions vs. the measured effects on T_{50} for both FoldX and consensus in Figure 2.1 using data from all 25 mutations tested so far.

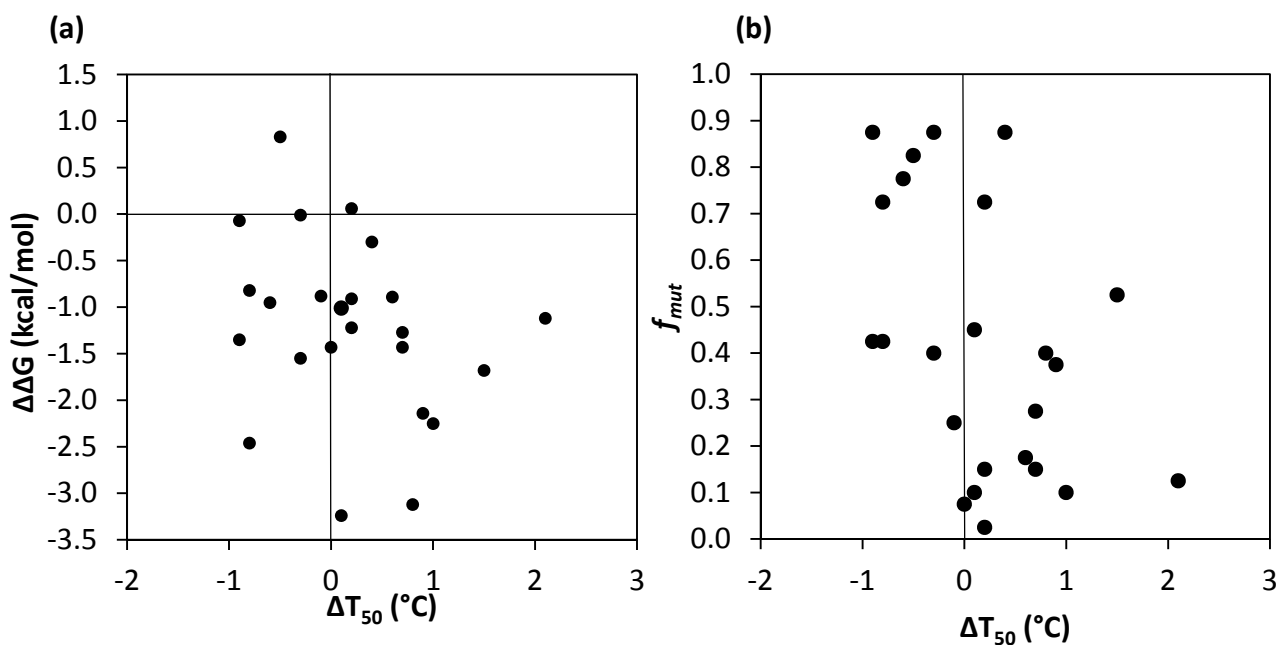


Figure 2.1. Plots of (a) FoldX stability predictions vs. actual effect on T_{50} and (b) mutation frequency vs. actual effect on T_{50} for the 25 mutations listed in Table 2.2 and Table 2.3.

From the plots, it is clear that there is little to no correlation. A good correlation for FoldX in Figure 2.1(a) would be a line with negative slope and most points appearing in quadrants II and IV. The presence of many points in quadrant III highlights the unreliability of FoldX alone. The cutoff chosen for mutation selection was a $\Delta\Delta G < -0.75$ kcal/mol, but moving this cutoff down to < -1.5 kcal/mol clearly increases the reliability. However, there are only a handful of mutations with predicted effects at

this level. Figure 2.1(b) shows that there is even less correlation between mutation frequency and changes in T_{50} . Increasing the cutoff level does not help, as most of the mutations with the highest frequency have destabilizing effects.

While FoldX is correct in its predictions only about half the time, this can be ample if there are enough mutations to test. At this point, we had tested most of the mutations that were predicted to be highly stabilizing in the majority of the CBHI structures used. However, there were many more mutations predicted to be stabilizing in fewer sequences. Therefore, we selected 19 more mutations predicted to be stabilizing in some structures, but with a wider spread in their predicted values between structures, reflected in the large standard deviation values shown in Table 2.4.

This group of mutations gave much poorer results than previous ones, with only 1 of 19 mutations resulting in an increase in T_{50} . This is not entirely surprising, as larger standard deviations in the FoldX predictions result from having stabilizing effects in some of the structures and destabilizing effects in others. We predicted that mutations predicted to be stabilizing in more of the structures would have a better chance of being stabilizing in the chimeras. The fact that this set of mutations with higher standard deviation values yielded far fewer stabilizing mutations supports this thinking.

Table 2.4. Mutations predicted by FoldX to be stabilizing in some but not all CBHI structures. The “ $\Delta\Delta G$ ” column gives the average value from the FoldX calculations on multiple CBHI structures with error given as the standard deviation between values from the different structures. The “ ΔT_{50} ” column gives the experimentally calculated effect on T_{50} of the mutation in a stable CBHI chimera and those in red gave a significantly stabilizing effect.

Mutation	Chimera	$\Delta\Delta G$ (kcal/mol)	ΔT_{50} (°C)
V226L	CBHI TS5	-1.52 ± 1.43	-0.8
T256I	CBHI TS5	-1.38 ± 0.59	-0.9
T256K	CBHI TS5	-1.17 ± 0.38	-2.0
T256V	CBHI TS5	-1.01 ± 0.34	-1.1
T272P	CBHI TS5	-1.08 ± 0.69	-0.9
D297K	CBHI TS5	-1.14 ± 0.82	-2.7
E322P	CBHI TS5	-1.25 ± 0.92	-1.8
A326G	CBHI TS5	-0.95 ± 1.78	-1.4
V328M	CBHI TS5	-1.47 ± 1.20	-3.5
T335P	CBHI TS5	-1.38 ± 1.19	-0.8
T335Q	CBHI TS5	-0.80 ± 0.40	-2.7
Q341M	CBHI TS5	-1.41 ± 0.78	-2.0
H354K	CBHI TS5	-2.03 ± 1.72	-1.7
H354R	CBHI TS5	-2.15 ± 1.63	-2.1
H354V	CBHI TS5	-2.08 ± 2.32	-3.2
T388I	CBHI TS5	-0.79 ± 1.25	0.5
T391P	CBHI TS5	-2.29 ± 0.62	NS
P395G	CBHI TS5	-1.40 ± 3.21	-1.8
V400A	CBHI TS5	-1.36 ± 1.94	-1.5

The one stabilizing mutation found in this group of mutations was incorporated into CBHI TS5 (described in Table 2.2) along with two of the stabilizing mutations in Table 2.3 (tyrosine was already present at position 206 in this chimera) to make CBHI TS8 (described in Table 2.5), which has a T_{50} of 72.0 ± 0.1 °C, an increase of 4.5 °C from

the best chimera and 9.1 °C from the most thermostable parent. This increase is again lower than the sum of the effects on stability of the individual mutations. A summary of all the stabilizing mutations and their effects on the stability of CBHI T50 is shown in Table 2.5 (note that the effects on stability of some of the mutations differs in CHI T50 than in the thermostable chimera in which it was first tested, i.e., the results listed in Table 2.2).

Table 2.5. Summary of the properties of stabilizing mutations in the CBHI T50 background. CBHI T50 contains block 7 from the *T. emersonii* CBHI with sub-block C from the *A. thermophilum* CBHI (hence 5* at block 7). The “ f_{mut} ” column gives the fraction of CBHI sequences in the multiple sequence alignment that contain the new amino acid. The “ $\Delta\Delta G$ ” column gives the average value from the FoldX calculations on multiple CBHI structures with error given as the standard deviation between values from the different structures. The “ ΔT_{50} ” column gives the experimentally calculated effect on T_{50} of the mutation compared to its “parent” enzyme (the enzyme listed before the mutations in the “enzyme” column).

Enzyme	f_{mut}	$\Delta\Delta G$ (kcal/mol)	T_{50} (°C)	ΔT_{50} (°C)
CBHI T50 (1515255*2)			67.4 ± 0.2	
CBHI T50 S13P	0.275	-1.43	67.8 ± 0.5	0.4
CBHI T50 Y60L	0.400	-3.12	67.8 ± 0.2	0.4
CBHI T50 S320P	0.375	-2.14	67.6 ± 0.7	0.2
CBHI T50 A378Y	0.175	-0.89	67.8 ± 0.3	0.4
CBHI T50 Y425F	0.525	-1.68	68.1 ± 0.1	0.7
CBHI T55 (CBHI T50 S13P, Y60L, S320P, A378Y, Y425F)	N. A.	N. A.	70.0 ± 0.2	2.6
CBHI T55 N92K	0.125	-1.12	70.8 ± 0.1	0.8
CBHI T55 D346V	0.100	-2.25	70.6 ± 0.1	0.6
CBHI T55 T388I	0.025	-0.79	70.5 ± 0.4	0.5
CBHI T55 N92K, D346V, T388I	N. A.	N. A.	72.1 ± 0.3	2.1

Overall, FoldX is useful for identifying stabilizing mutations when many can be tested in parallel, but it is not effective at quantitatively predicting the stability effects of individual mutations. This could in part be because of the lack of an accurate structure. The use of multiple homologous structures helps, and it appears that mutations predicted to be stabilizing in many structures are more likely to be stabilizing in the chimeras, whose structures are most likely similar to the parent structures at some, but not all regions. These results are in line with previously published works evaluating the effectiveness of FoldX and similar computational methods which concluded that the methods are good on average but not on a per mutation basis¹¹.

F. Structural Analysis of Stabilizing Mutations

An investigation into the structural changes caused by the selected mutations is hindered by the fact that crystal structures for the thermostable chimeras are not available. However, there are crystal structures of the parent CBHIs from *H. jecorina* and *T. emersonii*, as well as several other CBHIs not used as parents but that are structurally similar. These additional structures also have some of the mutant amino acids that are not present in the *H. jecorina* or *T. emersonii* CBHI sequences.

All of the crystal structures maintain the same overall fold, however significant deviation can be seen in large sections of the structures, including areas surrounding some of the mutations. These deviations suggest that without a chimera crystal structure, it is difficult to accurately identify the structural basis for a mutation's

stabilizing effects. Using multiple structures gives more chances to find one with an explanation for a mutation's effects, but it does not guarantee this.

The mutations tested, both destabilizing and stabilizing, are spread throughout the protein. Only two of the mutations (V215I and V226L, both destabilizing) are located near the enzyme's catalytic residues. Both positions face away from the substrate channel and so were presumed not to interfere with catalysis. With the exception of S13P, N92K, and Y425F, the stabilizing mutations are distant enough from each other to assume little interaction. This would suggest that the five mutations distant from each other would combine additively with each other, and the deviation from additivity seen when combining the mutations could be due to the three mutations close to each other. These three are located in antiparallel strands of a twisting beta sheet with N92K and Y425F in neighboring strands and S13P two strands from N92K (and three strands from Y425F), as seen in Figure 2.2. These mutations cause significant changes to the amino acids' structure (S13P), charge (N92K), or polarity (Y425F) and so likely have impacts on the local environment of the beta sheet. The combination of these effects so nearby could account for absence of additivity when combining these mutations.

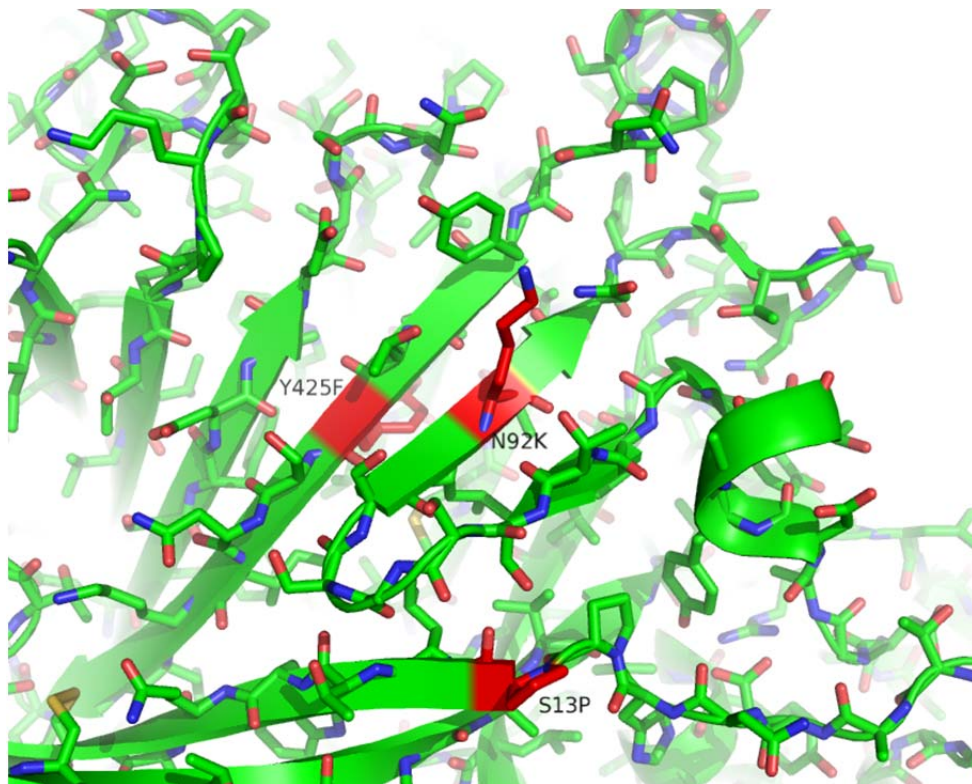


Figure 2.2. Crystal structure of the CBHI from *T. emersonii* (PDB ID 1Q9H) showing the location of the three stabilizing mutations S13P, N92K, and Y425F (red) that are located near each other in a beta sheet.

Of the stabilizing mutations, only Y425F is not located on the surface of the protein, meaning the majority of them involve interaction with the solvent. The beta sheet that contains Y425F is on the surface of the protein, but the residue is facing internally and is part of a hydrophobic pocket as seen in Figure 2.3. Tyrosine and phenylalanine have similar sizes but phenylalanine is more hydrophobic and thus packs better in this pocket.

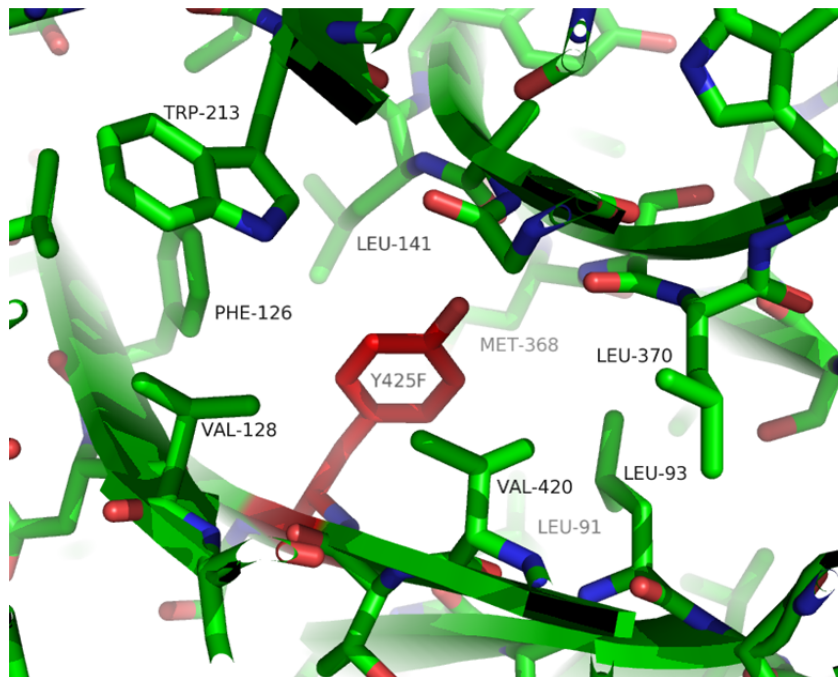


Figure 2.3. The hydrophobic pocket surrounding residue 425 (shown as a red tyrosine) in the *T. emersonii* CBHI crystal structure (PDB ID 1Q9H).

The mutation A378Y substitutes an aromatic amino acid in the correct orientation for a favorable pi-stacking interaction with tyrosine 247 that could account for its increase in stability. Mutations S320P, D350V and T388I all substitute hydrophobic amino acids near other hydrophobic residues (phenylalanine 270, tyrosine 429, and valine 393 respectively) that could form more favorable packing. In addition, when serine is present at position 320, its side chain points toward serine 254, leading to unfavorable electrostatic repulsion between the two polar oxygen atoms. The mutation N92K replaces a polar side chain with a positively charged one close to several polar residues, as seen in Figure 2.4. Three polar residues and a tyrosine are within 6 Å

of residue 92, and the negatively polarized oxygen atom of each has favorable electrostatic interactions with the positively charged nitrogen atom of the arginine in the mutation. While the arginine is too far from any one residue to form a salt bridge, the presence of so many polar residues nearby likely stabilizes the region. From examining the crystal structures, it is not clear why the mutations S13P and Y60L result in an increase in stability.

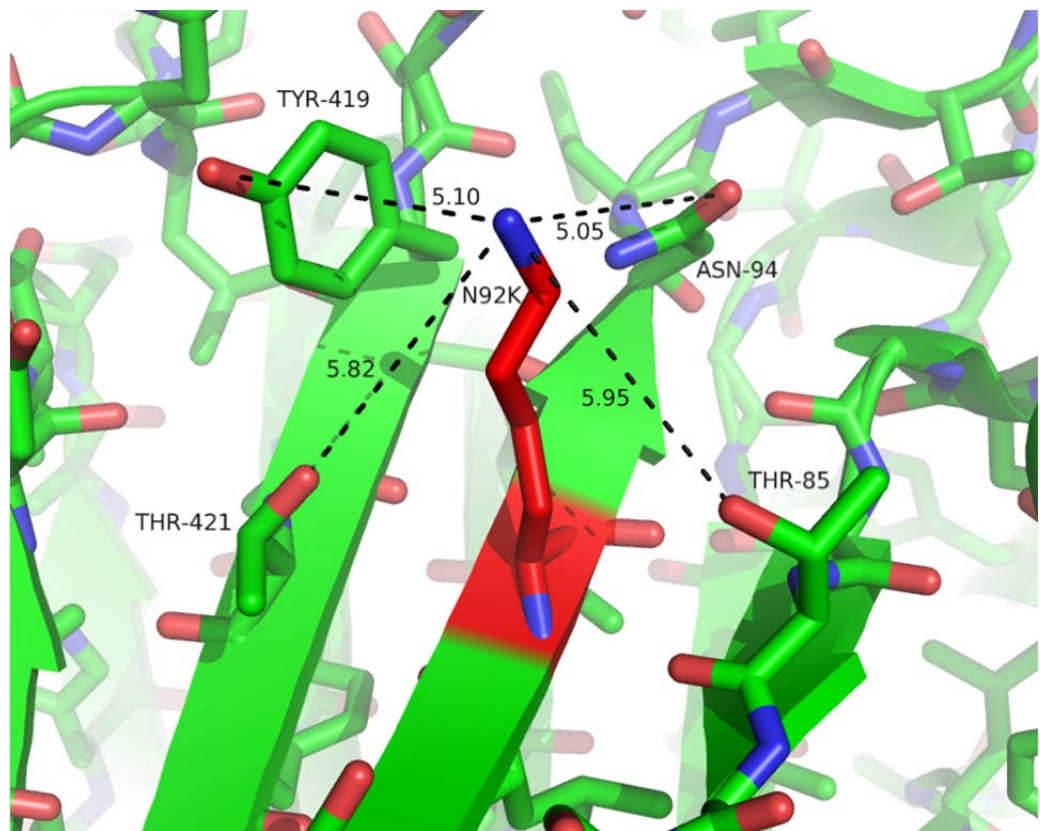


Figure 2.4. Mutation N92K (red) and residues within 6 Å electrostatically interacting with the arginine. *T. emersonii* CBHI crystal structure (PDB ID 1Q9H) shown.

G. Activity of Thermostable CBHIs

The overall goal of this work is not to simply create stable CBHIs, it is to create CBHIs with greater activity at higher temperatures. While we succeeded in producing stable CBHIs, we further characterized these proteins to verify that the increased stability translates into increased activity at higher temperature.

We tested the best chimera (CBHI TS0), CBHI TS0 with the first five stabilizing mutations (CBHI TS5), and CBHI TS0 with all eight stabilizing mutations (CBHI TS8) against the most stable parent, the CBHI from *T. emersonii*. We ran temperature profiling experiments (Section H of Materials and Methods) with yeast secretion culture on 4-methylumbelliferyl lactopyranoside (MUL). Samples were diluted to have equivalent hydrolysis rates at 45 °C where stability is not a factor and then reacted at 45, 50, 55, 60, 65, and 70 °C as shown in Figure 2.5. From the graph, it is clear that the optimal reaction temperature increases with stability, with the optimal temperature of the *T. emersonii* CBHI occurring around 55 °C and that of the most stable protein, CBHI TS8, occurring around 65 °C. The actual activity of the stable chimeras at their optimal temperatures is higher than that of the *T. emersonii* CBHI at its optimal temperature, but by less than 10%.

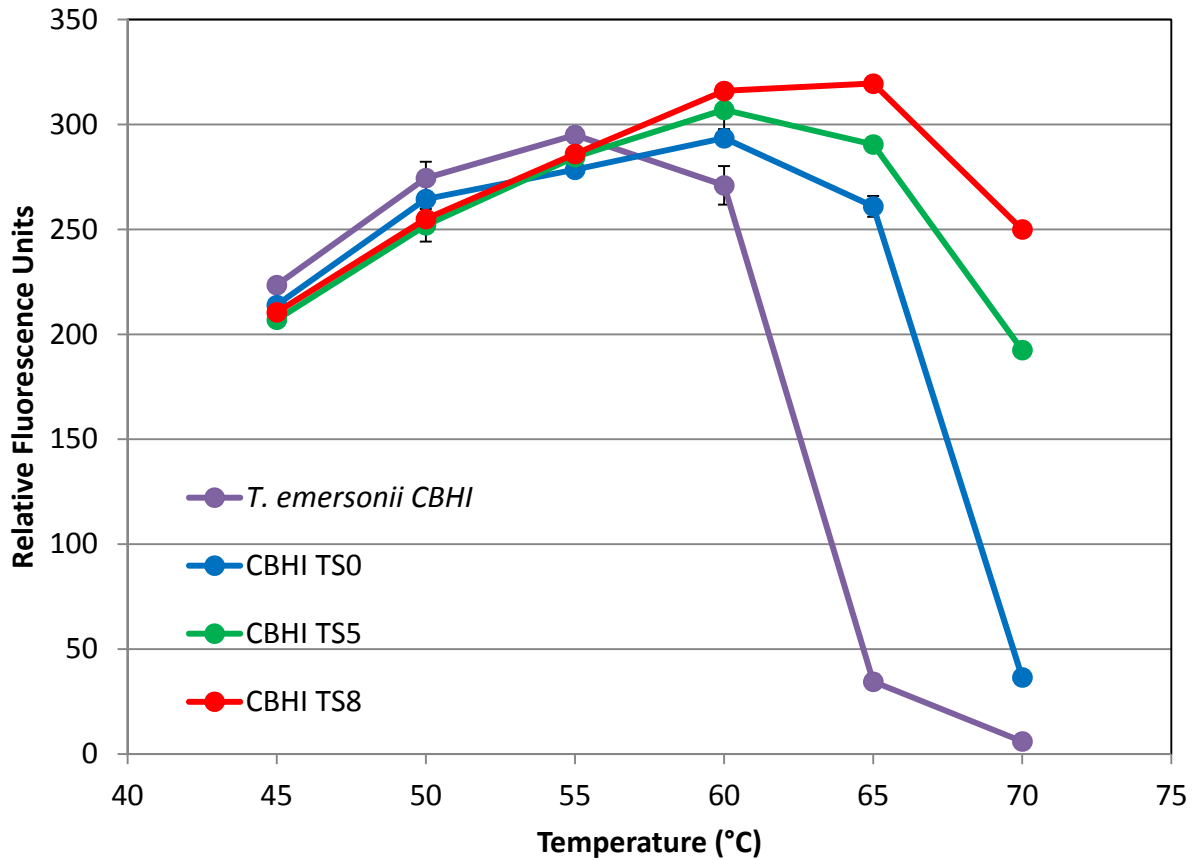


Figure 2.5. Temperature profiles of the *T. emersonii* CBHI, most stable chimera, and most stable chimera with thermostabilizing mutations on the soluble MUL substrate.

To further characterize the stable CBHIs under more stringent conditions, we purified the four enzymes and determined their concentrations as described in section C of Materials and Methods. We also switched to microcrystalline cellulose, a substrate that more closely resembles the industrial feedstock of pretreated biomass. We reacted the CBHIs at 45, 50, 55, 60, 65, and 70 °C as described in section I of Materials and Methods and measured the soluble sugar liberated as described in section J of Materials and Methods to produce the graph in Figure 2.6. Again we see an increase in optimal reaction temperature with increasing stability from the *T. emersonii* CBHI's optimal of

around 55 °C to CBHI TS8's optimal of around 65 °C. In the case of this solid substrate, increased stability leads to a large increase in product formation, with CBHI TS8 producing nearly 50% more sugar at its temperature optimum than the *T. emersonii* CBHI at its optimum. This is a significant increase and validates our strategy of increasing the enzyme's stability to promote higher activity at higher temperature.

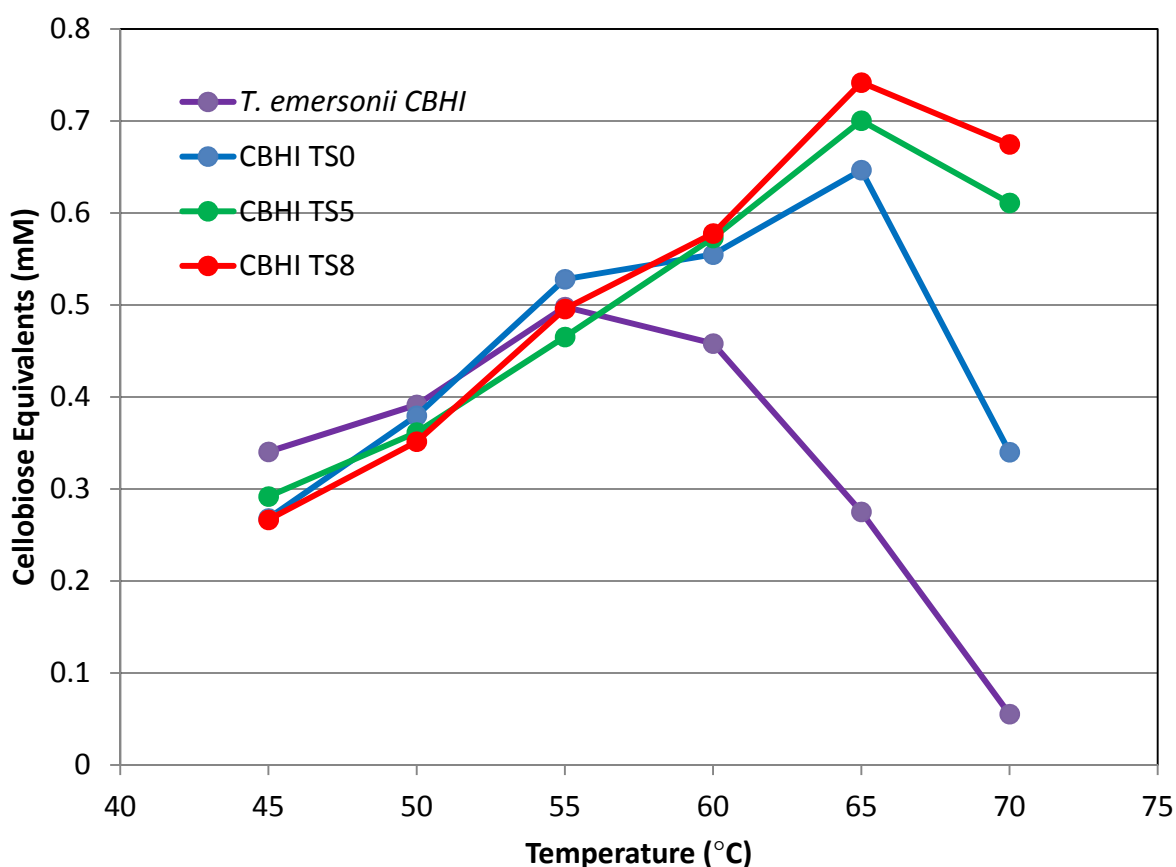


Figure 2.6. Temperature profiles of purified *T. emersonii* CBHI, the most stable chimera, and the most stable chimera with thermostabilizing mutations on solid microcrystalline cellulose.

The fact that there is significant increase in product formation in the temperature profiling on solid substrate and not on soluble substrate is most likely due

to the choice of reaction conditions, specifically the length of reaction. The temperature profiling on soluble substrate was done over 30 min, while that on the solid substrate was done over 20 hr. As stated earlier, increasing the stability of the enzyme and reacting it at higher temperatures allows the enzyme to react at a faster rate but also allows it to be active for longer periods of time, both resulting in increased product formation. To separate out the contributions to product formation of each effect, readings at earlier time points would be needed. These could be used to see if most of the product is formed early in the reaction (due to increased reaction rates) or continuously over the full course of the reaction (due to increased enzyme lifetime) or a combination of both.

H. Conclusions

Expanding on our previous work¹², we explored using a combination of predictive methods to further increase the stability of thermostable chimeras generated through SCHEMA structure-guided recombination. Combining regression data, a large consensus sequence alignment, and FoldX $\Delta\Delta G$ predictions, we identified eight individual mutations that further increased the T_{50} of the most stable chimera by 4.5 °C. While use of the methods led to the stabilizing mutations, it also predicted increased stability for a larger number of mutations that proved to be either neutral or destabilizing. The results show that these predictive methods can be used to select a set of mutations for testing, but they are unreliable in predicting the effects of individual mutations. Compared to the most stable recombination parent, this nearly 10 °C

increase in T_{50} translated into a corresponding 10 °C increase in optimal reaction temperature and a 50% increase in total sugar production at the optimal temperature. Further work must be done to see if the increase in sugar production is due to higher activity or longer enzyme lifetimes: e.g., a time point experiment could reveal whether the increase comes from high initial activity, a sustained activity over a long period of time, or a combination of both.

I. Supplemental Information

Hypocrea virens -----MYQKLAVISAFI-AAARAQQVCTQQAETHPPLTWQKCSSSG 40
Hypocrea lixii -----MYRKLAVISAFI-AAARAQQVCTQQAETHPPLTWQKCTASG 40
Hypocrea jecorina -----XSACTLQSETHPPLTWQKCSSGG 23
Aspergillus oryzae -MASLSLSKICRNALILSSVL-STAQQQVGTQYQTEHPSTWQTCNGGG 48
Neosartorya fischeri -MASAISFQVYRSALILSAFLPSITQAAQIGTYTTEHPSTWETCTSSG 49
Penicillium occitanis -MSALNSFNMYKSALILGSL-ATAGAQQIGTYTAETHPSLSWSTCKSGG 48
Penicillium funiculosum -MSALNSFNMYKSALILGSL-ATAGAQQIGTYTAETHPSLSWSTCKSGG 48
Penicillium marneffei -MSALNSFTMYKSALILGSL-ATAGAQQIGTLTTEHPPLTWSTCKSGG 48
Talaromyces stipitatus -MSALNSFKMYKNALILGSL-ATAHAQQIGNLTAETQPSLSWSTCTSSG 48
Botryotinia fuckeliana -----MTSRIALVSLF-AAVYQQVGTQYQTEHPSLTWQSCITAKG 39
Dictyostelium discoideum -----MYRILKSFILLSLVN-MSLSQKIGKLTPEVHPMTFQKCSSEGG 42
Thermoascus aurantiacus -----MYQRALLFS-FFLAAARAQQAGTVAENHPSLWQCCSSGG 40
Talaromyces emersonii -----MLRRALLSSSAILAVKAQQAGTATAENHPPLTWQECTAPG 41
Aspergillus fumigates -----MHQRALLFS-ALAVANAQQVGTQPEHPPLTWQKCTAAG 40
Aspergillus terreus -----MHQRALLFS-ALVGAVRAQQAGLTTEVHPPLTWQCTADG 40
Aspergillus nidulans -----MYQRALLFS-ALLSVSRAQQAGTAQEVHPSLWQRCESAG 40
Aspergillus niger -----MHQRALLFS-ALLTAVRAQQAGLTTEVHPSLWQKCTSEG 40
Penicillium oxalicum -MKGSISYQIYKALLSLLASVS-AQQAGTLTAESHPLTWQKCSAGG 48
Penicillium janthinellum -MKGSISYQIYKALLSALLNSVS-AQQVGTTLTAETHPLTWQKCTAGX 48
Penicillium chrysogenum -MASALSFKIYKNALLAAFLGAAQ-AQQVGTSTAETHPSLWQKCTAGG 48
Aspergillus aculeatus -MVDSFS--IYKTALLLS-MLATSN-AQQVGTQYTAETHPSLWQTCSSGG 45
Irpex lacteus -----MFHKAVLVAFSLVTIVHGQQAGTQTAENHPQLSSQKCTAGG 41
Phanerochaete chrysosporium -----MFRATLLAFTMAAMVFGQQVGTNTAENHRLTSQKCTKSG 41
Coniophora puteana -----MFPKSILLAFAPAAATSQQIGTSTAETHPPLTWQCTSSG 41
Lentinula edodes -----MFRTAALLSFAYLAVVYQQAGTSTAETHPPLTWQCTSSG 41
Volvariella volvacea -----MFPAATLFAFSLFAAVYQQVGTQLAETHPRLTWQKCTRSF 41
Agaricus bisporus -----MFPRSILLALSITAVALGQQVGTNMAENHPSLWQRCSTSSG 41
Marssonina brunnea -----MISSTFALASLFVAANAQQAGTNKPEHPPLEVSTCTASG 40
Aspergillus clavatus MLPSTISYRIYKNALFFA-ALFGAVQAQKVGTSKAEVHPSMAWQTCADG 49
Magnaporthe grisea -----MIRKITTLLAALVGVVRGQAACSLTAETHPSLWQKCSSGG 40
Acremonium thermophilum -----MYTKFAALAALVATVRGQAACSLTAETHPSLWQKCTAPG 40
Gibberella pulicaris -----MYRAIATASALIAAVRAQQVCSLTPETKPALSWKCTSSG 40
Fusarium venenatum -----MYRAIATASALIAAVRAQQVCSLTPETKPALSWKCTSSG 40
Fusarium poae -----MYRAIATASALIAAVRAQQVCSLTTETKPALTWKCTSSG 40
Gibberella avenacea -----MYRAIATASALIAAARAQQVCTLTETKPALTWKCTSSG 40
Nectria haematococca -----MYRAISLATALIASVRAQQACTSTQETHPPLTWKCTSSG 40
Humicola grisea -----MRTAKFATLAAALVASAAAQACSLTTERHPSLWKNCTAGG 41
Chaetomium thermophilum -----TETHPRLTWKRCSTSSG 16
Neurospora crassa -----MLAKFAALAALVASANAQAVCSLTAETHPSLWKNCTSSG 40
Alternaria alternata -----MTWQSCITAKG 10
: . *

Hypocrea virens --CTAQSGSVVLDANWRWTHDVKSTTNCYDGNWWSKTLCPDDATCAKNCC 88
Hypocrea lixii --CTPQQSGSVVLDANWRWTHDKSTTNCYDGNWWSSTLCPDDATCAKNCC 88
Hypocrea jecorina -TCTQQTGSSVIDANWRWTHATNSSTNCYDGNWWSSTLCPDNETCAKNCC 72
Aspergillus oryzae -SCSTNQSGSVVLDANWRWVHQGTSSSNCYTGNKWDTSYCSTNDACAQKCA 97
Neosartorya fischeri -SCATNQSGSVMDANWRWVHQVGTSTNCYTGNTWDTSIDTDETCATECA 98
Penicillium occitanis -SCTTNSGAITLDANWRWVHGVTSTNCYTGNTWNSAICDTDASCAQDCA 97
Penicillium funiculosum -SCTTNSGAITLDANWRWVHGVTSTNCYTGNTWNTAICDTDASCAQDCA 97
Penicillium marneffei -SCTSNSGSIIVLDANWRWVHVSNSSTNCYTGNTWNTNICDTDASCAQDCA 97
Talaromyces stipitatus -SCTSKSASITLDANWRWVHVSNGSTNCYTGNTWDTSIDTDTSCAQDCA 97
Botryotinia fuckeliana -SCTTNTGSIIVLDGNWRWTHGVGTSTNCYTGNTWDTLCPDDATCAQDCA 88
Dictyostelium discoideum -SCTETIQGEVVVDANWRWVHSAQQGQ-NCYTGNTWNPITCIPDDETCABENCY 90
Thermoascus aurantiacus -SCTTQNGKVVLDANWRWVHTTSGYTNCTYTGNTWDTSIDCPDDVTCAQNCA 89
Talaromyces emersonii -SCTTQNGAVVLDANWRWVHDVNGYTNCTYTGNTWDPYCPDDETCANCA 90
Aspergillus fumigates -SCSQSGSVVIDANWRWLHSTKDTTNCYTGNTWNTLCPDNECAQNCA 89
Aspergillus terreus -SCTEQSGSVVIDSNWRWLHSTNGSTNCYTGNTWDESLCPDNECAQNCA 89
Aspergillus nidulans -SCTEVAGSVVLDANWRWVHSDGYTNCTYTGNTWDTLCPDNECAQNCA 89
Aspergillus niger -SCTEQSGSVVIDANWRWVHSDNSTNCYTGNTWDTLCPDDETCANCA 89
Penicillium oxalicum -SCTPVSGSVVIDANWRWVHDKNG-KNCYTGNTWDTLCPDDKTCANCA 96
Penicillium janthinellum --CSQVSGSVVIDANWPXVHSTSGSTNCYTGNTWDTLCPDDVTCAANCA 96
Penicillium chrysogenum -SCTEQSGKVVVIDSNWRWLHSTNGYTNCTYTGNTWDTLCPDDVTCAANCA 97
Aspergillus aculeatus -SCTTTSGSVVIDANWRWVHEVGGYTNCTYSGNTWDSICSTDTTCASECA 94
Irpex lacteus -SCTSASTSVVLDANWRWVHTTSGYTNCTYTGNTWDTASICSDPVSCANCA 90

Phanerochaete chrysosporium -GCSNLNLIKIVLDANWRWLHSTSGYTNCYTGNQWDATLCPDGKTC AANCA 90
Coniophora puteana -CTTESSGSVVL DANWRWLHTVDGYTNCYTGNEDWTTICTSAEVC AEQCA 90
Lentinula edodes -SCTTQSSSVLDSNWRWTHVVGYYTNCYTGNENWTTVC PDGTTCAANCA 90
Volvariella volvacea GCQTQSNGAIVLDANWRWVHNVGGYTNCYTGNWNTSLCPDGATCAKNCA 91
Agaricus bisporus -CQN-VNGKVTLDANWRWTHRINDFTNCYTGNEDWTSICPDGVTCAENCA 89
Marssonina brunnea --CTTSAQSIVVDANWRWLHSTTGYTNCYTGNWWDATLCPDGATCAENCA 88
Aspergillus clavatus -TCTTKNGKVVVIDANWRWVHDVKGYTNCYTGNWNAELCPDNESCAENCA 98
Magnaporthe grisea S-CTNVAGSVTIDANWRWHTTSGYTNCYTGNKWDTSICSTNADCA SKCC 89
Acremonium thermophilum S-CTTVSGQVVIDANWRWLHQTNSSSTNCYTGNEDWTSICSSD TDCATKCC 89
Gibberella pulicaris --CSNVQGSVVIDANWRWTHQLSGSTNCYTGNKWDTSICTSGKVC AEKCC 88
Fusarium venenatum --CSNVQGSVVIDANWRWTHQLSGSTNCYTGNKWDTSICTSGKVC AEKCC 88
Fusarium poae --CSNVQGSVVIDANWRWTHQVSGSTNCHTGNKWDTSVCTSGKVC AEKCC 88
Gibberella avenacea --CTDVKGSVVIDANWRWTHQTSSTNCYTGNKWDTSVCTSGE TCAQKCC 88
Nectria haematococca --CTEVSQSVVVDANWRWTHQVSGSTNCYTGNKWDTSICPDGKTC AEKCC 88
Humicola grisea Q-CQTVQASITLDSNWRWTHQVSGSTNCYTGNKWDTSICTDAKSCAQNCC 90
Chaetomium thermophilum N-CSTVNGAVTIDANWRWHTVSGSTNCYTGNEDWTSICSDGKSCAQTCC 65
Neurospora crassa --CTNVAGSITVDANWRWTHITSGSTNCYSGNEWDTSLCSTNTDCATKCC 88
Alternaria alternate S-CTNKNGKIVIDANWRWLHKKEGYDNCYTGNEDWATA CPDNKACAANCA 59
: : * : * * * * * * * * * * * *

Hypocrea virens LDGAAYSSTYGITTTSSDSLTFVFTQS---NVGARLYLMA--TDSYQE 132
Hypocrea lixii LDGANYSSTYGVTTSGDALTLQFVTAS---NVGSRLYLMA--NDTSYQE 132
Hypocrea jecorina LDGAAYASTYGVTTSGNSLSIGFVTQSA-QKNVGARLYLMA--SDTTYQE 119
Aspergillus oryzae LDGADYSSTYGITTTSGNSLRLNFVTSNSNGKNVGSRVYMA--DDTHYEV 145
Neosartorya fischeri VDGADYESTYGVTTSGSQIRLNFVTQNSNGANVGSRLYMA--DNTHYQM 146
Penicillium occitanis LDGADYSSTYGITTTSGNSLRLNFVT---GSNVGSRTYLMA--DNTHYQI 141
Penicillium funiculosum LDGADYSSTYGITTTSGNSLRLNFVT---GSNVGSRTYLMA--DNTHYQI 141
Penicillium marneffeii LDGADYSSTYGITTTSGNSLRLNFVT---SSNVGSRTYLMA--DNTHYQM 141
Talaromyces stipitatus VDGADYSSTYGITTTSGNSLRLNFVT---GSNVGSRTYLMA--DDTHYQL 141
Botryotinia fuckeliana LEGADYSSTYGITTTSGNSLRLNFVTQS-ANKNIGSRVYMA--DTTHYKT 135
Dictyostelium discoideum LDGANYESVYGVTTSEDSVRLNFVTQS-QGKNIGSRFLMS--NESNYQL 137
Thermoascus aurantiacus LDGADYSSTYGVTTSGNALRRLNFVTQS-SGKNIGSRLLYLQ--DDTTYQI 136
Talaromyces emersonii LDGADYESTYGVTTSSGSSLKLNFTV---GSNVGSRLLYLQ--DDSTYQI 134
Aspergillus fumigates LDGADYAGTYGVTTSGSELEKLSFVT---GANVGSRLYLMQ--DDTYQH 133
Aspergillus terreus LDGADYESTYGITTTSGDALTLTFVT---GENVGSRVYLMAE--DDESYQT 134
Aspergillus nidulans VDGADYESTYGITTSNGDSLTLKFVT---GSNVGSRVYLMQ--DDETYQM 133
Aspergillus niger LDGADYESTYGVTTDGDLSLTLKFVT---GSNVGSRLLYMDT--SDEGYQT 134
Penicillium oxalicum VDGASYASTYGVTTSGNSLRLNFVTQA-SQKNIGSRLLYLE--NDTTYQK 143
Penicillium janthinellum VDGARRQHLR-VTTSGNSLRLNFVTTA-SQKNIGSRLLYLE--NDTTYQK 142
Penicillium chrysogenum LDGADYKGTYGVTASGSSLRRLNFVTQA-SQKNIGSRLLYMA--DDSKYEM 144
Aspergillus aculeatus LEGATYESTYGVTTSGSSLRRLNFVTTA-SQKNIGSRLLYLLA--DDSTYET 141
Irpex lacteus LDGADYAGTYGITTTSGDALTLKFVTGS---NVGSRVYLMED--ETNYQM 134
Phanerochaete chrysosporium LDGADYGTYGITASGSSLKLFVFTGS---NVGSRVYLMAD--DTHYQM 134
Coniophora puteana LDGADYESTYGITTTSGDALTLKFVFTQS-QKNVGSRVYLMAD--DTHYQM 137
Lentinula edodes LDGADYESTYGITTSGNALTLKFVTASA-QTNVGSRVYLMAPGSETEYQM 139
Volvariella volvacea LDGANYSSTYGITTTSGNALTLKFVFTQSE-QKNIGSRVYLLSE--DTKYQL 138
Agaricus bisporus LDGADYAGTYGVTTSGTALTLKFVTESQ-QKNIGSRLLYLMAD--DSNYEI 136
Marssonina brunnea LEGAEYASTYGITTTSGNSLKLFSVTKSA-QTNVGSRVYLMAPGSETKYQM 137
Aspergillus clavatus LEGADYAAATYGVTTSGNALSLKFVFTQS-QQKNIGSRLYMMK--DDNTYET 145
Magnaporthe grisea VDGANYQTYGASTSGNALSQYVFTQ--SSGKNVGSRLYLLSE--SENKYQM 136
Acremonium thermophilum LDGADYGTYGVTASGNLSLNLKFVFTQGPYSKNIGSRMYLME--SESKYQG 137
Gibberella pulicaris IDGAEYASTYGITSSGNQLSLSFVTKGAYGTNIGSRTYLME--DENTYQM 136
Fusarium venenatum IDGAEYASTYGITSSGNQLSLSFVTKGTYGTNIGSRTYLME--DENTYQM 136
Fusarium poae VDGADYASTYGITSSGNQLSLSFVTKGSYGTNIGSRTYLME--DENTYQM 136
Gibberella avenacea LDGADYAGTYGITSSGNQLSLGFVTKGSFSTNIGSRTYLME--NENTYQM 136
Nectria haematococca VDGADYAAATYGVTTSGDQLSLSFVTKGAYATNVGSRVYLMQ--DDETYQM 136
Humicola grisea VDGADYESTYGITTTNGDSLTLKFVTKGQHSTNVGSRTYLMD--GEDKYQT 138
Chaetomium thermophilum VDGADYSSTYGITTTSGDSLNLKFVTKHQHGTNVGSRVYLMQ--NDTKYQM 113
Neurospora crassa VDGAEYESTYGITTTSGNSLRLNFVTKGYSYTNIGSRTYLMA--GADAYQM 136
Alternaria alternate VDGADYSSTYGITTAGNSLKLKFVTKGYSYSTNIGSRTYLMA--DDTTYEM 107
: : * * : : * : * * * * * * * * * * * *

Hypocrea virens FTLSGN-EFSFDVDVSQLPCGLNGALYFVSMADGGQSKYPTNAAGAKYQ 181
Hypocrea lixii FTLSGN-EFSFDVDVSQLPCGLNGALYFVSMADGGQSKYPTNAAGAKYQ 181
Hypocrea jecorina FTLLGN-EFSFDVDVSQLPCGLNGALYFVSMADGGVSKYPTNTAGAKYQ 168
Aspergillus oryzae YKLLNQ-EFTFDVDVSKLPCGLNGALYFVVMADGGVSKYPTNNAAGAKYQ 194

Neosartorya fischeri FKLLNQ-EFTFDVDSNLP CGLNGALYFVTMDEDGGVSKYPNNKAGA QYG 195
Penicillium occitanis FDLLNQ-EFTFTVDVSHLP CGLNGALYFVTMDADGGVSKYPNNKAGA QYG 190
Penicillium funiculosum FDLLNQ-EFTFTVDVSNLPCGLNGALYFVTMDADGGVSKYPNNKAGA QYG 190
Penicillium marneffei FDLLNQ-EFTFTVDVSNLPCGLNGALYFVTMDADGGVSKYPNNKAGA QYG 190
Talaromyces stipitatus FNLLNQ-EFTFTVDASTLPCGLNGALYFVSMADGGVSKQPNNKAGA QYG 190
Botryotinia fuckeliana FNLLNQ-EFTFDVDSNLP CGLNGALYFANLPADGGIS--STNTAGAE YG 182
Dictyostelium discoideum FHVLGQ-EFTFDVDSNLD CGLNGALYLVSMDS DGG SARFPTNEAGAK YG 186
Thermoascus aurantiacus FKLLGQ-EFTFDVDSNLP CGLNGALYFVAMDADGGLSKYPGNKAGAK YG 185
Talaromyces emersonii FKLLNR-EFSFDVDSNLP CGLNGALYFVAMDADGGVSKYPNNKAGAK YG 183
Aspergillus fumigates FNLLNH-EFTFDVDSNLP CGLNGALYFVAMDADGGMSKYP SNKAGAK YG 182
Aspergillus terreus FDLVGN-EFTFDVDSNLP CGLNGALYFTSMADGGVSKYPANKAGAK YG 183
Aspergillus nidulans FDLLNN-EFTFDVDSNLP CGLNGALYFTSMADGGLSKYEGNTAGAK YG 182
Aspergillus niger FNLLDA-EFTFDVDSNLP CGLNGALYFTAMDADGGVSKYPANKAGAK YG 183
Penicillium oxalicum FNLLNQ-EFTFDVDSNLP CGLNGALYFVDMADGGMAKYPTNKAGAK YG 192
Penicillium janthinellum FNLLNQ-EFTFDVDSNLP CGLNGALYFVDMADGGMAKYPTNKAGAK YG 191
Penicillium chrysogenum FQLLNQ-EFTFDVDSNLP CGLNGALYFVAMDEDDGGMARYPTNKAGAK YG 193
Aspergillus aculeatus FKLFNR-EFTFDVDSNLP CGLNGALYFVSMADGGVSRFPTNKAGAK YG 190
Irpex lacteus FKLMNQ-EFTFDVDSNLP CGLNGALYFVQMDQDGGTSKFPNNKAGAK FG 183
Phanerochaete chrysosporium FQLLNQ-EFTFDVDSNLP CGLNGALYLSAMDADGGMAKYPTNKAGAK YG 183
Coniophora puteana FNPLNQ-EFSFTVDVSQLPCGLNGALYFSQMDADGGLSKYSTNKAGA QYG 186
Lentinula edodes FNPLNQ-EFTFDVDSALPCGLNGALYFSEMDADGGLSEYPTNKAGAK YG 188
Volvariella volvacea FNPLNQ-EFTFDVDSALPCGLNGALYFSAMDADGGMSKFPNNAAGAK YG 187
Agaricus bisporus FNLLNK-EFTFDVDSKLP CGLNGALYFSEMAADGGMS--STNTAGAK YG 183
Marssonina brunnea FKLLNK-EFTFDVDSKMP CGVNGALYFSEMEDGGMARHP TNKAGAK YG 186
Aspergillus clavatus FKLLNQ-EFTFDVDSNLP CGLNGALYFVSMADGGLSRYTGNAGAK YG 194
Magnaporthe grisea FNLLGN-EFTFDVDSKLG CGLNGALYFVSMADGGQSKYSGNKAGAK YG 185
Acremonium thermophilum FTLLGQ-EFTFDVDSNLGCGLNGALYFVSMDDLGGVSKYTTNKAGAK YG 186
Gibberella pulicaris FQLLGN-EFTFDVDSNIGCGLNGALYFVSMADGGKAKYPGNKAGAK YG 185
Fusarium venenatum FQLLGN-EFTFDVDSNIGCGLNGALYFVSMADGGKAKYPGNKAGAK YG 185
Fusarium poae FQLLGN-EFTFDVDSNIGCGLNGALYFVSMADGGKAKYPGNKAGAK YG 185
Gibberella avenacea FQLLGN-EFTFDVDSNIGCGLNGALYFVSMADGGKARYPANKAGAK YG 185
Nectria haematococca FSLLGN-EFTFDVDSQISCGVNGALYFVSMDEDDGKAKADGNKAGAK YG 185
Humicola grisea FELLGN-EFTFDVDSNIGCGLNGALYFVSMADGGLSRYPGNKAGAK YG 187
Chaetomium thermophilum FELLGN-EFTFDVDSNIGCGLNGALYFVSMADGGMSKYSGNKAGAK YG 162
Neurospora crassa FELLGN-EFTFDVDSGTGCGLNGALYFVSMDDLGGKAKYTNNKAGAK YG 185
Alternaria alternata FKFTGNQ-EFTFDVDSNLP CGFNALYFVSMADGGLKKYSTNKAGAK YG 157
: . *:* * * * * * . * : * * * * * * * *

Hypocrea virens TGYCDSQCPRDLKFIHQANVDGWQPSNNANTGIGGHGSCCSEMDIWEA 231
Hypocrea lixii TGYCDSQCPRDLKFIHQANVEGWEPSSNNANTGVGGHSCCSEMDIWEA 231
Hypocrea jecorina TGYCDSQCPRDLKFIHQANVEGWEPSSNNANTGIGGHGSCCSEMDIWEA 218
Aspergillus oryzae TGYCDSQCPRDLKFIHQANVEGWVSTNNANTGTGNHGSCEALDIWES 244
Neosartorya fischeri VGYCDSQCPRDLKFIHQANVEGWTSSNNNTGLGNYGSCCAELDIWES 245
Penicillium occitanis VGYCDSQCPRDLKFIHQANVEGWTSSANNANTGIGNHGSCCAELDIWEA 240
Penicillium funiculosum VGYCDSQCPRDLKFIHQANVEGWTSSNNNTGIGNHGSCCAELDIWEA 240
Penicillium marneffei VGYCDSQCPRDLKFIHQANVEGWAPSSNNNTGIGNHGSCCSELDIWEA 240
Talaromyces stipitatus VGYCDSQCPRDLKFIHQANVEGWQPSNNNSNTGLGNYGSCCAELDIWEA 240
Botryotinia fuckeliana TGYCDSQCPRDMKFIHQANVDGWVPSNNANTGVGNHGSCEALDIWEA 232
Dictyostelium discoideum TGYCDAQCPRLKFIHQANVDGWIPSTNNPNTGYNLSGCCAEMDLWEA 236
Thermoascus aurantiacus TGYCDSQCPRDLKFIHQANVEGWQPSANDPNAGVGNHGSCEALDIWEA 235
Talaromyces emersonii TGYCDSQCPRDLKFIHQANVEGWQPSNNANTGIGDHGSCCAEMDIWEA 233
Aspergillus fumigates TGYCDSQCPRDLKFIHQANVEGWEPSSSDKNAGVGGHGSCEALDIWEA 232
Aspergillus terreus TGYCDSQCPRDLKFIHQANVEGWTSSNDKNAGVGGHGSCEALDIWEA 233
Aspergillus nidulans TGYCDSQCPRDLKFIHQANVEGWEPSSSDANAGVGGMGTCPEALDIWEA 232
Aspergillus niger TGYCDSQCPRDLKFIHQANVDGWEPSSNNNTGIGNHGSCCEALDIWEA 233
Penicillium oxalicum TGYCDSQCPRDLKFIHQANVEGWTSSNDPNSGVGGHGTCCALDIWEA 242
Penicillium janthinellum TGYCDSQCPRDLKFIHQANVDGWTSSKNDVNSGIGNHGSCCAEMDIWEA 241
Penicillium chrysogenum TGYCDAQCPRLKFIHQANVEGWEPSSSDVNGGTGSYGSCCAEMDIWEA 243
Aspergillus aculeatus TGYCDSQCPRDLKFIHQANVEGWEPSSSDVNTAGTGNHGSCEALDIWEA 240
Irpex lacteus TGYCDSQCQDIKFIHQANIVDWTASAGDANSGTGSFGTCCQEMDIWEA 233
Phanerochaete chrysosporium TGYCDSQCPRDLKFIHQANVEGWNATS--ANAGTGNYGTCCTEMDIWEA 231
Coniophora puteana TGYCDSQCPRDLKFIHQANVQLNWT--STSTNSGTGSLGSCCEALDIWEA 234
Lentinula edodes TGYCDSQCPRDLKFIHQANVEGWTSSSTSPNAGTGGTGI CCNEMDIWEA 238
Volvariella volvacea TGYCDSQCPRDLKFIHQANVEGWTSSNDTNTAGTGNHGSCEALDIWEA 237
Agaricus bisporus TGYCDSQCPRDLKFIHQANVEGWTSSNDVNTAGTGNHGSCEALDIWEA 233
Marssonina brunnea TGYCDAQCARDVKFIHQANVEGWTSSNDVNTAGTGNHGSCEALDIWEA 236

Dictyostelium discoideum RMGNTSFFGPNK--MIDTNSVITVVVTFITDDGSSDGLKLSIKRRLVYQDG 332
Thermoascus aurantiacus RQGNHSFYGPGQ--IVDTSSKFTVVVTFITDDGTPSGTLTEIKRFRVYQNG 331
Talaromyces emersonii RMGNTSFYGPBK--IIDTTKPFVTVVTFITDDGTDGTGLSEIKRFRYIQNS 329
Aspergillus fumigates RMGNESFYGPBK--IVDTKSKMTVVVTFITADGTDGALSEIKRFRVYQNG 328
Aspergillus terreus RMGNTSYYGPK--IVDTNSVMTVVVTFIFIG---DGGSLSEIKRRLVYQNG 325
Aspergillus nidulans RMGNTSFYGPGA--IIDTSSKFTVVVTFIFIA---DGGSLSEIKRFRVYQNG 324
Aspergillus niger RMGNDSFYGPGK--TIDTGSKMTVVVTFITD---GSGSLSEIKRFRVYQNG 326
Penicillium oxalicum RQGVTFNYGPGM--TVDTKSPFTVVVTFITDDGTDGTGLSEIKRFRVYQNG 338
Penicillium janthinellum RMGVTNFYGPGE--TIDTKSPFTVVVTFITNDGTSTGTGLSEIKRFRVYQNG 337
Penicillium chrysogenum RMGNQSFYGPBK--IVDTESPFTVVVTFITNDGTSTGTGLSEIKRFRVYQNG 339
Aspergillus aculeatus RFGNTNFYGPBK--TVDNSKPFVTVVTFITHDGTDTGTGLTEIRRLVYQNG 338
Irpex lacteus RMGNTEFYGKGL--TVDTSQKFTIVVTFISDDGTADGNLAEIRRFVYQNG 327
Phanerochaete chrysosporium RMGDQTFGLKGL--TVDTSKPFVTVVTFITNDGTSAGTLTEIRRLVYQNG 321
Coniophora puteana RMGDTTFYGSGE--TVDTSQPFVTVVTFITSDNTTTGTGLSEIRRLVYQNG 328
Lentinula edodes RMGDTSFYGPGL--TVDTTSKITVVVTFITSDNTTTGDLTAIRRIYVQNG 331
Volvariella volvacea RMGDKSFYGPGL--TVNTQQKFTVVVTFITNNNSSSGTLREIRRLVYQNG 331
Agaricus bisporus RMGDKSFYGPGM--TVDTNQPIVVVTFITDNGSDNGLQEIRRIYVQNG 326
Marssonina brunnea RMGVTDFYGEK--KIDTSKMTVVVTFITDDNTDTGTGLVDIRRFVYQNG 329
Aspergillus clavatus RQGNKSFYGPBK--TVDTKKKMTVVVTFITNDGTATGTGLSEIKRFRVYQNG 340
Magnaporthe grisea RQGNRTFYGPGSNFVNDSSKKVTVVTFIFIS---SGQLTDIKRFRVYQNG 329
Acremonium thermophilum RMGDTSFYGPBK--TVDTGSKFTVVVTFITG---SDGNLSEIKRFRVYQNG 329
Gibberella pulicaris RQGNKTFYGPGSFNVDTTKKVTVVTFIFHKG---SNGRLSEITRLVYQNG 330
Fusarium venenatum RQGNKTFYGPGSFNVDTTKKVTVVTFIFHKG---SNGRLSEITRLVYQNG 330
Fusarium poae RQGNKTFYGPGSFNVDTTKKVTVVTFIFHKG---SNGRLSEITRLVYQNG 330
Gibberella avenacea RQGNKTFYGRGSDFNVDTTKKVTVVTFIFKKG---SNGRLSEITRLVYQNG 330
Nectria haematococca RQGNTEFYGPGEFTVDTTKKVTVVTFIFIKG---TSGGLSEIKRFRVYQNG 330
Humicola grisea RQGNKTFYGK--MTVDTTKITVVVTFIFLKD---ANGDLGEIKRFRVYQNG 330
Chaetomium thermophilum RQGDKTFYGK--MTVDTTKMTVVVTFIFKKN---SAGVLEIKRFRVYQNG 305
Neurospora crassa RMGNTFYGEK--KTVDTSSKFTVVVTFIFIKD---SAGDLAEIKRFRVYQNG 328
Alternaria alternata RMGVKDFYGK--TVDTSSKFTVVVTFIFIG---TGDAMEIKRFRVYQNG 298
* * : * . : : . * : * * * * * * * * * * .

Hypocrea virens VKFQQPNAQLSGYSG--NTLNSDYCAAEQAAFGGT--SFTDKGGLTQFNKA 368
Hypocrea lixii VKFQQPNAQVGSYSG--NTINTDYCAAEQTAFFGGT--SFTDKGGLAQINKA 368
Hypocrea jecorina VTFQQPNAELGSYSG--NELNDYCTAEAEAFGGG--SFSDKGGLTQFKKA 355
Aspergillus oryzae VTYPQPSADVSLG--NTINSEYCTAENTLFEBSGSFAKHGGLAGMGEA 388
Neosartorya fischeri VTYAQPDSISGITG--NAINADYCTAENTVDFGPGTFAKHGGSFAMSEA 389
Penicillium occitanis VVIPQSSKISGISG--NVINSDYCAAEISTFGGTASFNKHGGLTNMAAG 384
Penicillium fusiculosum VVIPQSSKISGISG--NVINSDFCAEELSAFGETASFNTNHGGLKNMGEA 384
Penicillium marneffei VVHPQSSKISGVS--NVINSDFCAEISTFGETASFNTNHGGLPKMSAG 384
Talaromyces stipitatus KVFAQPSKIDIGIS--NAINSDYCSAEISTFGGNPSFTKHGGLAGVSTA 384
Botryotinia fuckeliana KVIPNSYSTISGVS--NSITTFPCDAQKTAFGDPTSFSDHGGLASMSAA 376
Dictyostelium discoideum NVISQSVSTIDGVEG--NEVNEEFCTNQKVFGEDEDSFTKHGGLAKMGEA 380
Thermoascus aurantiacus KVIPQSESTISGVTG--NSITTEYCTAQKAAFGDNTGFFTHGGLQKISQA 379
Talaromyces emersonii NVIPQPSNDISGVTG--NSITTEFCTAQKQAFGDTDDFSQHGGLAKMGAA 377
Aspergillus fumigates KVIANSVSNVAGVSG--NSITSDFCTAQKKAFGDEDIFAKHGGLSGMGKA 376
Aspergillus terreus KVIANAQSNVDGVTG--NSITSDFCTAQKTAFGDQDIFSKHGGLSGMGDA 373
Aspergillus nidulans EVIPNSESNISGVEG--NSITSEFCTAQKTAFGDEDIFAQHGGLSAMGDA 372
Aspergillus niger NVIANADSNISGVTG--NSITDFCTAQKKAFGDEDIFAEHNLKLAGISDA 374
Penicillium oxalicum KVIGQPQSTVAGVSG--NSITDSFCKAQKAAFGDITDDFTKHGALAGMGAA 386
Penicillium janthinellum KVIGNPQSTIVGVS--NSITDSWCNAQKSAFGDTNEFSKHGGMAGMGAG 385
Penicillium chrysogenum KVIPQSVSTISAVTG--NSITDSFCSAQKTAFKDITDVFVAKHGGMAGMGAG 387
Aspergillus aculeatus VVIGNGPSTYTAASG--NSITSEFCKAEKTLFGDTNVFETHGGLSAMGDA 386
Irpex lacteus KVIPNSVQITGIDP--VNSITDFCTQKTVFGDITNNFAAKGGLKQMGEA 376
Phanerochaete chrysosporium KVIQNSSVKIPGIDL--VNSITDNFCSQKTAFGDTNYFAQHGLKQVGEA 370
Coniophora puteana KVIQNSNTDISGLST--YNSITDDYCTAQKTAFGDITDSFSSHGGLAKMGDS 377
Lentinula edodes QVIQNSMSNIAGTP--TNETTDFCDQKTAFGDITNTFSEKGGTLTGMGA 380
Volvariella volvacea RVIQNSKVNIPGMPSTMDSVTTEFCNAQKTAFNDFSFQKGGMANMSEA 381
Agaricus bisporus VTIQNSNVNIPGIDS--GNSISAEFCDAQEAFGDERSFQIDRGLSGMGEA 375
Marssonina brunnea VTFANPNSTVAGVTE--NSLTDSEFCEAQKTAFGDNNIFKEKGGLAAMGES 377
Aspergillus clavatus KVIANSESTWPNLGG--NSLTDNDFCKAQKTVFGDMDFSKHGMEGMGAA 388
Magnaporthe grisea KVIPNSQSTITGVTG--NSVTQDYCDKQKTAFGDQNVFNQRGGLRQMGDA 377
Acremonium thermophilum KVIPNSESKIAGVSG--NSITDFCTAQKTAFGDITNVFEERGGLAQMKA 377
Gibberella pulicaris KVIANSESKIAGVPG--SSLTPEFCTAQKVFVGDITDDFAKKGAWGMSDA 378
Fusarium venenatum KVIANSESKIAGVPG--SSLTPEFCTAQKVFVGDITDDFAKKGAWGMSDA 378
Fusarium poae KVIANSESKIAGNPG--SSLTDFCTQKVFVGDITDDFAKKGAWGMSDA 378

Gibberella avenacea KVIANS**ESKIP**GN**SG**--SSL**TADFC**S**KQKSV**FG**DIDDF**S**KKGG**WS**GM**SDA 378
Nectria haematococca **KVIGNPESK**VAS**NP**G--NS**VTEEF**C**SAQKKA**F**GDDDD**F**VAKGG**FS**Q**MSDA 378
Humicola grisea **KIIPNSE**ST**IP**VE**G**--NS**ITQ**W**CDRQ**K**VAF**GD**IDDF**N**RKGG**M**KQ**MGKA 378
Chaetomium thermophilum **KIIANAESK**IPGN**P**G--NS**ITQ**EW**CD**A**QKVA**FG**DIDDF**N**RKGG**M**AQ**MSKA 353
Neurospora crassa **KVIENSQ**SN**VD**GV**S**G--NS**ITQ**S**FCNA**Q**KTA**FG**DIDDF**N**KKGG**L**KQ**MGKA 376
Alternaria alternate **KTIAQ**PASAV**P**VE**G**--NS**ITTK**F**CD**Q**QKAV**FG**DIT**Y**TFK**D**KGG**MAN**MA**KA 346

: . : . : * * * : . .

Hypocrea virens L**SGG**-M**VLV**MS**LW**DD**YYAN**ML**WLD**ST**YPT**NAT-AST**PGAKR**G**SC**ST**SS**GV 416
Hypocrea lixii F**QGG**-M**VLV**MS**LW**DD**YAVN**ML**WLD**ST**YPT**NAT-AST**PGAKR**G**SC**ST**SS**GV 416
Hypocrea jecorina T**SGG**-M**VLV**MS**LW**DD**YYAN**ML**WLD**ST**YPT**NET-SST**PGAVR**G**SC**ST**SS**GV 403
Aspergillus oryzae M**STG**-M**VLV**MS**LW**DD**YYAN**ML**WLD**SN**YPT**NES-T**SKPG**V**AR**G**TC**ST**SS**GV 436
Neosartorya fischeri M**STG**-M**VLV**MS**LW**DD**YYAD**ML**WLD**ST**YPT**NAS-SST**PGAVR**G**SC**ST**DS**GV 437
Penicillium occitanis M**EAG**-M**VLV**MS**LW**DD**YAVN**ML**WLD**ST**YPT**NAT-G-T**PGAAR**G**TC**AT**TS**GD 431
Penicillium funiculosum L**EAG**-M**VLV**MS**LW**DD**YSVN**ML**WLD**ST**YPA**NET-G-T**PGAAR**G**SC**PT**TS**GN 431
Penicillium marneffei I**SAG**-M**VLV**MS**LW**DD**YDVN**ML**WLD**ST**YPT**NAT-G-T**PGAAR**G**SC**AT**TS**GD 431
Talaromyces stipitatus L**KNG**-M**VLV**MS**LW**DD**YSVN**ML**WLD**ST**YPT**NAT-G-T**PGAAR**G**TC**ST**SS**GS 431
Botryotinia fuckeliana F**EAG**-M**VLV**L**SLW**DD**YYAN**ML**WLD**ST**YPV**G**KT**-S--AG**GP**R**G**TC**DT**SSGV 422
Dictyostelium discoideum L**KDG**-M**VLV**L**SLW**DD**YQAN**ML**WLD**SS**YPT**TSS-P**TD**P**G**V**AR**G**TC**PT**TS**GV 428
Thermoascus aurantiacus L**AQG**-M**VLV**MS**LW**DD**HAAN**ML**WLD**ST**YPT**DAD-P**DP**T**P**G**V**AR**G**TC**PT**TS**GV** 427
Talaromyces emersonii M**QGG**-M**VLV**MS**LW**DD**YAAQ**ML**WLD**SD**YPT**DAD-P**TP**P**G**I**AR**G**TC**PT**DS**GV 425
Aspergillus fumigates L**SE**--M**VL**IMS**I**W**DD**H**HSS**MM**WLD**ST**YPT**DAD-P**SK**P**G**V**AR**G**TC**E**H**G**AD** 423
Aspergillus terreus M**SA**--M**VL**L**IS**I**W**DD**HNS**MM**WLD**ST**YPE**DAD-A**SEP**G**V**AR**G**TC**E**H**GV**GD 420
Aspergillus nidulans A**SA**--M**VL**L**IS**I**W**DD**HSS**MM**WLD**SS**YPT**DAD-P**SQ**P**G**V**AR**G**TC**E**H**G**AD** 419
Aspergillus niger M**SS**--M**VL**L**IS**L**W**DD**YYAS**M**EW**L**SD**Y**PE**NAT-A**TD**P**G**V**AR**G**TC**D**SE**SGV 421
Penicillium oxalicum F**EEG**-M**VLV**MS**LW**DD**HNS**NML**WLD**ST**YPT**TAS-ST**TL**G**AKR**G**SC**D**ISS**GA 434
Penicillium janthinellum L**ADG**-M**VLV**MS**LW**DD**HAS**DML**WLD**ST**YPT**NAT-ST**TP**G**AKR**G**TC**D**IS**R-R 432
Penicillium chrysogenum L**AEG**-M**VLV**MS**LW**DD**HAAN**ML**WLD**ST**YPT**SAS-ST**TP**G**AKR**G**SC**D**ISS**GE 435
Aspergillus aculeatus L**GDG**-M**VLV**L**SLW**DD**HAAD**ML**WLD**SD**YPT**TSC-ASS**P**G**V**AR**G**TC**PT**T**TGN** 434
Irpex lacteus V**KNG**-M**VL**L**AL**SL**W**DD**YAAQ**ML**WLD**SD**YPT**TAD-P**SQ**P**G**V**AR**G**TC**PT**TS**GV 424
Phanerochaete chrysosporium L**RTG**-M**VL**L**AL**SI**W**DD**YAA**NML**WLD**SN**YPT**NKD-P**ST**P**G**V**AR**G**TC**AT**TS**GV 418
Coniophora puteana F**AAG**-V**VL**V**L**SV**W**DD**YAAQ**ML**WLD**SD**YPT**TAD-AST**PG**V**AR**G**TC**AT**TS**GA 425
Lentinula edodes F**SRG**-M**VLV**L**SI**W**DD**AA**EM**L**WLD**ST**YPV**G**KT**-G--P**GAAR**G**TC**AT**TS**GV 426
Volvariella volvacea L**RRG**-M**VLV**L**SI**W**DD**HA**AN**ML**WLD**SN**YPT**DRP-AS**Q**P**G**V**AR**G**TC**PT**SS**GK 429
Agaricus bisporus L**DRG**-M**VLV**L**SI**W**DD**HA**VN**ML**WLD**SD**YPL**DAS-P**SQ**P**G**I**S**R**G**TC**S**R**DS**GK 423
Marssonina brunnea L**G**R**G**-V**VL**V**MS**I**W**DD**HAAN**ML**WLD**SS**YPT**TKD-P**SAP**G**V**TR**G**TC**AP**TS**GV** 425
Aspergillus clavatus L**AEG**-M**VLV**MS**LW**DD**HNS**NML**WLD**SN**SPT**TGT-ST**TP**G**V**AR**G**SC**D**I**SS**GD 436
Magnaporthe grisea L**AKG**-M**VLV**MS**VW**DD**HHS**Q**ML**W**L**D**ST**Y**P**T**S**---T**AP**G**AKR**G**SC**ST**SS**GK 423
Acronium thermophilum L**AEP**-M**VLV**L**SV**W**DD**HA**VN**ML**WLD**ST**YPT**DS---T**K**P**G**A**R**G**TC**PT**TS**GV 423
Gibberella pulicaris L**EAP**-M**VLV**MS**LW**HD**HNS**NML**WLD**ST**YPT**DS---T**KL**G**AQ**R**G**SC**ST**SS**GV** 424
Fusarium venenatum L**EAP**-M**VLV**MS**LW**HD**HNS**NML**WLD**ST**YPT**DS---T**KL**G**AQ**R**G**SC**ST**SS**GV** 424
Fusarium poae L**EAP**-M**VLV**MS**LW**HD**HNS**NML**WLD**ST**YPT**DS---T**AL**G**SQ**R**G**SC**ST**SS**GV** 424
Gibberella avenacea L**ESPP**M**VLV**MS**LW**HD**HNS**NML**WLD**ST**YPT**DS---T**KL**G**AQ**R**G**SC**AT**TS**GV** 425
Nectria haematococca L**AAP**-M**VL**T**MS**L**W**DD**HKAN**L**WLD**ST**YPV**DS---T**G**A**G**S**KR**G**TC**ST**DS**GV 424
Humicola grisea L**AGP**-M**VLV**MS**I**W**DD**H**ASN**ML**WLD**ST**FP**V**DAA**--G**K**P**G**A**ER**G**AC**PT**TS**GV 425
Chaetomium thermophilum L**EGP**-M**VLV**MS**VW**DD**HYAN**ML**WLD**ST**YPI**D**KA**--G**TP**G**AER**G**AC**PT**TS**GV 400
Neurospora crassa L**AKP**-M**VLV**MS**I**W**DD**HA**AN**ML**WLD**ST**YPV**E---G**GP**G**AYR**G**EC**PT**TS**GV 421
Alternaria alternate L**ANG**-M**VLV**MS**LW**DD**HYS**NML**WLD**ST**YPT**D**KNP**D**LD**L**TGR**G**EC**ET**SS**GV 395

: * * : * * * * : : * * * * * * * * *

Hypocrea virens P**SQ**I**ESQ**SP**NAK**V**VF**SN**IR**FG**PI**ST**GG**ST**GN**PP**P**GT**ST**T**RLP**----- 459
Hypocrea lixii P**AQ**VE**AQ**SP**NSK**VI**Y**SN**IR**FG**PI**ST**GG**T**GS**N**PP**GT**ST**T**RAP**----- 459
Hypocrea jecorina P**AQ**VE**SQ**SP**NAK**V**TF**SN**IK**FG**PI**ST**GN**PS**G**----- 434
Aspergillus oryzae P**SE**VE**ASN**P**SAY**V**AYS**NI**K**V**G**PI**ST**FK**S**----- 465
Neosartorya fischeri P**AT**IE**SE**SP**DSY**V**TY**SN**IK**V**G**PI**ST**F**SSG**S**G**S**G**S**G**S**G**S**G**S**G**S----- 480
Penicillium occitanis P**K**TV**ESQ**SG**SSY**V**TF**SD**IR**V**GF**PN**ST**F**SG**S**ST**GG**STTT**T**ASR**----- 474
Penicillium funiculosum P**K**TV**ESQ**SG**SSY**V**VF**SD**IK**V**GF**PN**ST**F**SG**GT**ST**GG**STTT**T**ASG**----- 474
Penicillium marneffei P**K**T**LES**Q**S**SG**SSY**V**Y**SD**IK**V**GF**PN**ST**F**SG**T**ST**GG**STTT**T**TTKPS**----- 476
Talaromyces stipitatus P**K**TV**EANS**P**NAH**V**IF**SD**IR**V**G**PL**N**ST**FSG**---S**T**ST**PG**GG**SS**----- 471
Botryotinia fuckeliana P**AS**VE**ASS**P**NAY**V**VY**SN**IK**V**G**A**IN**ST**YG**----- 450
Dictyostelium discoideum P**SK**VE**Q**NP**NAY**V**VY**SN**IK**V**G**P**ID**ST**YK**K----- 457
Thermoascus aurantiacus P**AD**VE**SQ**PN**SY**V**IY**SN**IK**V**G**P**IN**ST**FTAN**----- 457
Talaromyces emersonii P**SD**VE**SQ**PN**SY**V**TY**SN**IK**FG**PI**ST**FTAS**----- 455
Aspergillus fumigates P**EN**VE**SQ**HP**DAS**Y**TF**SN**IK**FG**PI**ST**YEG**----- 452
Aspergillus terreus P**ET**VE**SQ**HP**GAT**V**TF**SK**IK**FG**PI**ST**YSSNSTA**----- 453
Aspergillus nidulans P**DV**VE**SE**H**ADAS**V**TF**SN**IK**FG**PI**ST**F**----- 446

<i>Hypocrea virens</i>	QVLNPFYSQCL---	505
<i>Hypocrea lixii</i>	QVLNPFYSQCL---	505
<i>Hypocrea jecorina</i>	-----	
<i>Aspergillus oryzae</i>	-----	
<i>Neosartorya fischeri</i>	QVQNA YYSQCL---	535
<i>Penicillium occitanis</i>	TVVNP YYSQCL---	529
<i>Penicillium funiculosum</i>	TVVNP YYSQCL---	529
<i>Penicillium marneffeii</i>	TVVNP YYSQCL---	536
<i>Talaromyces stipitatus</i>	TVINP YYSQCL---	526
<i>Botryotinia fuckeliana</i>	-----	
<i>Dictyostelium discoideum</i>	-----	
<i>Thermoascus aurantiacus</i>	-----	
<i>Talaromyces emersonii</i>	-----	
<i>Aspergillus fumigates</i>	-----	
<i>Aspergillus terreus</i>	-----	
<i>Aspergillus nidulans</i>	-----	
<i>Aspergillus niger</i>	-----	
<i>Penicillium oxalicum</i>	TKQNE YYSQCL---	545
<i>Penicillium janthinellum</i>	TKQND WYSQCL---	537
<i>Penicillium chrysogenum</i>	QKQGD YYSQCL---	529
<i>Aspergillus aculeatus</i>	TKQND YYSQCL---	540
<i>Irpex lacteus</i>	HVLNP YYSQCY---	526
<i>Phanerochaete chrysosporium</i>	HVLNP YYSQCY---	516
<i>Coniophora puteana</i>	-----	
<i>Lentinula edodes</i>	TSSGP YYSQCL---	516
<i>Volvariella volvacea</i>	-----	
<i>Agaricus bisporus</i>	HVIND FYSQCF---	506
<i>Marssonina brunnea</i>	-----	
<i>Aspergillus clavatus</i>	TKQND YYSQCL---	539
<i>Magnaporthe grisea</i>	-----	
<i>Acremonium thermophilum</i>	KYSND WYSQCL---	523
<i>Gibberella pulicaris</i>	KKIND FYSQCQ---	510
<i>Fusarium venenatum</i>	KKIND FYSQCQ---	507
<i>Fusarium poae</i>	KKIND FYSQCQ---	511
<i>Gibberella avenacea</i>	KKIND FYSQCQ---	513
<i>Nectria haematococca</i>	KEQNT WYSQCVASA	519
<i>Humicola grisea</i>	TKLND WYSQCL---	525
<i>Chaetomium thermophilum</i>	TELNP WYSQCL---	505
<i>Neurospora crassa</i>	QKIND YYSQCV---	521
<i>Alternaria alternate</i>	-----	

Figure 2.7. ClustalW2 multiple sequence alignment of 40 CBHI sequences used to determine amino acid frequencies at each residue position¹³. The consensus symbols have the following meaning: an * (asterisk) indicates positions which have a single, fully conserved residue, a : (colon) indicates conservation between groups of strongly similar properties, and a . (period) indicates conservation between groups of weakly similar properties. The colors have the following meaning: red indicates small and hydrophobic, blue indicates acidic, magenta indicates basic, and green indicates everything else.

1. References

1. Tawfik, D. S., Tokuriki, N., Stricher, F., Schymkowitz, J. & Serrano, L. (2007). The stability effects of protein mutations appear to be universally distributed. *Journal of Molecular Biology* **369**, 1318-1332.

2. Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C. & Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 606-611.
3. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5869-5874.
4. Jackel, C., Bloom, J. D., Kast, P., Arnold, F. H. & Hilvert, D. (2010). Consensus Protein Design without Phylogenetic Bias. *Journal of Molecular Biology* **399**, 541-546.
5. Steipe, B., Schiller, B., Pluckthun, A. & Steinbacher, S. (1994). Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *Journal of Molecular Biology* **240**, 188-192.
6. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology* **320**, 369-387.
7. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research* **33**, W382-W388.
8. Petukhov, M., Cregut, D., Soares, C. M. & Serrano, L. (1999). Local water bridges and protein conformational stability. *Protein Science* **8**, 1982-1989.
9. Munoz, V. & Serrano, L. (1994). Intrinsic Secondary Structure Propensities of the Amino-Acids, Using Statistical Phi-Psi Matrices - Comparison with Experimental Scales. *Proteins-Structure Function and Genetics* **20**, 301-311.
10. Abagyan, R. & Totrov, M. (1994). Biased Probability Monte-Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins. *Journal of Molecular Biology* **235**, 983-1002.
11. Schreiber, G., Potapov, V. & Cohen, M. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering Design & Selection* **22**, 553-560.
12. Heinzelman, P., Komor, R., Kanaan, A., Romero, P., Yu, X. L., Mohler, S., Snow, C. & Arnold, F. (2010). Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Engineering Design & Selection* **23**, 871-880.
13. Higgins, D. G., Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D. & Gibson, T. J. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Chapter 3

Construction and Structural Analysis of a Thermostable P450 Enzyme with Broad Substrate Specificity

Abstract

The fact that most random mutations are destabilizing constrains evolution, with the stability of most enzymes dropping to just above the minimum stability threshold with accumulation of enough mutations. This prevents further evolution, as additional mutations are highly likely to cause the protein to unfold and no longer function. Excess stability prolongs a protein's ability to accept mutations by creating a stability buffer that can absorb mutations' destabilizing effects without negatively affecting the protein's fitness. An enzyme with high thermostability and broad substrate specificity is ideal for directed evolution for improved activity on varied compounds: it has the required stability to accept a large number of mutations and initial activity on varied compounds that can be increased and optimized by those mutations. Here, we identify a P450_{BM3} variant, 9-10A, with broad substrate specificity and then thermostabilize it without negatively affecting its substrate scope to produce enzyme 9-10ATS. Enzyme 9-10ATS with the F87A mutation is used as parent for active site mutagenesis at eight positions to produce variants with activity on a number of structurally diverse compounds, including protected sugars, alkaloids, and steroids. These variants are then further improved through random mutagenesis. We determined the structure of 9-10ATS using x-ray crystallography and examination clearly shows the structural basis of the gains in thermostability. Comparison with other P450_{BM3} structures reveals that 9-10ATS has a widened substrate channel, allowing larger substrates to enter the active

site. 9-10ATS also shows changes in the dynamic portions of the protein that move upon substrate binding.

A. Introduction

Protein stability is intricately tied to many properties, often in ways not fully understood. One area that has recently attracted interest is the interplay between protein stability and evolution. This interplay comes from the fact that stability is one of the defining components of protein fitness. For an enzyme, fitness (W) can be defined as the flux of the catalyzed reaction given by

$$W = [E]_0 f \quad \text{Eq. 3.1}$$

where $[E]_0$ is the concentration of functional protein and f is its function (which includes k_{cat} , K_d , and any other appropriate parameters, depending on the particular protein)^{1, 2}.

The main determinate of $[E]_0$ is stability, which is measured as the difference in free energy of the native and unfolded state, ΔG . This is a thermodynamic measurement of stability and does not reflect kinetic stability, but it is a good approximation for stability within cellular environments and is straightforward to measure and predict³.

The relationship between fitness and stability is then given by

$$W \propto [E]_0 = \mathbf{1} - \frac{1}{e^{\frac{\Delta G}{RT} + 1}} \quad \text{Eq. 3.2}$$

which gives a sigmoidal curve with midpoint at $\Delta G = 0$ ³. A striking characteristic of this type of curve is that the same change in ΔG can have drastically different effects on

fitness depending on the initial value of ΔG . Changes have larger effects on fitness the closer the initial ΔG is to 0, meaning that marginally stable proteins are much more affected by destabilizing mutations, while the fitness of highly stable proteins is little affected by destabilizing mutations.

Evolution affects protein fitness largely due to the fact that most random mutations are deleterious. Estimates of the fraction of random mutations that are deleterious range from 33% to 40%^{4; 5; 6} measured in different experimental protein systems to 70%⁷ when using the experimentally validated FoldX algorithm to calculate effects on stability alone^{8; 9}. While “deleterious” can mean negatively affecting the proteins’ fitness in any number of ways (fold, function, etc.), it is known that most mutations affect protein stability and not function³. In fact, it is estimated that 80% of deleterious mutations are due to loss of stability¹⁰. While stability is a property affected by myriad factors, here we use thermostability as a proxy for stability at the temperatures of the experiments, as it is straightforward to measure and accounts for a large portion of protein stability.

The high percentage of destabilizing mutations makes highly stable sequences rare and results in most naturally occurring proteins being only marginally stable¹¹. Modest stability leads to a high percentage of folded molecules³, resulting in little selective pressure to increase stability beyond this marginal level. The ΔG of folding for most proteins is between -3 and -10 kcal/mol^{12; 13} which is equivalent to one or a few intermolecular bonds (the energy of a single hydrogen bond is 2-10 kcal/mol). Thus, the

effect on stability of even a single mutation in the average protein is on the same order of magnitude as the protein's net folding stability. Looking back at Eq. 3.2, this puts most proteins in the regime where they are highly affected by individual deleterious mutation and leads to an exponential drop in fitness with random mutations given by

$$W \approx e^{-\alpha n} \quad \text{Eq. 3.3}$$

where α is the fraction of random mutations that are deleterious and n is the number of random mutations³. Using a value of $\alpha = 0.36$ (the average of experimental studies), after the accumulation of five random mutations the fitness of an average protein will decline to <20%. This low tolerance to accumulating multiple mutations has led researchers to conclude that protein stability is the limiting factor of organism genetic diversity and evolutionary rates¹⁴.

A natural extension of this model is that proteins with stability above the minimum level will have an extra buffer before the exponential decline¹⁵. In this buffer region, mutations will decrease stability but not fitness (since even modest stability gives such a high percentage of folded proteins, excess stability does not appreciably increase this percentage), allowing the accumulation of more mutations and thus more genetic diversity. Once the stability buffer has been exhausted, additional mutations will cause the exponential drop in fitness described in Eq. 3.3. This stability buffer is said to increase a protein's mutational robustness, which in turn increases its evolvability^{15; 16}. This theory has been validated using both simulations^{15; 17; 18} and experiments^{18; 19}.

The experimental results have given insight into the nature of different mutations and their effects on protein stability and function. Bloom et al. showed that a more stable P450 variant yielded more folded mutants when subjected to random mutation than its marginally stable counterpart¹⁸. More interesting was the fact that extra stability increased the number of functionally improved mutants much more than it increased the number of mutants that retain parental function. This suggests that mutations that improve activity are more destabilizing than average mutations.

The apparent trade-off between stability and function has been a widely debated topic with the current consensus being that while individual mutations may result in a tradeoff¹⁸, there is no *inherent* incompatibility between high stability and function^{20; 21; 22; 23}. Some of the confusion on the subject can be explained by the fact that the effects of random mutations on stability in most proteins (excluding viral proteins which are poorly packed in some cases) forms an overlay of two Gaussian distributions that is very similar over a range of sizes (50–330 amino acids) and folds, shown by Tawfik and coworkers⁷. These two distributions can be explained by dividing the mutations by their location in the proteins' structure. Calculating the accessible surface area that, based on the 3D structure, indicate to what extent an amino acid residue is exposed to the solvent and applying a cutoff that separates the mutations into those that occur on the surface or core of the protein produces two Gaussian distributions as shown in Figure 3.1(b).

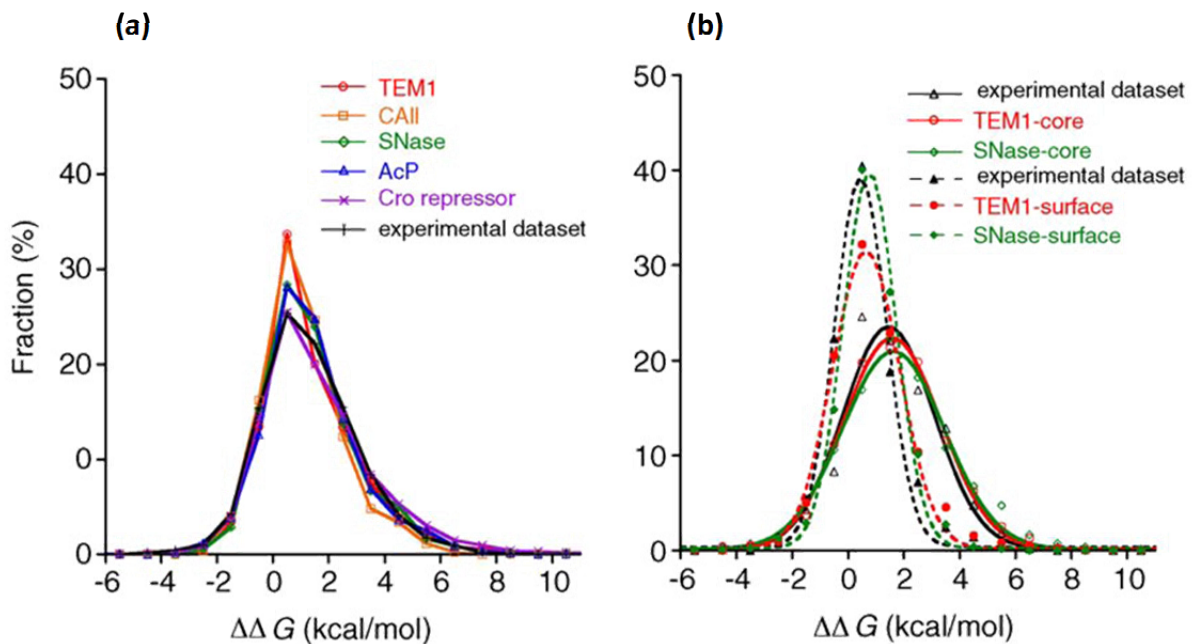


Figure 3.1. (a) $\Delta\Delta G$ values predicted by FoldX for all possible mutations in many proteins (shown are few characteristic examples, and the experimentally measured $\Delta\Delta G$ values for 1285 mutations all give similar asymmetric distributions with larger destabilizing shoulders ($\Delta\Delta G > 0$). (b) Separated $\Delta\Delta G$ distributions of core and surface residues. Residues were divided according to their accessible surface area values, and the $\Delta\Delta G$ values for all possible mutations were arranged in histograms and fitted to a single Gaussian.^{3; 7}

Dividing residues in this way is akin to treating a protein as composed of a hydrophobic core and a hydrophilic surface (an “oil droplet in water”)²⁴. The core plays a key role in protein folding and stability, and therefore core mutations are more deleterious than surface mutations²⁵. This treatment is supported by the fact that, in general, mutations are less destabilizing in smaller proteins, whose ratio of surface to core residues is larger⁷. In fact, Tawfik showed that the only types of proteins that do not show this dual distribution are proteins with loosely packed cores, such as viral proteins, where on average residues make relatively few contacts. In these poorly

packed enzymes, mutations are predicted to exhibit smaller destabilizing effects, especially in the core²⁶. This represents an entirely different method of increasing mutational robustness. Instead of extra stability to absorb destabilizing mutations, loosely packed proteins lower the destabilizing effects of the mutations themselves by minimizing the number interresidue contacts that can be disrupted by mutation²⁶.

Mutations altering protein function often occur in or around the protein's active site, which is usually buried in the core of the protein. These mutations therefore have a higher probability of disrupting the core packing and destabilizing the protein²⁷. However, as mentioned before, this is not inherently true of all mutations affecting function, as mutations that enhance stability while retaining or even increasing activity are not difficult to find²⁸.

Proteins with high stability are useful not only for their ability to accept more mutations, but for their ability to accept highly destabilizing mutations. Besenmatter and coworkers showed that thermostable proteins allow for more variation at critical residues (positions showing significantly restricted amino acid variation), and in fact have fewer critical residues compared to their mesostable homologs¹⁹. These critical residues are often in the interior of the protein, such as in a hydrophobic pocket where nonhydrophobic residues could be highly destabilizing. Mutations at these kinds of locations are most likely what lead to the formation of active sites in many proteins. Active site organization is often inherently unfavorable, with polar or charged functional residues embedded in hydrophobic clefts, sometimes with proximal like charges. In

addition, key catalytic residues often possess unfavorable backbone angles²⁷. Proteins with extra stability buffers have a higher chance of maintaining their structural scaffolds while incorporating mutations in these locations.

New function often arises from altering the protein's substrate specificity, typically by increasing the affinity and rates for weak promiscuous substrates²⁷. Mutations that cause these changes usually occur in the periphery of the active site, within the second or third shell of residues that surround the catalytic residues²⁷. Often, new function mutations will affect conformations of residues and loops distinct from the core catalytic machinery, such as in the substrate channel or loops surrounding the active site²⁹. These locations still have a high probability of being in the packed core of the protein and therefore being destabilizing when mutated. As such, new-function mutations often occur alongside stabilizing compensatory mutations²⁷, with the stabilizing mutations necessary to restore the stability threshold lost due to the new function mutation³⁰.

Excess stability helps in allowing proteins to accept potentially destabilizing new function mutations, but it could also make it more difficult for these types of mutations to appear. A large excess in stability could hinder evolution by rigidifying the protein³, a common result of increasing thermostability. High flexibility can allow multiple conformers to be populated, each one binding a different, unrelated ligand and fixing that conformation when the ligand is present, resulting in specificity for multiple substrates³¹. Therefore, enzymes need a balance between rigidity for stability and

flexibility for activity^{32; 33}. Tawfik suggests that protein flexibility is especially important in evolving new function, as it allows for broad substrate specificity and promiscuous functions. A simple example of flexibility correlating with substrate specificity is that of the human P450s. Skopalik measured the flexibility of three human P450s and found that CYP3A4, the most promiscuous CYP known, has the highest flexibility, while CYP2A6, with a narrow substrate range, is the most rigid, and CYP2C9 is intermediate in both properties³⁴.

Substrate promiscuity is especially important for evolving new-function using directed evolution, as it is difficult to produce entirely new activity without at least a small amount present; the probability of finding mutations that result in completely novel function is prohibitively low³⁵. Most directed evolution projects targeted at increasing activity instead rely on increasing a weak promiscuous activity, which is often straightforward. Multiple directed evolution experiments can be done in parallel, starting with the same parental enzyme but targeting different substrates each time to produce enzyme superfamilies³⁶. In fact, it has been speculated that early enzyme ancestors exhibited broad substrate specificity so they could maximize their catalytic versatility while dealing with limited enzyme resources³⁷. It is thought then, that gene duplication and selective pressure led to specialized enzymes³⁸.

A promiscuous, highly stable enzyme would be an ideal starting point for directed evolution of new-function. Promiscuity means activity on multiple substrates, and activity on each could be selected for and optimized to create an enzyme highly

active and specific for that substrate. High stability would allow incorporation of more mutations, especially in the core of the protein.

B. Identification of a Cytochrome P450 with Broad Substrate Specificity

The cytochrome P450 superfamily database currently contains over 11,000 members in nearly 1000 families³⁹. These enzymes catalyze the insertion of oxygen into unactivated C-H bonds using a protoporphyrin iron heme cofactor with electrons supplied from NADPH via a reductase domain⁴⁰. While the catalytic mechanism is same for the different P450s, they are active on a wide variety of substrates, including steroids, bile acids, eicosanoids, fat-soluble vitamins, and xenobiotics for the human P450s alone⁴¹. The specificities P450s run the gamut from very specific to highly promiscuous, with only 15 human P450s responsible for the breakdown of xenobiotics, including one enzyme, CYP3A4, which is alone responsible for 30% to 50% of drug metabolism^{42; 43; 44; 45}.

Due to their promiscuous nature, human P450s, especially CYP3A4, seem like an ideal starting enzyme for directed evolution. However, human P450s are difficult to express heterologously due to their multicomponent nature (separate reductase), which limits their potential reaction rates. They are also insoluble, limiting their expression to the membrane surface. Although there are examples of successful heterologous human P450 expression for the production of authentic human metabolites, these systems are costly and have low productivities⁴⁶.

Because of their soluble nature and high expression levels in bacterial hosts, several other P450s are preferred for biotechnological applications. P450_{BM3} (CYP102A1) is a class II P450 from *Bacillus megaterium*, meaning it requires a FAD/FMN-containing diflavin reductase for catalysis⁴⁷, and can be expressed in *E. coli* at up to 12% of the dry cell mass⁴⁸. The property that sets P450_{BM3} apart from the other soluble P450s is that its reductase domain is directly fused to its heme domain via a short linker, making it a single polypeptide. This could account for its extremely fast reaction rates (thousands per minute) on its preferred fatty acid substrate⁴⁹.

P450_{BM3} shares 52% homology and 22% identity with CYP3A4⁵⁰ and both enzymes have highly flexible active sites, a property that allows for the binding of many substrates with different structures⁵¹. Past members of the Arnold lab have shown that P450_{BM3} variants have activity on structurally diverse compounds such as buspirone⁵², propranolol⁵³, and compactin⁵⁴. To identify highly promiscuous P450_{BM3} variants, we screened 120 diverse P450_{BM3} variants (Table 3.6 in Supplemental Material) for activity on the four structurally diverse compounds shown in Figure 3.2.

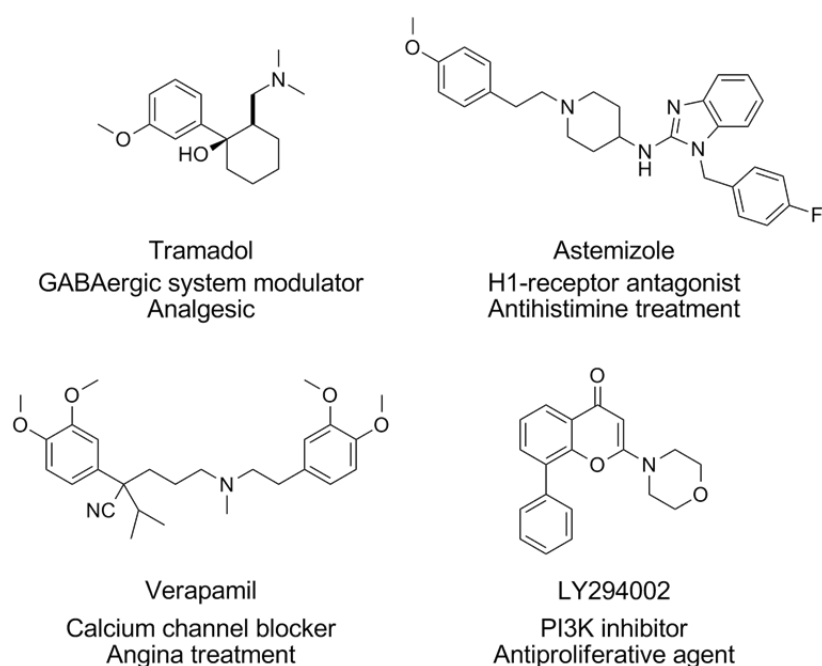


Figure 3.2. The structures, names, biological activities, and pharmacological uses of four structurally diverse compounds used to identify promiscuous P450_{BM3} variants.

Screening revealed that enzyme 9-10A and its close variants (within three mutations) had detectable activity on several of the compounds. Of 45 9-10A variants, 13 had activity on at least three compounds, as seen in Figure 3.3. In addition, 20 showed activity on at least two compounds, and nearly all of them had activity on at least one compound. These variants have different selectivity for the four compounds, and from HPLC/LCMS analysis, different regioselectivity on each compound. These data suggest that 9-10A has the broad substrate specificity necessary to make it an ideal parent for directed evolution studies aimed at producing enzymes with activity on a wide variety of compounds.

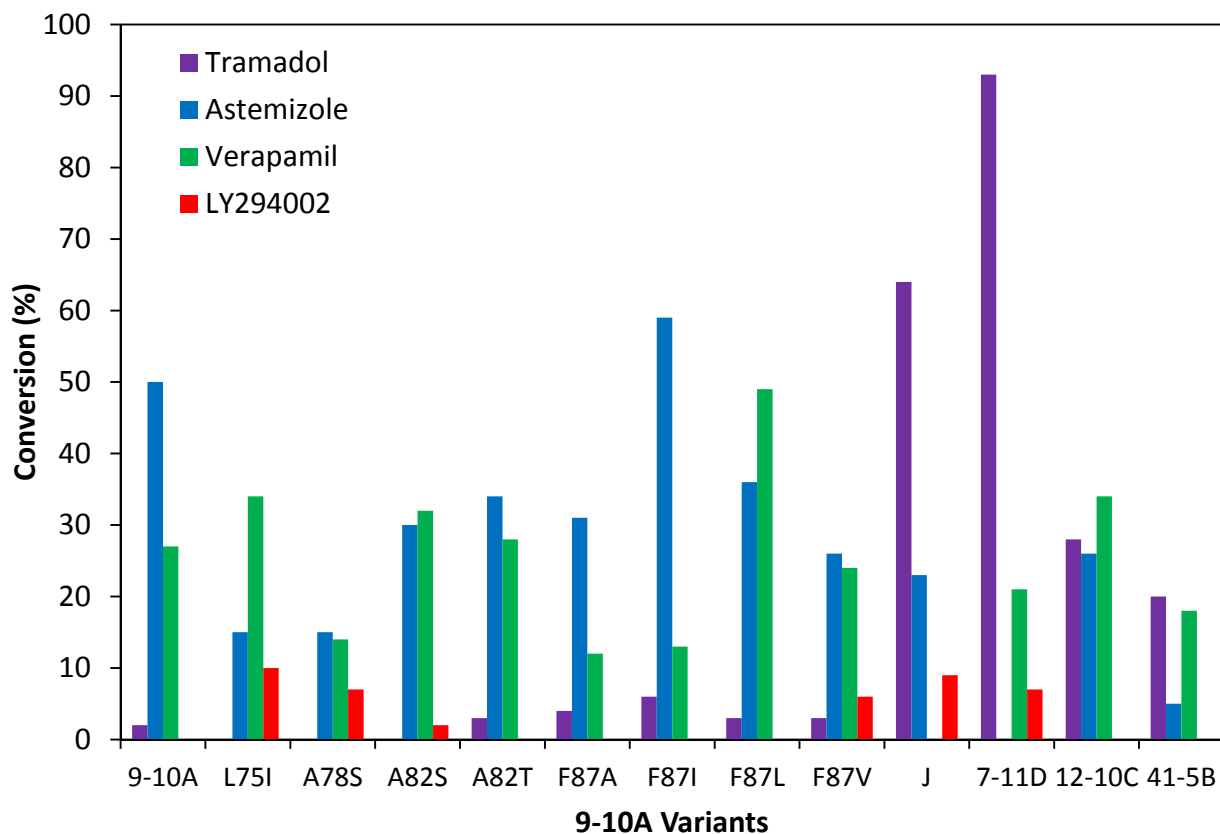


Figure 3.3. Conversions of cytochrome P450_{BM3} 9-10A variants on the four structurally diverse compounds shown in Figure 3.2. See Table 3.6 in Supplemental Material for sequences.

C. Thermostabilization of 9-10A

Enzyme 9-10A (see Table 3.1 for sequences) is part of the evolutionary lineage leading from wild type (WT) P450_{BM3} to the propane monooxygenase P450_{PMO}⁵⁵. It was selected for its activity on propane and differs from WT at 13 positions⁵⁶. While its broad substrate specificity makes 9-10A an attractive parent for directed evolution studies on a wide range of substrates, it suffers from a low thermostability. 9-10A has a T₅₀ (temperature at which half of the enzyme is inactivated after a ten-minute incubation) of 45 °C, which is 10 °C below that of WT and just above the threshold for proper folding and function under cellular conditions, estimated to be 43 °C for P450s⁵⁷.

This low stability makes 9-10A a poor evolution parent, as the number of mutations it can accept is limited.

This lack of stability and thus mutational robustness proved to be a problem later on in the P450_{PMO} lineage, when after just one more round of evolution to produce enzyme 35-E11, further improvements in activity could not be found⁵⁸. Analysis of the lineage's stability led to the conclusion that 35-E11's stability was too low to withstand further mutation. Stabilization was needed before activity improvements could be found⁵⁵. Analysis of the lineage's thermostabilities revealed that while selecting for increased catalytic activity on propane, the thermostability of the enzyme steadily drifted downward. This is of course a consequence of the fact that most mutations are destabilizing.

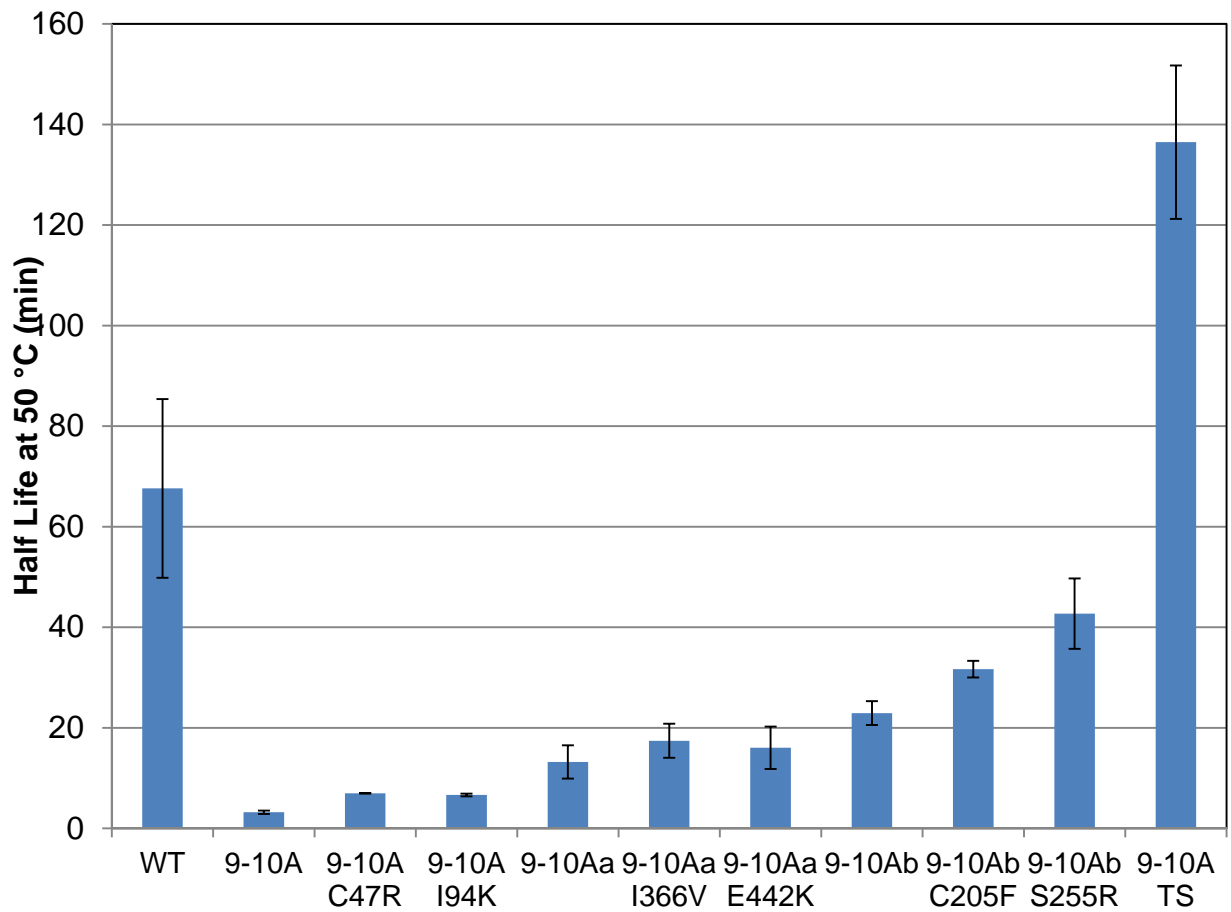
To create a more ideal parent for directed evolution, we set out to thermostabilize 9-10A while simultaneously maintaining its broad substrate specificity. To do this, we looked at mutations already known to be stabilizing in other P450_{BM3} variants. As the proteins are all within 20 mutations of each, their structures should be very similar and we would expect the stabilizing effects of these mutations to translate well between the variants²³.

9-10A was chosen for its increased activity on propane, but there is a significant drop in stability compared to its parent (a decrease in T_{50} of nearly 5 °C), which differs by three mutations. Reversion of two of these three mutations (C47R and I94K) was already known to stabilize a variant further along in the P450_{PMO} lineage. These

mutations were reverted in 9-10A and each resulted in an increase in stability. Their combination further increased stability and resulted in enzyme 9-10Aa as seen in Figure 3.4. We next examined eight mutations that had been discovered while stabilizing a P450_{BM3} peroxygenase variant, enzyme 21B3⁵⁹. Of the eight, three appeared to be alleviating local structural perturbations caused by earlier mutations and so were most likely context specific. These were avoided. The remaining five mutations (L52I, L324I, V340M, I366V, and E442K) were examined in 35-E11, and two of these were found to be stabilizing, resulting in the stability increase seen in Figure 3.4. We therefore examined all five mutations in 9-10Aa and found that two were stabilizing and their combination even more stabilizing, resulting in enzyme 9-10Ab (Figure 3.4). However, these two stabilizing mutations were not the same two found to be the best combination in 35-E11. Finally, an additional two reversions were found to be stabilizing in another variant further along in the P450_{PMO} lineage. See Table 3.1 for full sequences. These sites were reverted in 9-10Ab individually and in combination to produce enzyme 9-10ATS. 9-10ATS is highly stable, with a half-life at 50 °C two orders of magnitude greater than that of 9-10A, and over twice as long as that of WT, as seen in Figure 3.4.

Table 3.1. Sequences of thermostable P450_{BM3} 9-10A variants

Enzyme	Sequence
9-10A	WT R47C V78A K94I P142S T175I A184V F205C S226R H236Q E252G R255S A290V A295T L353V
9-10Aa	9-10A C47R I94K
9-10Ab	9-10A C47R I94K I366V E442K
9-10ATS	9-10A C47R I94K I366V E442K C205F S255R

**Figure 3.4. Half-lives at 50 °C of 9-10A and thermostabilized variants.**

Stabilizing 9-10A makes it more robust to mutation, but it is important that this does not diminish its catalytic promiscuity. To verify that 9-10ATS still retains broad substrate specificity, we screened 9-10A, 9-10ATS and all of their intermediates on several diverse substrates, including astemizole, verapamil (Figure 3.2), and another compound on which 9-10A has activity: dimethyl ether. As seen in Figure 3.5, all of the variants maintain activity on each of the compounds.

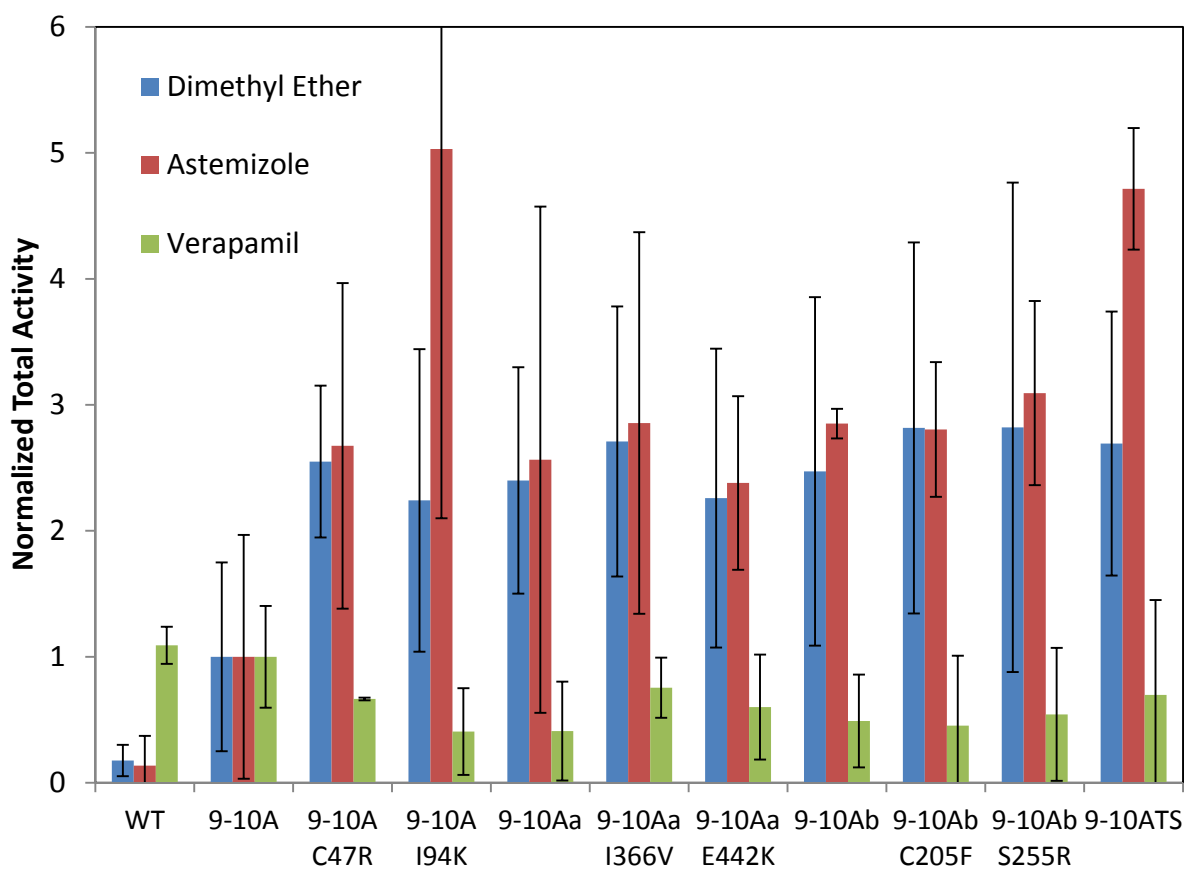


Figure 3.5. Total activities on dimethyl ether, astemizole, and verapamil, normalized to 9-10A's activity on those substrates.

9-10ATS is a highly thermostable protein with broad substrate specificity. It likely has a good balance of rigidity for stability and flexibility for activity. As such, it makes an ideal parent for directed evolution toward increased activity on a wide range of compounds.

D. 9-10ATS as a Parent in Directed Evolution

This section presents work published in *ChemBiochem*, 2010. **11**(18): p. 2502-2505 by Lewis, J. C., Mantovani, S. M., Fu, Y., Snow, C. D., Komor, R. S., Wong, C. H., & Arnold, F. H.⁶⁰.

Our lab previously reported the use of P450_{BM3} variants in the synthesis of difficult-to-synthesize monosaccharide derivatives⁶¹. However, these enzymes' activity was limited to methoxymethyl (MOM)-protected pentoses. Activity on MOM-protected hexoses and other bulky compounds (Figure 3.6) could not be obtained even with from random mutant libraries generated by error-prone PCR.

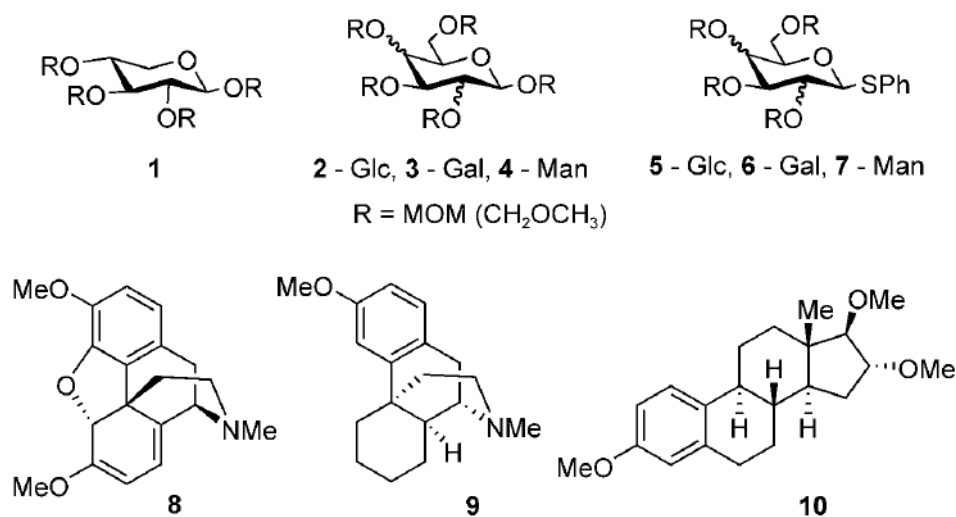


Figure 3.6. Structures of compounds utilized in enzyme library design and screening: MOM-protected xylose, hexoses, and thioglycosides (1-7); alkaloids thebaine and dextromethorphan (8 and 9); trimethyl estriol (10).

site, it was decided to combinatorially replace active site residues with alanine. This extensive mutagenesis requires a thermostable parental enzyme as a starting point. Enzyme 9-10ATS F87V was chosen as the unstabilized version, 9-10A F87V, shows activity on compound 1 in Figure 3.6⁶¹. The F87V mutation has been shown to increase P450_{BM3} variants activity on a number of aromatic compounds^{62; 63; 64}. 9-10ATS F87A showed weak activity on the compounds in Figure 3.6. Structures of the most stable conformers of these compounds were generated using Omega⁶⁵ and docked into an active site model of 9-10ATS F87V. This led to the identification of eight residues for replacement with alanine: K69, L75, M177, L181, T260, I263, T268, and L437. The 2⁸ (256) member library was generated using splicing by overlap overlap extension PCR (see Section K of Materials and Methods). CO binding analysis (see Section M of Materials and Methods) revealed that 65% of the enzymes were properly folded and

sequencing indicated unbiased incorporation of alanine at all of the desired sites with an average alanine substitution of 3.9 per sequence.

The library was screened against the compounds in Figure 3.6 by detecting formaldehyde released during P450-catalyzed removal of methyl or MOM groups (Section M of Materials and Methods). Consistent with the hypothesis that an expanded active site would provide an advantage on bulky substrates, library variants showed increased activity on the larger substrates (Table 3.2) but not on the smaller pentose substrate (compound 1 in Figure 3.6).

Table 3.2. Sequence-activity relationship for alanine substitution. The fold improvement is the ratio of A_{550} measurement for reaction of each variant to that of the parent (9-10ATS F87A). Ratio for substrate with maximum improvement shown. The improvement in the reaction with steroid compound 10 was identified by visual inspection as substrate insolubility complicated plate reader measurement.

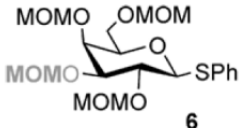
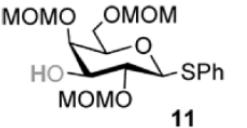
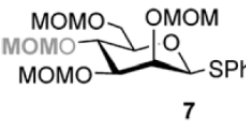
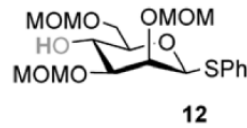
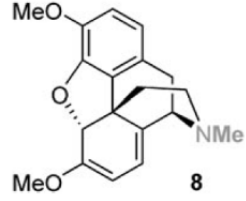
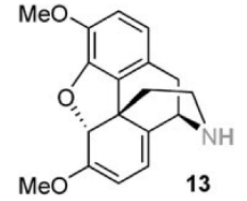
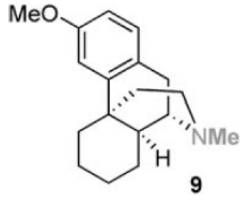
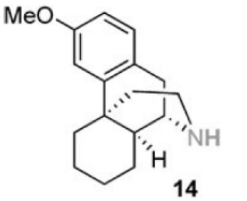
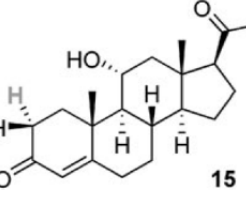
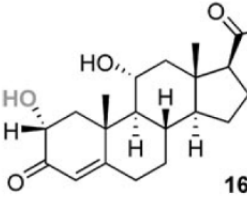
Substrate	Variant	Alanine Substitution (+) at residue								Fold Improvement
		69	75	177	181	260	263	268	437	
Thioglycosides (compounds 5-7 in Figure 3.6)	2A1	-	+	-	+	-	-	-	+	4.4
	4H9	-	-	-	+	+	-	-	+	4.1
	8C7	-	+	-	+	-	-	-	-	7.9
Alkaloids (compounds 8-9 in Figure 3.6)	4H5	-	+	+	+	-	-	-	-	2.7
	4H9	-	-	-	+	+	-	-	+	3.9
	7A1	-	+	+	+	+	-	-	-	2.8
	8C7	-	+	-	+	-	-	-	-	2.7
	8F11	-	-	-	-	-	-	-	+	3.3
Steroid (10)	8F11	-	-	-	-	-	-	-	+	N.A.

Steroid hydroxylation is a particularly valuable reaction due to the biological activity of these compounds and their common occurrences as metabolites⁶⁶, often

using P450 enzymes to produce said metabolites⁶⁷. However, most of these reactions require the use of whole-cell biocatalysts⁶⁶ or multicomponent enzyme systems⁶⁸. Variant 8F11 exhibited activity with the steroid derivative trimethyl estriol, potentially providing a convenient single enzyme platform for steroid hydroxylation. A random mutagenesis library of variants of 8F11 was generated using error-prone PCR with screening for improved demethylation of trimethyl estriol leading to enzyme F1, with four mutations and 1.6-fold improved activity over 8F11. F1 also showed activity with additional steroids, including 11- α -hydroxyprogesterone and testosterone acetate.

To demonstrate the synthesis utility of the enzymes obtained, preparative scale bioconversions were conducted with reaction conditions previously developed in our laboratory⁶¹. Yields and selectivities are shown in Table 3.3. The site-selective deprotection of MOM-protected hexoses (entries 1 and 2 in Table 3.3) expands the utility of previously reported chemoenzymatic monosaccharide elaboration⁶¹. Demethylation and N-substitution of opiate alkaloids (entries 3 and 4 in Table 3.3) is commonly used to vary the properties of these compounds. The importance of steroid hydroxylation was stated above and the high selectivity of variant F1 (entry 5 in Table 3.3) produced from a single round of random mutagenesis suggests that additional catalysts with high selectivities on other substrates can be developed by further directed evolution.

Table 3.3. Substrate scope and reaction selectivity. Conversion of starting material determined by HPLC or GC analysis of crude reaction mixture. Selectivity, percentage of desired product relative to additional products, determined by HPLC or GC analysis of crude reaction extracts. Yield determined by isolated yield of the pure product.

Substrate	P450 _{BM3} Variant	Product	Conversion (%)	Selectivity (%)	Yield (%)
 6	2A1	 11	80	90	75
 7	8C7	 12	93	80	70
 8	8C7	 13	88	72	60
 9	4H5	 14	54	98	50
 15	F1	 16	28	82	20

Together these results demonstrate the utility of combinatorial alanine substitution in a thermostable parent for generation of variants with activity against bulky, synthetically useful substrates. We have demonstrated that the resulting

enzymes have novel activities that can be further optimized by directed evolution. Monosaccharides, alkaloids, and steroids were all viable substrates despite their large size.

E. 9-10ATS Structure Determination

After verifying that 9-10ATS is an ideal parent for directed evolution, we wanted to obtain a three-dimensional crystal structure of the enzyme in order to examine the structural effects of the mutations that resulted in increased stability and expanded substrate scope. The Molecular Observatory for Structural Molecular Biology is an effective and convenient resource available to Caltech students and faculty. It consists of the on-campus Macromolecular Crystallography Facility in the Beckman Institute, and a high intensity, automated synchrotron radiation beam line at the Stanford Synchrotron Radiation Laboratory (beam line 12.2).

Both 9-10A and 9-10ATS were purified as described in Section O of Materials and Methods, and sent to the high throughput crystallographic screening facility. Using the facility, we screened both proteins against 480 common crystallization conditions from commercially available screens listed in Section Q of Materials and Methods. While 9-10A did not form crystals in any of the conditions, after 30 days, poorly formed crystals of 9-10ATS appeared in three related conditions listed in Table 3.4. Still using the high throughput crystallographic screening facility, we combinatorially screened conditions varying the concentration of ammonium sulfate, concentration of lithium

sulfate, identity of the buffering agent, and pH. This screening identified 2.0 M ammonium sulfate, 0.2 M lithium sulfate, and 0.1 M tris, pH 7.0-7.5 as yielding the best-formed crystals. These crystals formed after 30 days and were harvested and sent to the beam line at Stanford. These crystals diffracted to 5-7 Å resolution, implying that further condition optimization was needed.

Table 3.4. Composition of the three conditions that resulted in crystals of 9-10ATS during high-throughput screening.

Buffer	Salt	Co-precipitate
0.1 M tris buffer, pH 8.5	2 M ammonium sulfate	none
0.1 M tris buffer, pH 7.0	2 M ammonium sulfate	0.2 M lithium sulfate
0.1 M HEPES buffer, pH 7.5	2 M ammonium sulfate	none

We used shards of harvested crystals to seed new nucleation sites and increase the speed of crystal formation. This technique reduced the time for crystal formation from over 30 days to 3-5 days for the same conditions with seeding. When crystals from the seeded conditions were sent to the beam line at Stanford for screening, they diffracted to 4 Å, a marked improvement but still not low enough for structure determination. To further improve crystallization conditions, we used a commercial additive screen to test the addition of 96 compounds, one at a time, to the current conditions. The addition of 3% 6-aminohexanoic acid resulted in improved crystal form, and crystals from these conditions were sent to the beam line at Stanford for screening. These crystals diffracted to 3 Å, suitable for structure determination.

The data were processed and a preliminary structure was generated using the molecular replacement method with the structure of WT as the template. The dimensions of the unit cell, with one axis much longer than the other two (Table 3.5), complicated solving the initial structure, but with extra time MOLREP⁶⁹ converged to a solution. The generated structure revealed the reason for the odd space group: the unit cell contained six monomers. This extended the time needed to refine the structure, but was useful in that the six chains could be compared to each other in areas where one lacked resolution. After ten rounds of manual refinement with COOT⁷⁰ and automated refinement with CCP4^{71; 72} the structure was solved with statistics listed in Table 3.5.

Table 3.5. Data collection and refinement statistics for 9-10ATS structure. All data sets were collected from single crystals. Highest-resolution shell is shown in parentheses.

PDB Accession Number	T.B.D.
Data Collection	
Space group	P 41 21 2
Wavelength (Å)	1.033
Resolution (Å)	39.7 – 3.0 (3.2 – 3.0)
R_{merge}	5.1(38.2)
$I / \sigma I$	25.5(5.2)
Completeness (%)	99.9(100.0)
Redundancy	6.6(6.8)
Cell Dimensions	
a, b, c (Å)	162.85, 162.85, 361.43
α, β, γ (°)	90.0, 90.0, 90.0
Refinement	
Resolution (Å)	39.5 – 3.0
Number of reflections	92760
$R_{\text{work}} / R_{\text{free}}$	0.19/0.25
Number of Atoms	
Protein	21188
Ligand/ion (Heme)	258
Ligand/ion (Sulfate)	80
Water	11
B-factors	
Protein	55.2
Ligand/ion (Heme)	45.0
Ligand/ion (Sulfate)	116
Water	36.5
R. M. S. Deviations	
Bond lengths (Å)	0.017
Bond angles (°)	1.77

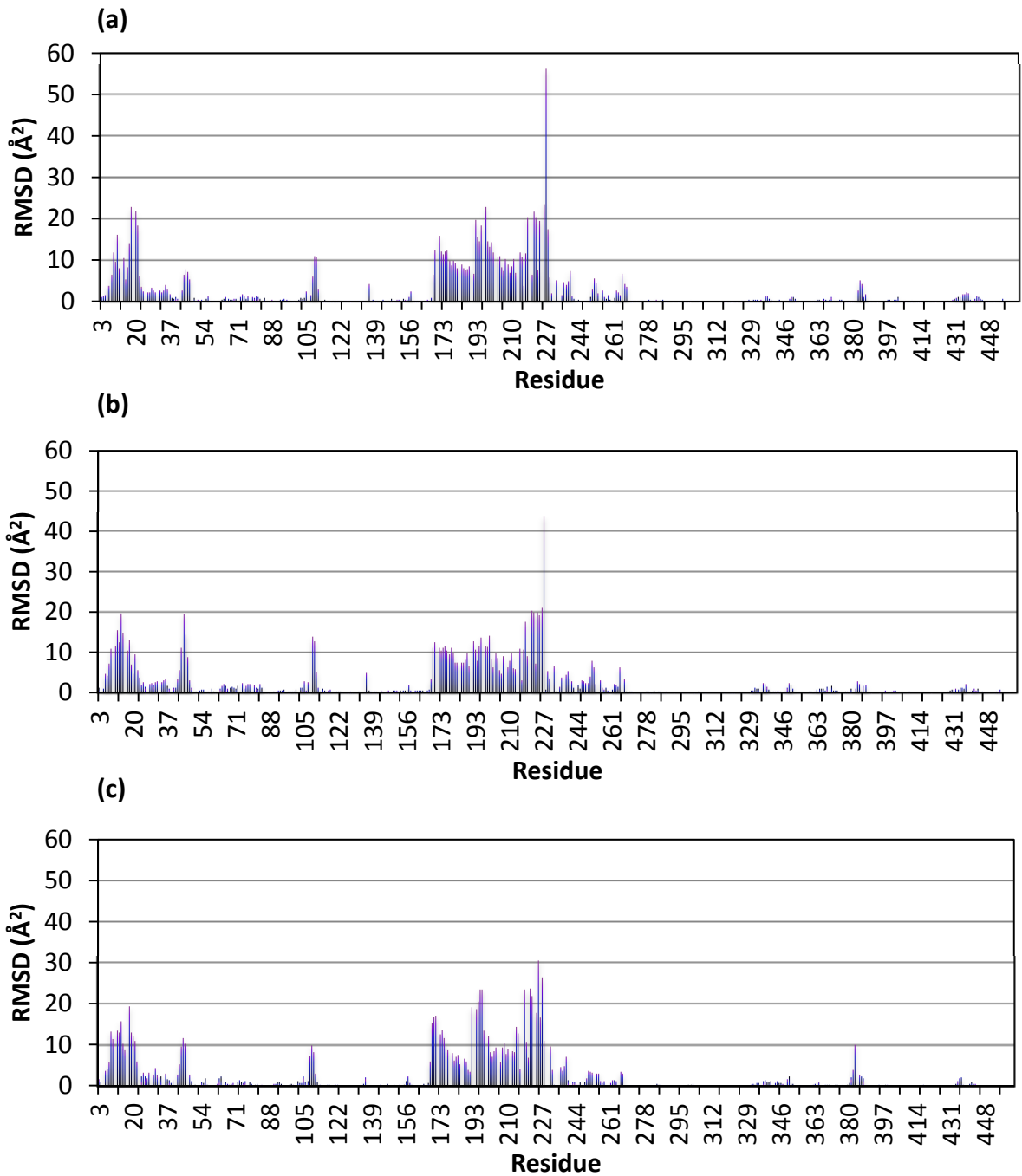
Visual inspection of the structure fit to the density map showed good resolution of most side chains. Less than 9% of the residues needed to be stubbed (side chain atoms removed) due to a lack of electron density. Density fit analysis shows that chains A and C fit the data best in addition to needing fewer stubbed residues. As such, structural analysis was performed on chains A and C.

F. 9-10ATS Structural Analysis

For comparison, WT structures in the open (PDB ID 2IJ2) and closed (PDB ID 1JPZ) conformation were used as well as the structure of 139-3 (PDB ID 3CBD), an evolutionary intermediate from the P450_{PMO} lineage between WT and 9-10A. 139-3 differs from WT and 9-10A at eleven and eight residues, respectively. The structure of 139-3 is in the closed conformation, with the same molecule, N-palmitoylglycine, bound in the active site as in the closed structure of WT.

Aligning the structures and calculating their pairwise RMSD at each residue produces the graphs in Figure 3.7. Comparison of the WT closed and open conformations (Figure 3.7 (a)) reveals that the major deviations occur from the N-terminus of the protein to residue 49, and at residues 168-232. The deviations at residues 168-232 are large in magnitude and encompass the F and G helices. These helices form the roof of the “hinge” which closes upon substrate binding and are among the most flexible parts of the proteins. Deviations in the same regions in Figure 3.7 (b) and (c) show that both the 139-3 and 9-10ATS structures are in the closed state as well. As stated before, the closed conformation WT and 139-3 structures both have N-palmitoylglycine bound in the active site, and inspection of the electron density map for the 9-10ATS structures shows enough density above the heme for a substrate to be present. The density is not clear enough to identify what is bound, but it appears to have a long carbon chain. It could be that 6-aminohexanoic acid, the additive compound that resulted in crystals that diffracted well enough to generate a structure, bound to

the heme and fixed the protein in a closed conformation, improving packing and thus diffraction.



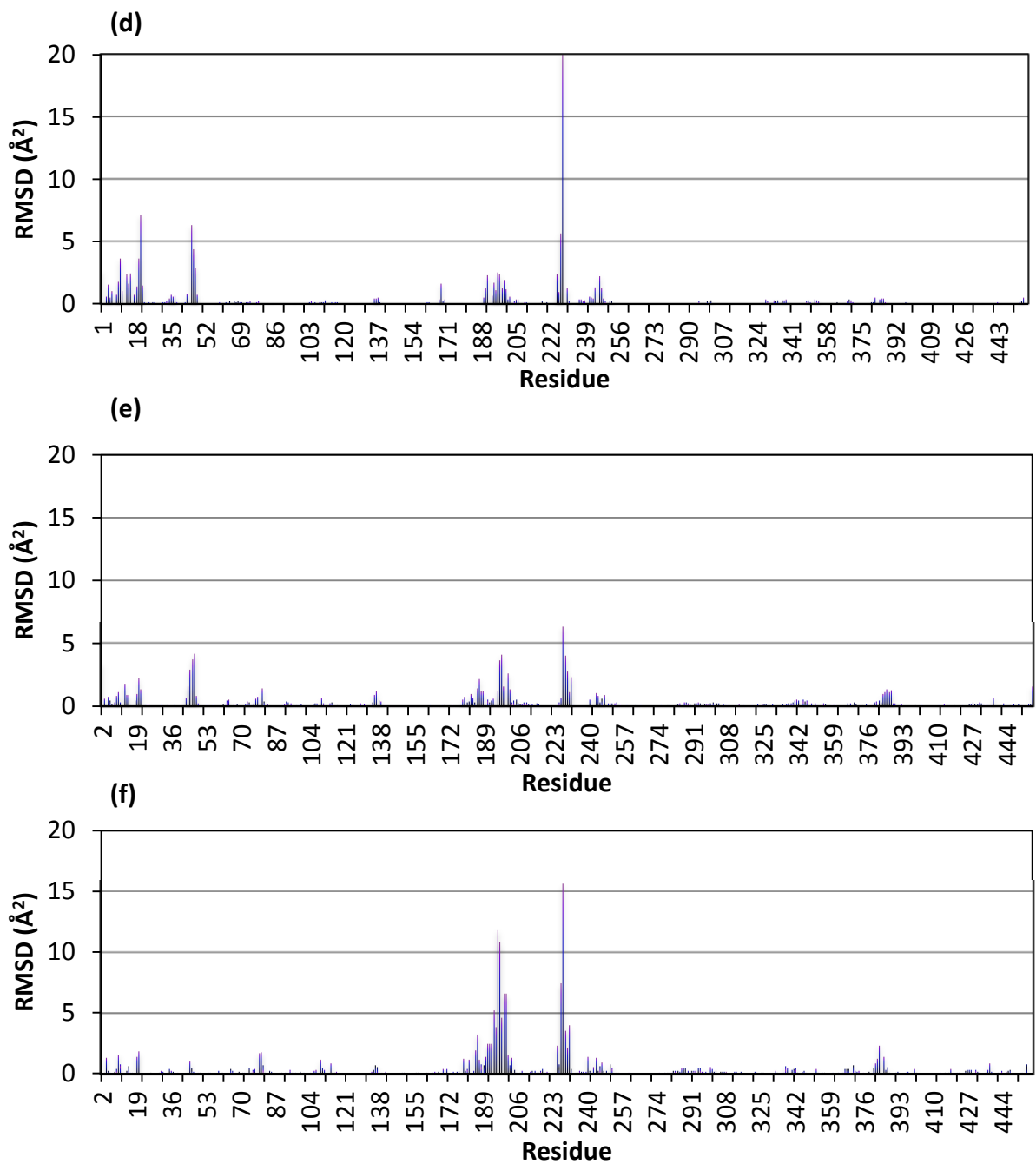


Figure 3.7. Pairwise, residue by residue RMSD values (Å²) for open conformation WT vs. (a) closed conformation WT, (b) 139-3, (c) 9-10ATS, closed conformation WT vs. (d) 139-3, (e) 9-10ATS, and (f) 139-3 vs. 9-10ATS.

Pairwise comparison of 9-10ATS to the closed conformation WT and 139-3 (Figure 3.7 (e) and (f), respectively) show much smaller RMSD values compared to 9-10ATS vs. the open conformation WT. In fact, the RMSD values for comparing the entire structure is 1.3 \AA^2 for 9-10ATS vs. the open conformation WT and $\sim 0.5 \text{ \AA}^2$ for 9-10ATS vs. the closed conformation WT or 139-3, further supporting the notion that the structure of 9-10ATS is in the closed conformation.

The graphs comparing 9-10ATS to the closed conformation structures (Figure 3.7 (e) and (f)) have peaks in the same locations, but with higher magnitude in the graph of 9-10ATS vs. 139-3. This is surprising, as 139-3 is closer in sequence to WT, although several mutations accumulated in 139-3 were reverted in 9-10ATS to regain stability. There are large structural deviations near many, but not all, of the mutations, both those responsible for increased stability and those presumably responsible for broad substrate specificity.

Most notable are the regions away from mutations that show significant deviations. The F and G helices both have mutations in their middles, but show considerable structural deviations along their length and especially where they connect to adjacent secondary structures. As stated before, the F and G helices are highly flexible and important for substrate binding. Changes in this part of the protein are likely to affect substrate binding and could be responsible for 9-10ATS's wide substrate range. The mutations T175I and A185V are present in 139-3, whose F and G helices closely align with those of WT. As such, the changes are most likely due to S226R, which is not

present in 139-3. Position 226 is located at the end of the G helix on the opposite side of where it meets the F helix, but is where the portion of the protein that moves meets the unmoving part.

B-factors are influenced by the quality of the crystal structure, but indicate the degree of mobility, with a higher B-factor indicating more flexibility⁷³. To compare B-factors between different structures, we normalized the C α B-factors with that of the heme to account for differences in resolution, similar to the strategy used by others to compare B-factors from different structures⁷³. Interestingly, the normalized B-factors of residues in the F and G helices are considerably lower in 9-10ATS than in the closed conformation WT or 139-3, as seen in Figure 3.8. A lower B-factor in the F and G helices could mean that the substrate is bound more tightly. Binding substrates tightly could account for increased activity, but more structures with different substrates bound and with no substrate are needed for further insight.

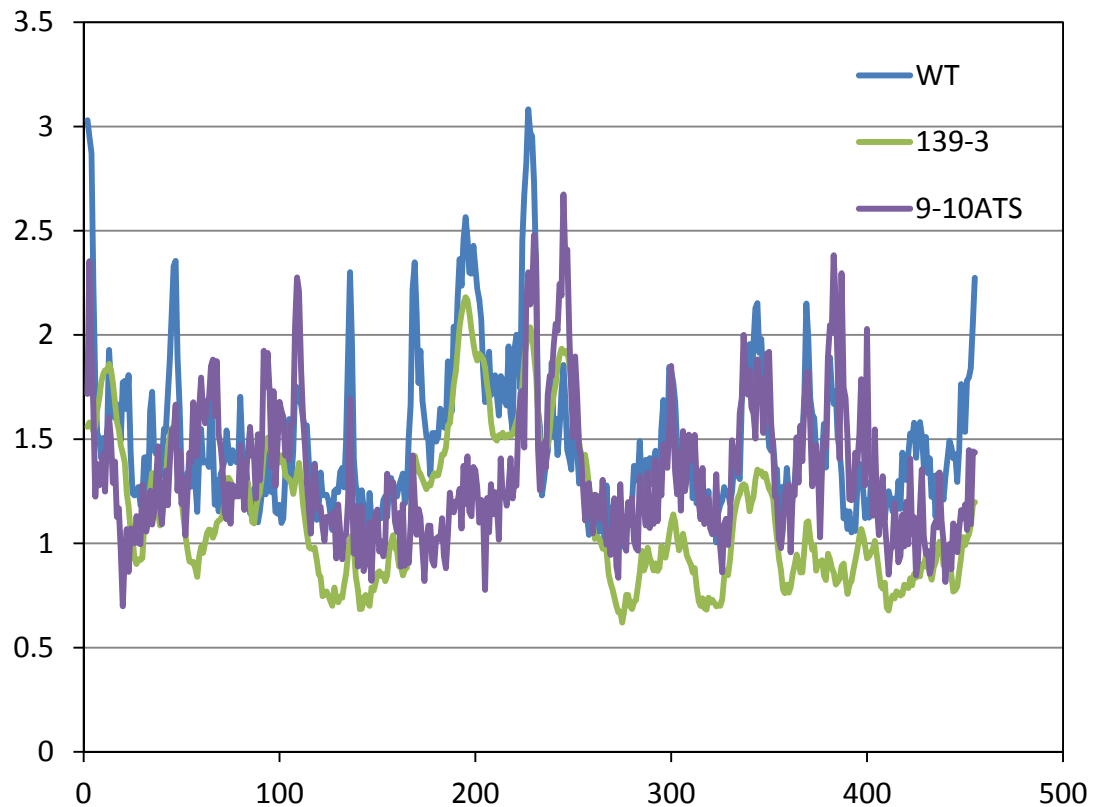


Figure 3.8. Residue by residue B-factors for the closed conformation WT, 139-3, and 9-10ATS structures. B-factors are normalized to that of the heme in each structure.

Compared to the other closed conformation structures, 9-10ATS has higher B-factors in the region below the heme (residues 381-408) on the opposite side from where the substrate binds and catalysis take place. The cysteine that covalently binds the heme is in this region and if it is flexible, it could allow the heme to have different orientations in the active site, accommodating substrates that otherwise would not fit. The mutation V78A is in the substrate channel and substitution of the smaller alanine for valine widens the channel, allowing larger substrates access to the active site. While

139-3 also has the V78A mutation, the C helix is moved farther back in 9-10ATS, as seen in Figure 3.9, widening the substrate channel even more.

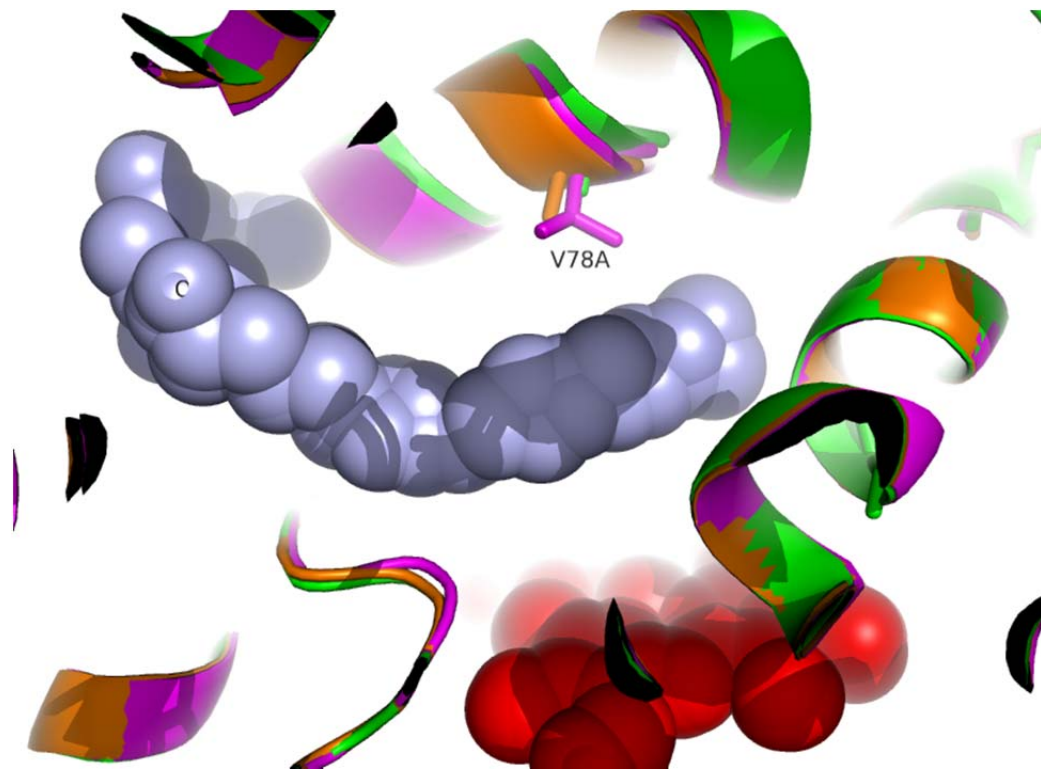


Figure 3.9. Structural alignment of closed conformation WT (magenta), 139-3 (orange), and 9-10ATS (green) showing the V78A mutation with the heme in red and N-palmitoylglycine shown in light blue. Alanine, present in 139-3 and 9-10ATS, creates more room for the substrate than valine. The entire C helix is moved back in 9-10ATS, creating even more space.

The structural cause of the increase in stability from the six stabilizing mutations is clear from inspecting the crystal structure. The mutation I366K is near the carbonyl oxygen of glutamine 55, which could have electrostatic interactions with lysine, but not isoleucine. Position 47 is located at the entrance to the substrate channel and its side chain is in direct contact with N-palmitoylglycine in both the closed conformation WT and 139-3 structures. This would seem a likely mutation for altering the substrate

specificity of the enzyme rather than its stability, but the opposite is true. When an arginine is present at this position, its side chain is within 6 Å of the carbonyl oxygen of glutamine 73, forming electronic interactions that would be disrupted by substitution with a cysteine, which would also repel the electronegative carbonyl oxygen. Position 47 is on the surface of the protein, and so a positively charged arginine could have interactions with the solvent as well.

Several of the surface mutations result in salt bridges with neighboring residues as seen in Figure 3.10 (a), (b), and (c). Lysine 94 forms a salt bridge with glutamate 247, arginine 255 with aspartate 217, and lysine 442 with aspartate 432. Arginine 255 and lysine 442 also have lower normalized B-factors than serine or glutamate at those positions, respectively. A lower B-factor suggests more rigidity which can improve stability.

Another reversion, C205F, is located on the side of the G helix in the midst of a group of hydrophobic residues including phenylalanine, isoleucine, and valine as seen in Figure 3.10 (d). Replacing phenylalanine with cysteine would disrupt the hydrophobic packing. Interestingly, even though all the structures have phenylalanine at this position, the normalized B-factor of the phenylalanine is much lower in the 9-10ATS structure than in that of WT or 139-3. Again, this could be related to tighter substrate binding.

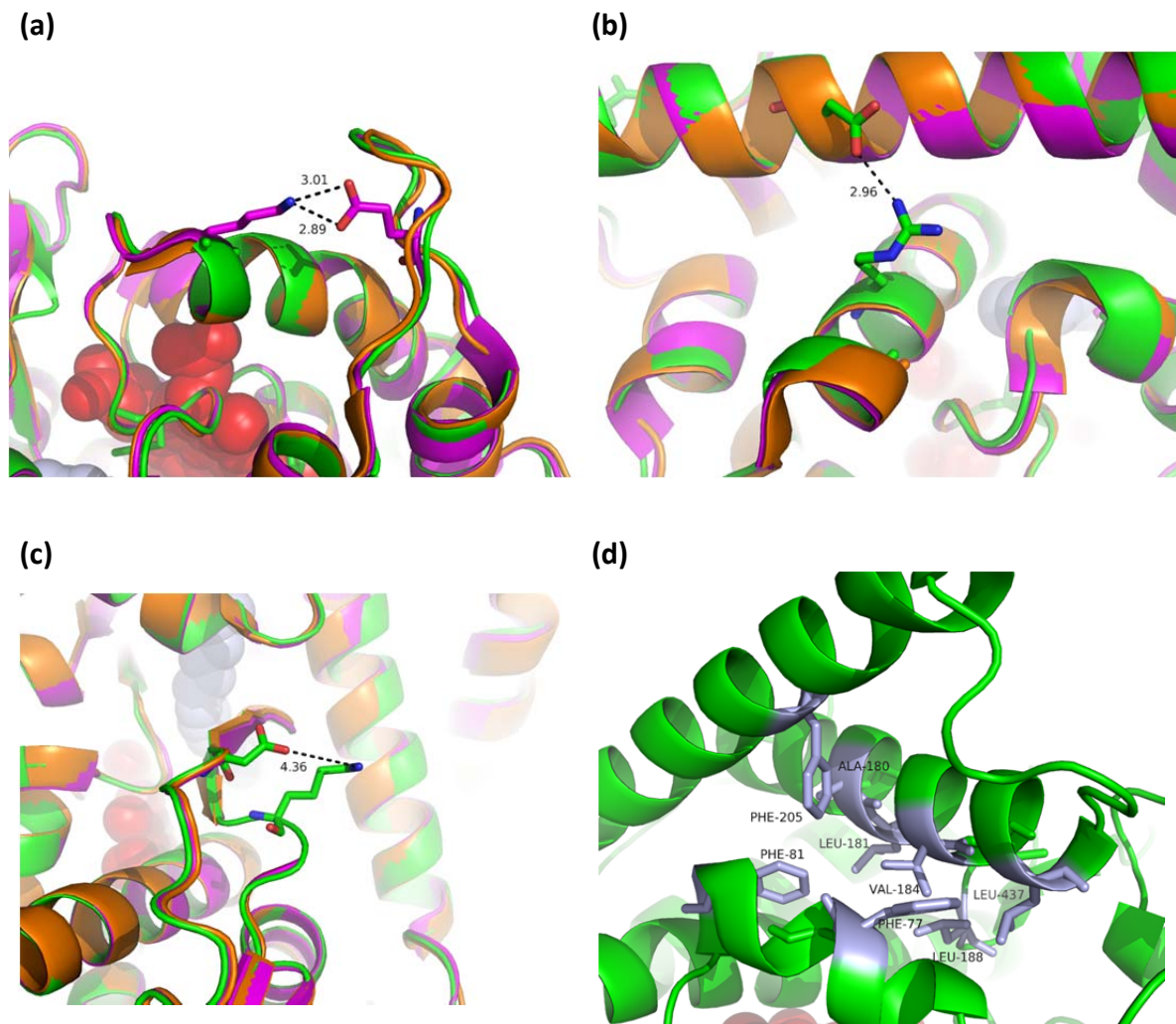


Figure 3.10. Structural alignment of closed conformation WT (magenta), 139-3 (orange), and 9-10ATS (green) with the heme in red. Mutations (a) I94K, (b) S255R, and (c) E442K form salt bridges. Measurements are given in angstroms. (d) Phenylalanine 205 is part of a hydrophobic pocket (light blue).

G. Conclusions

We identified variant 9-10A as having broad substrate specificity after screening a panel of P450_{BM3} variants on structurally diverse compounds. Using putatively stabilizing mutations, we increased the stability of 9-10A to greater than that of WT,

while maintaining its wide substrate scope. 9-10ATS F87A was used as the parent for alanine substitution at eight positions in the active site in order to increase activity on a number of bulky substrates, including protected sugars, alkaloids, and steroids. This produced a number of variants with improved activity, one of which was further improved by random mutagenesis. The three-dimensional structure of 9-10ATS was solved using x-ray crystallography and compared to structures of WT and 139-3, an evolutionary intermediate. The structural basis of the stabilizing effects of the mutations is clearly visible, as are structural changes to the catalytic machinery of the enzyme. The substrate channel is significantly wider in 9-10ATS and the F and G helices (the portions of the protein that move upon substrate binding) have moved. We have shown that 9-10ATS is an ideal parent for directed evolution projects aimed at increasing activity on a number of structurally diverse compounds. Besides giving insight into the structural basis of 9-10ATS's unique properties, its structure will aid in planning future directed mutagenesis studies.

H. Supplemental Material

Table 3.6. Mutants screened for activity on the compounds listed in Figure 3.2. Sequences are listed with the parent sequence in bold followed by the mutations in normal type. Chimeras are written according to fragment composition: 32313233-R1, for example, represents a protein which inherits the first fragment from parent CYP102A3, the second from CYP102A2, the third from CYP102A3, and so on. R1 connotes the presence of the reductase domain from parent A1, indicating that this chimera is a monooxygenase.

Entry	Name	Sequence	Mutations vs. WT	Selection Criteria	Oxidant
1	P450 _{BM3}	CYP102A1	0	NA	O ₂
2	CYP102A2	CYP102A2	0	NA	O ₂
3	CYP102A3	CYP102A3	0	NA	O ₂
4	A1	CYP102A1 Heme Domain Only	0	NA	H ₂ O ₂
5	A2	CYP102A2 Heme Domain Only	0	NA	H ₂ O ₂
6	WT F87A	WT F87A	1	NA	O ₂
7	9C1	WT I58V H100R F107L A135S M145V N239H S274T K434E V446I I102T A145V L324I I366V E442K	14	Propranolol	H ₂ O ₂
8	D6H10	9C1 L75H V78E A82P	17	Propranolol	H ₂ O ₂
9	DE10	9C1 A74V A82L A87G	16	Propranolol	H ₂ O ₂
10	2C11	DE10 K24R R47H	18	Propranolol	H ₂ O ₂
11	Chimera	11113311	35	None	H ₂ O ₂
12	Chimera	12112333	96	None	H ₂ O ₂
13	Chimera	21112233	98	None	H ₂ O ₂
14	Chimera	21112331	85	None	H ₂ O ₂
15	Chimera	21112333	89	None	H ₂ O ₂
16	Chimera	21113312	97	None	H ₂ O ₂
17	Chimera	21113333	75	None	H ₂ O ₂
18	Chimera	21212233	89	None	H ₂ O ₂
19	Chimera	21212333	87	None	H ₂ O ₂
20	Chimera	21311231	81	None	H ₂ O ₂
21	Chimera	21311233	97	None	H ₂ O ₂
22	Chimera	21311311	63	None	H ₂ O ₂
23	Chimera	21311313	95	None	H ₂ O ₂
24	Chimera	21311331	81	None	H ₂ O ₂
25	Chimera	21311333	81	None	H ₂ O ₂
26	Chimera	21312133	100	None	H ₂ O ₂
27	Chimera	21312211	76	None	H ₂ O ₂
28	Chimera	21312213	99	None	H ₂ O ₂
29	Chimera	21312231	94	None	H ₂ O ₂
30	Chimera	21312233	96	None	H ₂ O ₂
31	Chimera	21312311	76	None	H ₂ O ₂
32	Chimera	21312313	98	None	H ₂ O ₂

33	Chimera	21312331	94	None	H ₂ O ₂
34	Chimera	21312332	83	None	H ₂ O ₂
35	Chimera	21312333	80	None	H ₂ O ₂
36	Chimera	21313111	58	None	H ₂ O ₂
37	Chimera	21313231	96	None	H ₂ O ₂
38	Chimera	21313233	82	None	H ₂ O ₂
39	Chimera	21313311	78	None	H ₂ O ₂
40	Chimera	21313313	84	None	H ₂ O ₂
41	Chimera	21313331	96	None	H ₂ O ₂
42	Chimera	21313333	66	None	H ₂ O ₂
43	Chimera	21333233	61	None	H ₂ O ₂
44	Chimera	22112233	81	None	H ₂ O ₂
45	Chimera	22112333	93	None	H ₂ O ₂
46	Chimera	22212333	88	None	H ₂ O ₂
47	Chimera	22223132	55	None	H ₂ O ₂
48	Chimera	22311233	92	None	H ₂ O ₂
49	Chimera	22311331	98	None	H ₂ O ₂
50	Chimera	22311333	85	None	H ₂ O ₂
51	Chimera	22312231	78	None	H ₂ O ₂
52	Chimera	22312233	79	None	H ₂ O ₂
53	Chimera	22312331	94	None	H ₂ O ₂
54	Chimera	22312333	84	None	H ₂ O ₂
55	Chimera	22313231	92	None	H ₂ O ₂
56	Chimera	22313233	86	None	H ₂ O ₂
57	Chimera	22313331	102	None	H ₂ O ₂
58	Chimera	22313333	70	None	H ₂ O ₂
59	Chimera	23132233	70	None	H ₂ O ₂
60	Chimera	32312231	101	None	H ₂ O ₂
61	Chimera	32312333	53	None	H ₂ O ₂
62	Chimera	32313233	55	None	H ₂ O ₂
63	Chimera	11113311-R1	35	None	O ₂
64	Chimera	12112333-R1	96	None	O ₂
65	Chimera	21113312-R1	97	None	O ₂
66	Chimera	21113312-R2	97	None	O ₂
67	Chimera	21311231-R1	81	None	O ₂
68	Chimera	21311233-R1	97	None	O ₂
69	Chimera	21313311-R1	78	None	O ₂
70	Chimera	21333233-R2	61	None	O ₂
71	Chimera	22132231-R1	77	None	O ₂
72	Chimera	22223132-R1	55	None	O ₂
73	Chimera	22312333-R1	84	None	O ₂
74	Chimera	22313233-R1	86	None	O ₂
75	Chimera	23132233-R1	70	None	O ₂
76	Chimera	23132233-R2	70	None	O ₂
77	Chimera	32312231-R1	101	None	O ₂
78	Chimera	32312333-R1	53	None	O ₂
79	Chimera	32313233-R1	55	None	O ₂
80	139-3	WT V78A H138Y T175I V178I A184V H236Q E252G R255S A290V A295T L353V	10	Octane	O ₂

81	J	139-3 Y138H I178V F205C S226R	10	Propane	O ₂
82	9-10A	J R47C K94I P142S	13	Propane	O ₂
83	9-10A A328F	9-10A A328F	14	Propane	O ₂
84	9-10A A328M	9-10A A328M	14	Propane	O ₂
85	9-10A A328V	9-10A A328V	14	Propane	O ₂
86	9-10A A78S	9-10A A78S	13	Propane	O ₂
87	9-10A A78T	9-10A A78T	13	Propane	O ₂
88	9-10A L75F	9-10A A78F	13	Propane	O ₂
89	9-10A A82C	9-10A A82C	14	Propane	O ₂
90	9-10A A82F	9-10A A82F	14	Propane	O ₂
91	9-10A A82G	9-10A A82G	14	Propane	O ₂
92	9-10A A82I	9-10A A82I	14	Propane	O ₂
93	9-10A A82L	9-10A A82L	14	Propane	O ₂
94	9-10A A82S	9-10A A82S	14	Propane	O ₂
95	9-10A A82T	9-10A A82T	14	Propane	O ₂
96	9-10A F87A	9-10A F87A	14	Propane	O ₂
97	9-10A F87I	9-10A F87I	14	Propane	O ₂
98	9-10A F87L	9-10A F87L	14	Propane	O ₂
99	9-10A F87V	9-10A F87V	14	Propane	O ₂
100	9-10A L75I	9-10A L75I	14	Propane	O ₂
101	9-10A L75W	9-10A L75W	14	Propane	O ₂
102	9-10A T260L	9-10A T260L	14	Propane	O ₂
103	9-10A T260N	9-10A T260N	14	Propane	O ₂
104	9-10A T260S	9-10A T260S	14	Propane	O ₂
105	9-10A T88L	9-10A T88L	14	Propane	O ₂
106	1-12G	9-10A A82L A328V	15	Propane	O ₂
107	7-11D	9-10A A82F A328V	15	Propane	O ₂
108	11-8E	9-10A F87V A328L	15	Propane	O ₂
109	12-10C	9-10A A82G F87V A328V	16	Propane	O ₂
110	13-7C	9-10A A78T A328V	14	Propane	O ₂
111	23-11B	9-10A A78T A82G F87V A328L	16	Propane	O ₂
112	29-10E	9-10A A82F A328F	15	Propane	O ₂
113	29-3E	9-10A A78F A82G A328F	14	Propane	O ₂
114	35-7F	9-10A A78F A82S A328L	14	Propane	O ₂
115	41-5B	9-10A A78F A82G A328V	14	Propane	O ₂
116	49-1A	9-10A A78T A82G F87V A328L	16	Propane	O ₂
117	49-9B	9-10A A82G F87L A328L	16	Propane	O ₂
118	53-5H	9-10A A78F A82S A328F	15	Propane	O ₂
119	68-8F	9-10A A78F A82G A328L	15	Propane	O ₂
120	77-9H	9-10A A78T A82G A328L	15	Propane	O ₂

I. References

1. Dean, A. M., Miller, S. P. & Lunzer, M. (2006). Direct demonstration of an adaptive constraint. *Science* **314**, 458-461.
2. Dean, A. M., Zhu, G. P. & Golding, G. B. (2005). The selective cause of an ancient adaptation. *Science* **307**, 1279-1282.
3. Tokuriki, N. & Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596-604.
4. Tawfik, D. S., Bershtein, S., Segal, M., Bekerman, R. & Tokuriki, N. (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929-932.
5. Camps, M., Herman, A., Loh, E. & Loeb, L. A. (2007). Genetic constraints on protein evolution. *Critical Reviews in Biochemistry and Molecular Biology* **42**, 313-326.
6. Raines, R. T. & Smith, B. D. (2006). Genetic selection for critical residues in ribonucleases. *Journal of Molecular Biology* **362**, 459-478.
7. Tawfik, D. S., Tokuriki, N., Stricher, F., Schymkowitz, J. & Serrano, L. (2007). The stability effects of protein mutations appear to be universally distributed. *Journal of Molecular Biology* **369**, 1318-1332.
8. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology* **320**, 369-387.
9. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research* **33**, W382-W388.
10. Moulton, J., Yue, P. & Li, Z. L. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology* **353**, 459-473.
11. Goldstein, R. A. & Taverna, D. M. (2002). Why are proteins marginally stable? *Proteins-Structure Function and Genetics* **46**, 105-109.
12. Pace, C. N. (1975). The stability of globular proteins. *CRC Crit Rev Biochem* **3**, 1-43.
13. Plaxco, K. W., Simons, K. T., Ruczinski, I. & David, B. (2000). Topology, stability, sequence, and length: Defining the determinants of two-state protein folding kinetics. *Biochemistry* **39**, 11177-11183.
14. Shakhnovich, E. I., Zeldovich, K. B. & Chen, P. Q. (2007). Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 16152-16157.
15. Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C. & Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 606-611.
16. Fersht, A. R., Nikolova, P. V., Wong, K. B., DeDecker, B. & Henckel, J. (2000). Mechanism of rescue of common p53 cancer mutations by second-site suppressor mutations. *Embo Journal* **19**, 370-378.
17. Bloom, J. D., Wilke, C. O., Arnold, F. H. & Adami, C. (2004). Stability and the evolvability of function in a model protein. *Biophysical Journal* **86**, 2758-2764.
18. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5869-5874.
19. Hilvert, D., Besenmatter, W. & Kast, P. (2007). Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. *Proteins-Structure Function and Bioinformatics* **66**, 500-506.

20. Arnold, F. H., Giver, L., Gershenson, A. & Freskgard, P. O. (1998). Directed evolution of a thermostable esterase. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 12809-12813.
21. Van den Burg, B., Vriend, G., Veltman, O. R. & Eijsink, V. G. H. (1998). Engineering an enzyme to resist boiling. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 2056-2060.
22. Arnold, F. H. & Zhao, H. M. (1999). Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Engineering* **12**, 47-53.
23. Serrano, L., Day, A. G. & Fersht, A. R. (1993). 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. *Journal of Molecular Biology* **233**, 305-312.
24. Kamtekar, S., Schiffer, J. M., Xiong, H. Y., Babik, J. M. & Hecht, M. H. (1993). Protein Design by Binary Patterning of Polar and Nonpolar Amino-Acids. *Science* **262**, 1680-1685.
25. Cordes, M. H. & Sauer, R. T. (1999). Tolerance of a protein to multiple polar-to-hydrophobic surface substitutions. *Protein Sci* **8**, 318-25.
26. Berezovsky, I. N., Tokuriki, N., Oldfield, C. J., Uversky, V. N. & Tawfik, D. S. (2009). Do viral proteins possess unique biophysical features? *Trends in Biochemical Sciences* **34**, 53-59.
27. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. (2008). How Protein Stability and New Functions Trade Off. *Plos Computational Biology* **4**.
28. Arnold, F. H., Wintrode, P. L., Miyazaki, K. & Gershenson, A. (2001). How enzymes adapt: lessons from directed evolution. *Trends in Biochemical Sciences* **26**, 100-106.
29. Tawfik, D. S., Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S. M. & Roodveldt, C. (2005). The 'evolvability' of promiscuous protein functions. *Nature Genetics* **37**, 73-76.
30. Shoichet, B. K., Wang, X. J. & Minasov, G. (2002). Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *Journal of Molecular Biology* **320**, 85-95.
31. Tawfik, D. S., James, L. C. & Roversi, P. (2003). Antibody multispecificity mediated by conformational diversity. *Science* **299**, 1362-1367.
32. Davail, S., Feller, G., Narinx, E. & Gerday, C. (1994). Cold Adaptation of Proteins - Purification, Characterization, and Sequence of the Heat-Labile Subtilisin from the Antarctic Psychrophile Bacillus Ta41. *Journal of Biological Chemistry* **269**, 17448-17453.
33. Fontana, A., De Filippis, V., de Laureto, P. P., Scaramella, E. & Zamboni, M. (1998). Rigidity of thermophilic enzymes. *Stability and Stabilization of Biocatalysts* **15**, 277-294.
34. Anzenbacher, P., Skopalik, J. & Otyepka, M. (2008). Flexibility of human cytochromes P450: Molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *Journal of Physical Chemistry B* **112**, 8165-8173.
35. Tawfik, D. S. & Peisajovich, S. G. (2007). Protein engineers turned evolutionists. *Nature Methods* **4**, 991-994.
36. Babbitt, P. C., Glasner, M. E. & Gerlt, J. A. (2006). Evolution of enzyme superfamilies. *Current Opinion in Chemical Biology* **10**, 492-497.
37. Jensen, R. A. (1976). Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology* **30**, 409-425.
38. Herschlag, D. & O'Brien, P. J. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology* **6**, R91-R105.

39. Nelson, D. R. (2009). The cytochrome p450 homepage. *Human genomics* **4**, 59-65.
40. Ortiz de Montellano, P. R. (1995). *Cytochrome P450: Structure, Mechanism, and Biochemistry*, Plenum Press, New York and London.
41. Guengerich, F. P. (2003). Cytochromes P450, Drugs, and Diseases. *Mol. Interv.* **3**, 194-204.
42. Guengerich, F. P. (2003). New overview of distribution, substrates, inhibitors, active sites, and clinical relevance of the fifty seven human P450s. In *Cytochrome P450: Structure, Mechanism, and Biochemistry* 3rd edit. (Ortiz de Montellano, P. R., ed.). Plenum Press, New York.
43. Gunaratna, C. (2000). Drug Metabolism & Pharmacokinetics in Drug Discovery: A Primer for Bioanalytical Chemists, Part 1. *Current Separations* **19**, 17-23.
44. Rendic, S. (2002). Summary of information on human CYP enzymes: Human P450 metabolism data. *Drug Metabolism Reviews* **34**, 83-448.
45. Evans, W. E. & Relling, M. V. (1999). Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science* **286**, 487-491.
46. Vail, R. B., Homann, M. J., Hanna, I. & Zaks, A. (2005). Preparative synthesis of drug metabolites using human cytochrome P450s 3A4, 2C9 and 1A2 with NADPH-P450 reductase expressed in *Escherichia coli*. *Journal of Industrial Microbiology & Biotechnology* **32**, 67-74.
47. Graham-Lorence, S. & Peterson, J. A. (1996). P450s: Structural similarities and functional differences. *Faseb Journal* **10**, 206-214.
48. Narhi, L. O. & Fulco, A. J. (1986). Characterization of a Catalytically Self-Sufficient 119,000-Dalton Cytochrome-P-450 Monooxygenase Induced by Barbiturates in *Bacillus-Megaterium*. *Journal of Biological Chemistry* **261**, 7160-7169.
49. Munro, A. W., Leys, D. G., McLean, K. J., Marshall, K. R., Ost, T. W. B., Daff, S., Miles, C. S., Chapman, S. K., Lysek, D. A., Moser, C. C., Page, C. C. & Dutton, P. L. (2002). P450BM3: the very model of a modern flavocytochrome. *Trends in Biochemical Sciences* **27**, 250-257.
50. Lewis, D. F. V., Eddershaw, P. J., Goldfarb, P. S. & Tarbit, M. H. (1996). Molecular modelling of CYP3A4 from an alignment with CYP102: Identification of key interactions between putative active site residues and CYP3A-specific chemicals. *Xenobiotica* **26**, 1067-1086.
51. Anzenbacherova, E., Bec, N., Anzenbacher, P., Hudecek, J., Soucek, P., Jung, C., Munro, A. W. & Lange, R. (2000). Flexibility and stability of the structure of cytochromes P450 3A4 and BM-3. *European Journal of Biochemistry* **267**, 2916-2920.
52. Landwehr, M., Carbone, M., Otey, C. R., Li, Y. G. & Arnold, F. H. (2007). Diversification of catalytic function in a synthetic family of chimeric cytochrome P450s. *Chemistry & Biology* **14**, 269-278.
53. Otey, C. R., Bandara, G., Lalonde, J., Takahashi, K. & Arnold, F. H. (2006). Preparation of human metabolites of propranolol using laboratory-evolved bacterial cytochromes P450. *Biotechnology and Bioengineering* **93**, 494-499.
54. Hochrein, L. (2006). Engineering Bacterial Cytochrome P450 BM-3 to Hydroxylate Drug Compounds, California Institute of Technology.
55. Fasan, R., Chen, M. M., Crook, N. C. & Arnold, F. H. (2007). Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting natively-like catalytic properties. *Angewandte Chemie-International Edition* **46**, 8414-8418.

56. Peters, M. W., Meinhold, P., Glieder, A. & Arnold, F. H. (2003). Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. *Journal of the American Chemical Society* **125**, 13442-13450.
57. Fasan, R., Meharena, Y. T., Snow, C. D., Poulos, T. L. & Arnold, F. H. (2008). Evolutionary History of a Specialized P450 Propane Monooxygenase. *Journal of Molecular Biology* **383**, 1069-1080.
58. Arnold, F. H., Meinhold, P., Peters, M. W., Chen, M. M. Y. & Takahashi, K. (2005). Direct conversion of ethane to ethanol by engineered cytochrome P450BM3. *Chembiochem* **6**, 1765-1768.
59. Salazar, O., Cirino, P. C. & Arnold, F. H. (2003). Thermostabilization of a cytochrome P450 peroxygenase. *Chembiochem* **4**, 891-893.
60. Lewis, J. C., Mantovani, S. M., Fu, Y., Snow, C. D., Komor, R. S., Wong, C. H. & Arnold, F. H. (2010). Combinatorial Alanine Substitution Enables Rapid Optimization of Cytochrome P450(BM3) for Selective Hydroxylation of Large Substrates. *Chembiochem* **11**, 2502-2505.
61. Lewis, J. C., Bastian, S., Bennett, C. S., Fu, Y., Mitsuda, Y., Chen, M. M., Greenberg, W. A., Wong, C. H. & Arnold, F. H. (2009). Chemoenzymatic elaboration of monosaccharides using engineered cytochrome P450(BM3) demethylases. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 16550-16555.
62. Landwehr, M., Hochrein, L., Otey, C. R., Kasrayan, A., Backvall, J. E. & Arnold, F. H. (2006). Enantioselective alpha-hydroxylation of 2-arylacetic acid derivatives and buspirone catalyzed by engineered cytochrome P450BM-3. *Journal of the American Chemical Society* **128**, 6058-6059.
63. Li, Q. S., Ogawa, J., Schmid, R. D. & Shimizu, S. (2001). Residue size at position 87 of cytochrome P450BM-3 determines its stereoselectivity in propylbenzene and 3-chlorostyrene oxidation. *Febs Letters* **508**, 249-252.
64. Li, Q. S., Ogawa, J., Schmid, R. D. & Shimizu, S. (2001). Engineering cytochrome P450BM-3 for oxidation of polycyclic aromatic hydrocarbons. *Applied and Environmental Microbiology* **67**, 5735-5739.
65. Bostrom, J., Greenwood, J. R. & Gottfries, J. (2003). Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *Journal of Molecular Graphics & Modelling* **21**, 449-462.
66. Furuya, T., Shibata, D. & Kino, K. (2009). Phylogenetic analysis of Bacillus P450 monooxygenases and evaluation of their activity towards steroids. *Steroids* **74**, 906-912.
67. Bureik, M. & Bernhardt, R. (2007). *Steroid Hydroxylation: Microbial Steroid Biotransformations Using Cytochrome P450 Enzymes*. Modern Biooxidation: Enzymes, Reaction, and Applications (Schmid, R. D. & Urlacher, V. B., Eds.), Wiley-VCH.
68. Zehentgruber, D., Hannemann, F., Bleif, S., Bernhardt, R. & Lutz, S. (2010). Towards Preparative Scale Steroid Hydroxylation with Cytochrome P450 Monooxygenase CYP106A2. *Chembiochem* **11**, 713-721.
69. Vagin, A. & Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. *Journal of Applied Crystallography* **30**, 1022-1025.
70. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D-Biological Crystallography* **60**, 2126-2132.
71. Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011).

Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D-Biological Crystallography* **67**, 235-242.

72. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). A graphical user interface to the CCP4 program suite. *Acta Crystallographica Section D-Biological Crystallography* **59**, 1131-1137.
73. Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K. & Zhu, G. (2003). Improved amino acid flexibility parameters. *Protein Science* **12**, 1060-1072.

Chapter 4

Materials and Methods

A. CBHI Gene Construction

Parent CBHI genes feature native codon usage and were synthesized by DNA2.0 (Menlo Park, CA, USA). The CBHI genes were digested by adding 1 μg of CBHI DNA to 1.3 μL NheI and Acc65I restriction enzymes from New England BioLabs (NEB) (Ipswich, MA, USA) with 1X NEBuffer 2 supplemented with 100 $\mu\text{g}/\text{mL}$ BSA in 50 μL total volume for 1 hr at 37 $^{\circ}\text{C}$. The digestion mixture was brought to 1X DNA loading buffer (NEB) and 1X SYBR Gold from Invitrogen (Carlsbad, CA, USA) and loaded onto a 1% agarose gel. Gels were run at 100 V for 30 min and the band corresponding to the CBHI sequence (~1500 kb) was excised and purified using a QIAquick Gel Extraction Kit from Qiagen (Valencia, CA, USA) following the manufacturer's protocol and eluted into 40 μL 0.1X buffer EB in water. The insert was then ligated into the yeast expression vector Yep352/PGK91-1- αss (Figure 4.1), similarly digested and purified, by adding 18 μL of the insert (~20 $\text{ng}/\mu\text{L}$) to 3 μL (~50 $\text{ng}/\mu\text{L}$) of the vector with 1.5 μL T4 DNA ligase (NEB) in the presence of 1X T4 DNA Ligase Reaction Buffer at 16 $^{\circ}\text{C}$ for 16 hr. Completed ligations were purified using the DNA Clean & Concentrator from Zymo Research (Irvine, CA, USA) and eluted into 8 μL water.

All 8 μL were then transformed into 50 μL *E. coli* DH5 α electro-competent cells by pulsing at 2.5 kV in 0.2 cm electroporation cuvettes, rescuing at 37 $^{\circ}\text{C}$ with shaking at 250 rpm for 1 hr in the presence of 1 mL SOC media. Transformation mixtures were then centrifuged at 5000 rpm for 1 min, the supernatant drawn off, and remaining ~100 μL plated on Luria-Bertani (LB) broth plates supplemented with 1 g/L ampicillin and 15 wt% agar. The mixture was spread with sterile glass beads and grown overnight at 37 $^{\circ}\text{C}$. The

next day, single colonies were picked and grown overnight in 5 mL liquid LB broth supplemented with 1 g/L ampicillin. Cultures were centrifuged at 4500 rpm for 5 min and the plasmid DNA extracted using a QIAprep Miniprep kit (Qiagen) and eluted into 40 μ L 0.1X buffer EB in water. To verify the sequences, 10 μ L of plasmid (\sim 200 ng/ μ L) was sent for sequencing (see Table 4.1 for primer sequences).

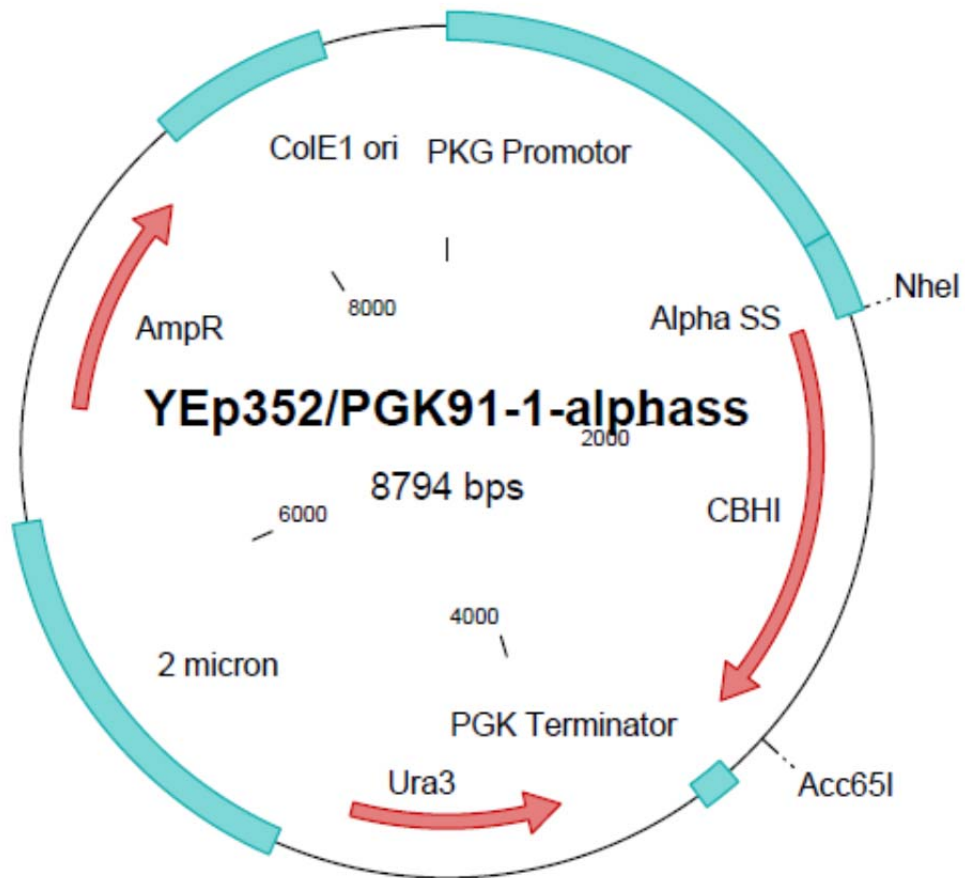


Figure 4.1. Plasmid construct used for expression of CBHI genes.

N-terminal His₆ CBHI constructs were made via PCR using Phusion High-Fidelity DNA Polymerase from Finnzymes (Vantaa, Finland) according to manufacturer's

protocol on PCR conditions with forward primers complementary to the appropriate CBHI N-terminal sequence with NheI and His₆ overhangs. After PCR, the constructs were prepared identically to the CBHI parental genes as described above.

CBHI chimeric genes were constructed by splicing by overlap extension PCR¹. Block junction primers consist of 30 bp complementary to the corresponding parents on either side of the block boundary. Block fragments were constructed using Phusion High-Fidelity DNA Polymerase as per the manufacturer's protocol on PCR conditions. The fragments were purified on a 1% agarose gel as described above. Up to five fragments were assembled at once using 1 µL of each purified fragment as template and Phusion PCR conditions. The resulting full-length CBHI gene was then prepared identically to the CBHI parental genes as described above.

Point mutations were introduced using the QuikChange Lightning Site-Directed Mutagenesis Kit from Agilent Technologies (Santa Clara, CA, USA) as per the manufacturer's protocol. Primers for the point mutations were composed of the new codon flanked on either side by 8-15 bp complementary to the parental sequence.

Table 4.1. Primer list for cellulase constructs.

Name	Sequence
Forward_Seq	GCTGAAGCTGTCATCGGTTACTTAG
Reverse_Seq	GCAACACCTGGCAATTCCTTA
TE_His6	TGAACCTGCTAGCCACCATCATCACCACCATCAGCAGGCCTGTACGGCGAC GGCAGAGAA
CT_His6	TGAACCTGCTAGCCACCATCATCACCACCATCAGCAGGCCTGTCTCCCTCAC CGCTGAGAA
TA_His6	TGAACCTGCTAGCCACCATCATCACCACCATCAGCAGGCCTGTACCGTAAC CGCAGAGAA
S13P (CT block 1)	GAGAACCACCCT CCT CTCACCTGGAAG
T41V (HJ block 2)	AACTGGCGCTGG GTT CACGCTACGAAC
Y60I (TE block 2)	TGGGACCCACG ATT TGCCCTGACGA
Y60L (TE block 2)	TGGGACCCACG CTT TGCCCTGACGA
V109L (CT block 3)	ACCAACGTCGGCTCCCGT CTT TATCTG
T162K (HJ block 5)	CCCACCAAC AAG GCTGGCGCCAAGTA
A199P (HJ block 5)	TCATCCAACAAC CCT AACACGGGCATT
T252I (TE block 7A)	TACGCGGGA ATT TGCGATCCTGACGGC
T252V (TE block 7A)	TACGCGGGA GTT TGCGATCCTGACG
T268K (TE block 7A)	CGCATGGGCAAC AAG TCTTTCTACGGG
S318P (AT block 7C)	GTCATCCCCAAC CCT GAGTCCAAGATC
S320P (TA block 7C)	GTAATCCCCAG CCT GAGTCGACGATC
A378Y (TA block 8)	TGGGACGATCAC TAT GCCAACATGCTC
Y425F (TA block 8)	TCATATGTTATC TTC TCCAACATCAAGGTCGG
N434G (TA block 8)	TCGGACCCATC GGA TCGACCTTCAC
S5T (CT1)	AGCAGGCTTGC ACC CTCACCGCTGAGAA
D52T (HJ2)	ACGAACTGCTAC ACG GGCAACACTTGG
S57D (HJ2)	GGCAACACTTGG GAC TCGACCTATGT
L108M (CT3 TE4)	CGTGTCTATCTG ATG CAGGACGACTCG
S125T (TE4)	AACCGCGAGTTC ACC TTTGACGTGAT
V215I (TA6)	GCTGAGATGGAT ATC TGGGAAGCCAAC
M364L (TE7F)	AAGATGGGAGCGGCC CTG CAGCAGGGTAT
N92K (CT3)	GGTGACTCCCTG AAG CTCAAGTTCGTC
N121G (TE4)	TTCAAGCTTCTG GGC CGCGAGTTCAGC
H206Y (TA6)	GGTAAC TAT GGTTCTGCTGCGCTGAG
S220K (TA6)	TGGGAAGCCAAC AAA ATCTCTACTGCG
D346V (TE7)	CGACACGGAC GTC TTCTCTCAGCAC
T403D (TA8)	CGCGGT GAT TGCCCCACGAC
TA5S175A	TGGTTACTGCGAC GCT CAGTGCCCTCGGGAT
TA5N185G	CTCAAGTTCATC GGT GGTCAGGCCAAT
TE6V226L	ATCTCCAATGCG CTT ACTCCGCACCCG
TE7T256I	ATCGCTACGCGGGA ATC TGCGATCCTGAC
TE7T256V	TAACGATCGCTACGCGGGA GTC TGCGATCCTGAC

TE7T256K	ATCGCTACGCGGGA AAG TGCGATCCTGAC
TE7T272P	CGCATGGGCAAC CCT TCTTTCTACGGG
TE7D297G	TTCCTCACTGAT GGC GGTACGGATACT
TE7D297K	TTCCTCACTGAT AAG GGTACGGATACT
AT7C E322P	GTCATCCCCAACCCCT CCC TCCAAGATC
AT7C A326G	TCCAAGATC GGA GCGTCTCCGGCAACT
AT7C V328M	TGAGTCCAAGATCGCCGGC ATG TCCGGCAAC
TE7T335P	AACTCGATCACG CCT GAGTTCTGCACT
TE7T335Q	AACTCGATCACG CAG GAGTTCTGCACT
TE7Q341M	TTCTGCACTGCT ATG AAGCAGGCCTTT
TE7H354K	GTCTTCTCTCAG AAG GGTGGCCTGGCCAA
TE7H354R	ACGTCTTCTCTCAG CGT GGTGGCCTGGCCAAGAT
TE7H354V	GTCTTCTCTCAG GTT GGTGGCCTGGCCAA
TA8T388I	TGGCTGGACAGC ATC TACCCGACTGAT
TA8T391P	TGGACAGCACCTACCCG CCT GATGCGGAC
TA8P395G	ATGCGGAC GGA GACACCCCTGGCGT
TA8V400A	ACACCCCTGGC GCT GCGCGCGGTACTT

B. Yeast Culturing

S. cerevisiae expression strain YDR483W BY4742 was made competent using the Frozen-EZ Yeast Transformation II kit (Zymo Research), transformed with 1 μ L plasmid DNA (~200 ng/ μ L), plated on synthetic dropout -uracil supplemented with 16 wt% agar, and grown at 30 °C for 2 days. Single colonies were picked on the second day, placed in 5 mL synthetic dextrose casamino acids (SDCAA) media (20 g/L dextrose, 6.7 g/L Difco yeast nitrogen base, 5 g/L Bacto casamino acids, 5.4 g/L Na₂HPO₄, 8.56 g NaH₂PO₄·H₂O), and grown overnight with shaking at 250 rpm. The following morning, cultures were expanded into 40 mL of yeast peptone dextrose (YPD) media (20 g Bacto peptone, 10 g Bacto yeast extract, and 20 g dextrose) in 250 mL Tunair flasks from Sigma-Aldrich (St. Louis, MO, USA) and grown for 48 hours with 250 rpm shaking. Cultures were

centrifuged at 4500 rcf for 15 min, then decanted, brought to 0.02% NaN₃ and 1/200X Protease Inhibitor Cocktail (Sigma-Aldrich) and stored at 4 °C until characterization.

C. Cellulase Purification

N-terminal His₆ CBHI constructs were grown as described in section B. After centrifugation, supernatants were filtered with 0.45 μM pore size filter units from Nalgene (Rochester, NY, USA), brought to 0.02% NaN₃ and 1/200X Protease Inhibitor Cocktail, and concentrated to under 1 mL with Vivaspin 20 ultrafiltration spin columns with a 30 kDa MWCO PES membrane from Sartorius Stedim (Aubagne Cedex, France). The concentrated supernatants were then purified using Ni-NTA spin columns (Qiagen) per the manufacturer's protocol and the proteins exchanged into 50 mM sodium acetate, pH 4.8, using the Vivaspin 20 spin columns. Purified protein concentration was determined using the Bradford Protein Assay from Bio Rad (Hercules, CA, USA) with a bovine serum albumin standard and concentrations determined by averaging readings of multiple dilutions for each sample. Post-Ni-NTA isolation CBHI yield estimates range from 500 μg/L culture for the poorly secreted *T. aurantiacus* parent CBHI to 50 mg/L for the most highly expressed chimera.

D. Total Yeast-Secreted CBHI Activity Assay

Total yeast-secreted CBHI activity toward to the soluble substrate 4-methylumbelliferyl lactopyranoside (MUL) (Sigma-Aldrich) was determined by adding 125 μL of culture supernatant to 25 μL of 1.8 mM MUL dissolved in 125 mM sodium

acetate, pH 4.8, 18% DMSO, incubating at 45 °C for 30 min and quenching with 150 μ L of 1 M Na_2CO_3 . MUL hydrolysis rates were determined by using a microplate reader to measure sample fluorescence with excitation at 364 nm and emission at 445 nm and comparing values to a standard curve prepared with 4-methylumbelliferone (Sigma-Aldrich). Error in the measurements was less than 10%. Our treating total yeast-secreted CBHI activity toward MUL as a proxy for CBHI expression is based on our observation that Ni-NTA affinity-isolated, N-terminally His₆-tagged CBHI parents and chimeras have similar specific activities toward MUL. We did not attempt to determine whether the nonsecreted or poorly secreted enzymes were expressed but retained within the host cells.

E. T₅₀ Assay

The T₅₀ value is defined as the temperature at which a ten-minute incubation in the absence of substrate causes loss of one-half of the activity, measured after reaction with MUL substrate, relative to a 100% activity reference sample that does not undergo incubation. For T₅₀ assays, culture supernatants were diluted using a supernatant from a negative control YPD yeast culture not containing secreted cellulase (it contains the Yep352/PGK91-1- α ss plasmid containing a CBHI gene with a frameshift) so that approximately equivalent MUL hydrolysis rates of 2.0×10^{-9} mol MUL/L/s was obtained for samples not incubated for thermal denaturation. These diluted samples were adjusted to 1 mM DTT and 125 mM sodium acetate, pH 4.8. Aliquots of 125 μ L were incubated for 10 min in a water bath across a range of temperatures bracketing the T₅₀

value. Water bath temperatures were measured using two different alcohol thermometers and observed to be consistent within 0.1 °C. After cooling, 25 µL of 1.8 mM MUL dissolved in 125 mM sodium acetate, pH 4.8, 18% DMSO, was added to the incubated and unheated samples, and these were reacted in a 45 °C water bath for 90 min. MUL hydrolysis was determined as in section D, and the T_{50} value was calculated by linear interpolation of data using Microsoft Excel.

F. Half-Life Assay

At a given temperature, a half-life is defined as the incubation time in the absence of substrate that causes loss of one-half of the activity, measured after reaction with MUL substrate, relative to a 100% activity reference sample that does not undergo incubation. For half-life assays, samples were prepared as in section E but were incubated at constant temperature for 15, 30, 60, 90, and 120 min in addition to an unheated sample. Samples were cooled, quenched, and read as in section E. Half-lives are calculated with Microsoft Excel by fitting a straight line to a graph of the natural logarithm of the residual activity vs. time. Measurements were repeated on separate days.

G. Circular Dichroism

Circular dichroism measurements were performed on an Circular Dichroism Spectrometer Model 62DS from Aviv (Lakewood, NJ, USA) in a 21-Q-1 Quartz Spectrophotometer Cell, Rectangular, Stopper, 1 mm from Starna Cells (Atascadero, CA,

USA). Wavelength scans before and after melting were performed in the far-UV spectral region (190-240 nm) useful for determination of secondary structure. Scans showed high signal at the characteristic wavelengths for both alpha-helices (208, 222 nm) and beta-sheets (210-220 nm), implying that folded protein was present. The unfolding curves were measured at 202 nm to measure the increase in concentration of random coils at their characteristic wavelength, using the temperature scan mode with a gradient of 2 °C/min from 20 to 100 °C. The measurements were performed in 10 mM sodium acetate, pH 4.8, with 3.85 µM protein concentration.

H. MUL Temperature Activity Profiles

CBHI activity measurements on MUL were performed at different temperatures. To verify that CBHIs maintained full activity at 45 °C, supernatants were incubated for 30 min at 45 °C and then reacted with MUL for 30 min at 45 °C. These incubated samples showed no difference in activity from unincubated samples. CBHI culture supernatants were therefore diluted as in section E based on MUL activity measurements at 45 °C for 30 min so that approximately equivalent MUL hydrolysis rates of 6.0×10^{-9} mol MUL/L/s were obtained. Duplicate samples of each dosed enzyme was reacted at several temperatures for 30 min, quenched, and read as in section E.

I. Solid Cellulose Temperature Activity Profiles

CBHI solid cellulose temperature activity profiles were obtained by assuming that all protein in the affinity-isolated CBHI samples was fully active CBHI and adding

5 μg to 500 μL of 50 mM sodium acetate, pH 4.8, containing 60 mg/mL Lattice NT cellulose from FMC (Philadelphia, PA, USA). After incubation for 20 hr in a water bath at the temperature of interest, supernatant reducing sugar was determined by the Nelson-Somogyi assay as described in J. Reactions were run in duplicate and repeated on separate days.

J. Nelson-Somogyi Reducing Sugar Assay

After incubation with solid cellulose, reaction tubes are centrifuged at 14000 rpm for 3 min to pellet the solid cellulose. The supernatant is drawn off and 200 μL is added to 200 μL of a 4:1 mixture of Somogyi reagent 1 (180 g/L Na_2SO_4 , 15 g/L Rochele salts, 30 g/L Na_2CO_3 , and 20 g/L NaHCO_3 in degassed water) and Somogyi reagent 2 (180 g/L Na_2SO_4 , 12.8 g/L anhydrous CuSO_4 in degassed water) by vortexing. The mixture is incubated at 98 $^\circ\text{C}$ for 20 min, vortexed, and added to 200 μL Nelson reagent (50 g/L ammonium molybdate, 6 g/L sodium arsenate, and 42 mL/L concentrated sulfuric acid in water, filtered and incubated at 37 $^\circ\text{C}$ for 28 hr). Product sugar concentration was determined using a microplate reader to measure absorbance at 520 nm and comparing values to a standard curve prepared with cellobiose².

K. P450 Gene Construction

Parent P450_{BM3} genes were taken from the Arnold lab culture collection contained downstream of a double tac promoter of the expression vector pCWori P450 specific plasmid as shown in Figure 4.2. Mutations were incorporated via splicing by

overlap extension PCR. Primers consisting of the mutation flanked by 15 bps on each side (see Table 4.2 for primer list) were used to construct overlapping fragments using Phusion High-Fidelity DNA Polymerase as per the manufacturer's protocol on PCR conditions. The fragments were brought to 1X DNA loading buffer (NEB) and 1X SYBR Gold (Invitrogen) and loaded onto a 1% agarose gel. Gels were run at 100 V for 30 min and the band corresponding to the fragment was excised and purified using a QIAquick Gel Extraction Kit (Qiagen) following the manufacturer's protocol and eluted into 40 μ L 0.1X buffer EB in water.

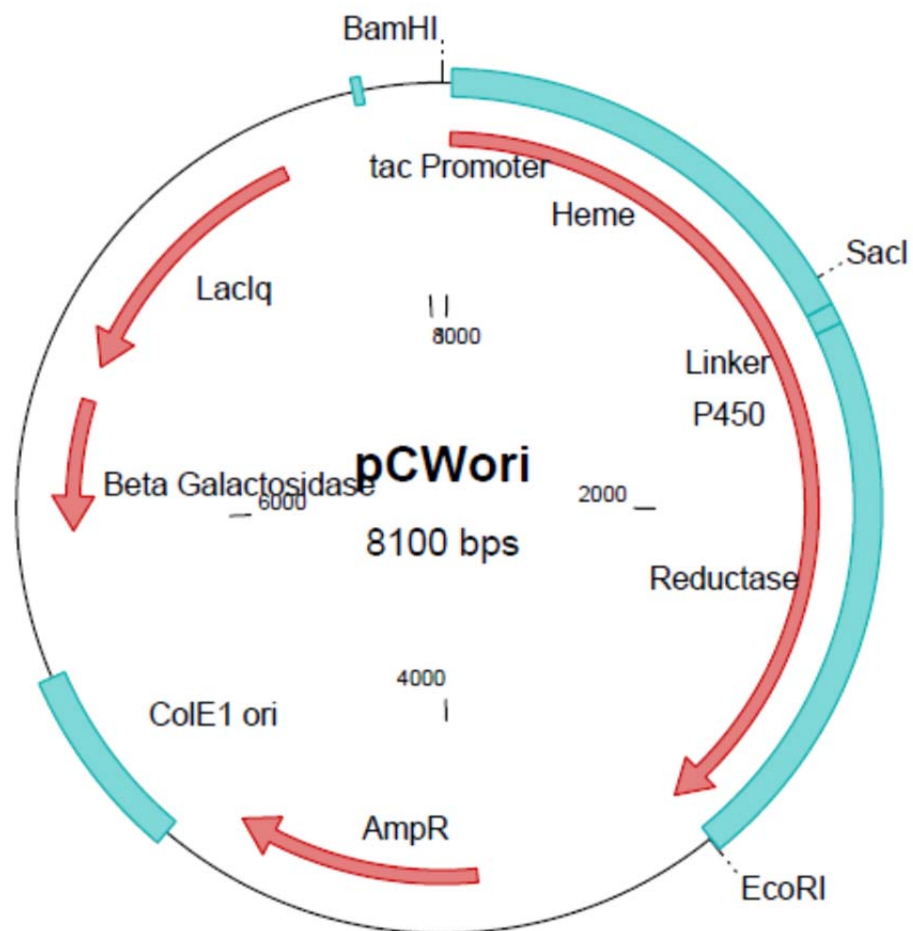


Figure 4.2. Plasmid construct used for expression of P450 genes.

Up to five fragments were assembled at once using 1 μL of each purified fragment as template and Phusion PCR conditions. The assembled inserts were digested by adding 1 μg of insert to 1 μL *SacI* and *BamHI* restriction enzymes (NEB) with 1X NEBuffer 1 supplemented with 100 $\mu\text{g}/\text{mL}$ BSA in 50 μL total volume for 1 hr at 37 $^{\circ}\text{C}$. The inserts were purified on a 1% agarose gel as described above and ligated into the pCWori vector, similarly digested and purified, by adding 18 μL of the insert ($\sim 20 \text{ ng}/\mu\text{L}$) to 3 μL ($\sim 50 \text{ ng}/\mu\text{L}$) of the vector with 1.5 μL T4 DNA ligase (NEB) in the presence of 1X T4 DNA Ligase Reaction Buffer at 16 $^{\circ}\text{C}$ for 16 hr. Completed ligations were purified using the DNA Clean & Concentrator (Zymo Research) and eluted into 8 μL water.

All 8 μL were then transformed into 50 μL *E. coli* DH5 α electro-competent cells by pulsing at 2.5 kV in 0.2 cm electroporation cuvettes, rescuing at 37 $^{\circ}\text{C}$ with shaking at 250 rpm for 1 hr in the presence of 1 mL SOC media. Transformation mixtures were then centrifuged at 5000 rpm for 1 min, the supernatant drawn off, and remaining $\sim 100 \mu\text{L}$ plated on Luria-Bertani (LB) broth plates supplemented with 1 g/L ampicillin and 15 wt% agar. The mixture was spread with sterile glass beads and grown overnight at 37 $^{\circ}\text{C}$. The next day, single colonies were picked and grown overnight in 5 mL liquid LB broth supplemented with 1 g/L ampicillin. Cultures were centrifuged at 4500 rpm for 5 min and the plasmid DNA extracted using a QIAprep Miniprep kit (Qiagen) and eluted into 40 μL 0.1X buffer EB in water. To verify the sequences, 10 μL of plasmid ($\sim 200 \text{ ng}/\mu\text{L}$) was sent for sequencing.

C-terminal His₆ heme domain constructs were made via PCR using Phusion High-Fidelity DNA Polymerase with reverse primers complementary to the appropriate P450 C-terminal sequence with EcoRI and His₆ overhangs. After PCR, the constructs were prepared as above but with EcoRI instead of SacI.

Table 4.2. Primer list for P450 constructs

Name	Sequence
BM3seq1_for	CGC ATA TGC TAA ACG GAA AAG ATC C
BM3seq2_for	GCA GAT ATT GCA ATG AGC AAA GG
BM3seq3_for	CGG TCT GCC CGC CGC ATA AAG
BM3seq4_rev	CAT GTG GAT TTT TCA CTA AG
BM3seq5_for	CAG GAA ACA GGA TCA GCT TAC TCC CC
C47R_for	CGA GGC GCC TGG TCG TGT AAC GCG CTA C
C47R_rev	GTA GCG CGT TAC ACG ACC AGG CGC CTC G
I94K_for	GCT GGA CGC ATG AAA AAA ATT GGA AAA AAG CG
I94K_rev	CGC TTT TTT CCA ATT TTT TTC ATG CGT CCA GC
L52I_for	CGC GCT ACA TAT CAA GTC AGC
L52I_rev	GCT GAC TTG ATA TGT AGC GCG
V340M_for	GAA GAT ACG ATG CTT GGA GGA G
V340M_rev	CTC CTC CAA GCA TCG TAT CTT C
I366V_for	CGT GAT AAA ACA GTT TGG GGA GAC G
I366V_rev	CGT CTC CCC AAA CTG TTT TAT CAC G
E442K_for	CGT TAA AAC CTA AAG GCT TTG TGG
E442K_rev	CCA CAA AGC CTT TAG GTT TTA ACG
L324I_for	CGA AGC GCT GCG CAT CTG GCC AAC T
L324I_rev	AGT TGG CCA GAT GCG CAG CGC TTC G
BamHIDel_for	CAC AGG AAA CAG GAT CCA TCG TGC TTA GG
SacI_rev	CTA GGT GAA GGA ATA CCG CCA AGC GGA
EcoRI_rev	GCT CAT GTT TGA CAG CTT ATC ATC G
C205F_for	CAA CGC CCA GTT TCA AGA AGA TAT CAA

C205F_rev	TTG ATA TCT TCT TGA AAC TGG CGC TTG
S255R_for	GAT GAC GGG AAC ATT CGC TAT CAA ATT ATT AC
S255R_rev	GTA ATA ATT TGA TAG CGA ATG TTC CCG TCA TC
F87A_for	GAG ACG GGT TAG CAA CAA GCT GGA C
F87A_rev	GTC CAG CTT GTT GCT AAC CCG TCT C
74-78NNK_for	AAC TTA AGT CAA NNK CTT AAA TTT NNK CGT GAT TTT
74-78NNK_rev	AAA ATC ACG MNN AAA TTT AAG MNN TTG ACT TAA GTT
78-82NNK_for	CTT AAA TTT NNK CGT GAT TTT NNK GGA GAC GGG
78-82NNK_rev	CCC GTC TCC MNN AAA ATC ACG MNN AAA TTT AAG
185-188NNK_for	GAT GAA GTA NNK AAC AAG NNK CAG CGA GCA
185-188NNK_rev	TGC TCG CTG MNN CTT GTT MNN TAC TTC ATC
74A_for	GAT AAA AAC TTA AGT CAA GCG CTT AAA TTT
74A_rev	AAA TTT AAG CGC AAG ACT TAA GTT TTT ATC
78A_for	CTT AAA TTT GCA CGT GAT TTT GCA GGA
78A_rev	TCC TGC AAA ATC ACG TGC AAA TTT AAG
A87F_for	GCA GGA GAC GGG TTA TTT ACA AGC TGG ACG CAT
A87F_rev	ATG CGT CCA GCT TGT AAA TAA CCC GTC TCC TGC
Wob_N134K_for	GGA GCG TCT AAA KGC AGA TGA GC
Wob_N134K_rev	GCT CAT CTG CMT TTA GAC GCT CC
Wob_Q206R_for	GCG CCA GTT TCR AGA AGA TAT CA
Wob_Q206R_rev	TGA TAT CTT CTY GAA ACT GGC GC
Wob_R226G_T235K_for	CGC AAA GCA RGG GGT GAA CAA AGC GAT GAT TTA TTA AMG CAG ATG CTA
Wob_R226G_T235K_rev	TAG CAT CTG CKT TAA TAA ATC ATC GCT TTG TTC ACC CCY TGC TTT GCG
Wob_L322Q_for	AAA CGA AGC GCW GCG CTT ATG GC
Wob_L322Q_rev	GCC ATA AGC GCS GCG CTT CGT TT
Wob_N381K_R398H_for	GAA AAK CCA AGT GCG ATT CCG CAG CAT GCG TTT AAA CCG TTT GGA AAC GGT CAG CRT GCG
Wob_N381K_R398H_rev	CGC AYG CTG ACC GTT TCC AAA CGG TTT AAA CGC ATG CTG CGG AAT CGC ACT TGG MTT TTC
pET_Heme_rev	ACGA CTC GAG AGTG CTA GGT GAA GGA ATAC
pCWori_Heme_rev	ACGA GAA TTC TCA GTG GTG GTG GTG GTG GTG AGT GCT AGG TGA AGG AAT AC
pCWori_Heme_rev2	ACGA GAA TTC TCA GTG GTG GTG GTG GTG GTG AGT GCT AGG TGA AGG AAT ACC GCC

L. P450 Expression

For high-throughput screening, P450 variants in the pCWori vector were transformed into the catalase-deficient strain of *E. coli* SN0037 (for peroxygenases) or *E. coli* DH5 α (for monooxygenases) and spread on LB plates supplemented with 100 $\mu\text{g}/\text{mL}$ ampicillin as described in section K. The next day, single colonies were placed into 400 μL liquid LB media supplemented with 100 $\mu\text{g}/\text{mL}$ ampicillin in 96 well plates and grown at 37 $^{\circ}\text{C}$ and 80% humidity for 24 hr with shaking at 250 rpm. For inoculation, 50 μL of the LB culture was transferred to 900 μL Terrific Broth (TB) media supplemented with 100 $\mu\text{g}/\text{mL}$ ampicillin and grown at 37 $^{\circ}\text{C}$ and 80% humidity with shaking at 250 rpm for 3 hr, before being induced with isopropyl β -D-1-thiogalactopyranoside (IPTG) and the heme precursor δ -aminolevulinic acid (δ -ALA) to a final concentration of 1 mM. The protein was expressed for 24 hr at 25 $^{\circ}\text{C}$ and 80% humidity with shaking at 250 rpm, then pelleted at 5000 rpm for 10 min at 4 $^{\circ}\text{C}$ and stored at -20 $^{\circ}\text{C}$.

For thermostability measurements, P450 constructs in the pCWori vector were transformed into *E. coli* DH5 α cells and spread on LB plates supplemented with 100 $\mu\text{g}/\text{mL}$ ampicillin as described in section K. The next day, single colonies were placed into 5 mL liquid LB media supplemented with 100 $\mu\text{g}/\text{mL}$ ampicillin and grown at 37 $^{\circ}\text{C}$ and 80% humidity overnight with shaking at 250 rpm. The next day, 50 mL TB supplemented with 100 $\mu\text{g}/\text{mL}$ ampicillin cultures were inoculated with all 5 mL of the starter culture and grown at 37 $^{\circ}\text{C}$ and 80% humidity with shaking at 250 rpm. Once an OD_{600} of 2 was reached (\sim 3 hr), the cultures were induced with IPTG and ALA to a final

concentration of 1 mM and grown for 24 hr at 25 °C and 80% humidity with shaking at 250 rpm, then pelleted at 5000 rpm for 10 min at 4 °C and stored at -20 °C.

For crystallography experiments, C-terminal His₆ heme domains constructs in the pCWori vector were transformed into *E. coli* DH5α cells and spread on LB plates supplemented with 100 µg/mL ampicillin as described in section K. The next day, single colonies were placed into 25 mL liquid TB media supplemented with 100 µg/mL ampicillin and grown at 37 °C and 80% humidity overnight with shaking at 250 rpm. When the culture reached an OD₆₀₀ of 15, three 500 mL TB supplemented with 100 µg/mL ampicillin cultures were inoculated with 7 mL of the starter culture and grown at 37 °C and 80% humidity with shaking at 250 rpm. Once an OD₆₀₀ of 2 was reached (~3 hr), the cultures were induced with IPTG and ALA to a final concentration of 1 mM and grown for 24 hr at 25 °C and 80% humidity with shaking at 250 rpm, then pelleted at 5000 rpm for 10 min at 4 °C and stored at -20 °C.

M. High Throughput Screening

Frozen pellets were thawed at room temperature for 1 hr, then resuspended in 600 µL of 0.1 M phosphate buffer, pH 8.0 (EPPS buffer, pH 8.2 buffer was used for peroxygenases) with 10 mM MgCl₂, 0.5 mg/mL lysozyme and 8 U/mL DNase I and incubated at 37 °C for 1 hr, then spun down at 5000 rpm for 10 min at 4 °C.

P450s show a characteristic Soret band at 450 nm when CO is bound to the iron heme, corresponding to the CO stretch frequency³. This can be used to quantify the amount of folded protein⁴. To each well of a microtiter plate, 100 µL of lysate and

100 μL of 140 mM sodium dithionite in 1.0 M phosphate buffer, pH 8.0 (EPPS buffer, pH 8.2 for peroxigenases) was added and pre-read at 450 and 490 nm. The plates were incubated in a vacuum chamber evacuated to 20 mmHg and filled with CO to 1 atm for 15 min before being read at 450 and 490 nm. The protein concentration, [P450], is then calculated as

$$[\text{P450}] = \frac{(\text{Abs}_{450_1} - \text{Abs}_{450_0}) - (\text{Abs}_{490_1} - \text{Abs}_{490_0})}{\epsilon} \quad \text{Eq. 4.1}$$

where ϵ is the molar extinction coefficient, taken as 91 mM cm^{-1} for P450_{BM3} variants.

Demethylation after hydroxylation at a methyl group is a common P450 reaction leading to loss of a formaldehyde molecule that can be detected using the scheme shown in Figure 4.3⁵. For holoenzyme reactions, 60 μL lysate was added to 110 μL 0.1 M phosphate buffer, pH 8.0, 10 μL 5 M substrate, and 20 μL 20 mM NADPH. For peroxygenase reactions, 50 μL lysate was added to 100 μL 0.1 M EPPS buffer, pH 8.2, 10 μL 5 M substrate, and 40 μL 5 mM hydrogen peroxide. The reactions were mixed and left to react for 2 hr at room temperature. The reactions were pre-read at 550 nm before 50 μL 168 mM purpald in 2 M NaOH was added. The samples were left to react for 20 min and then read at 550 nm.

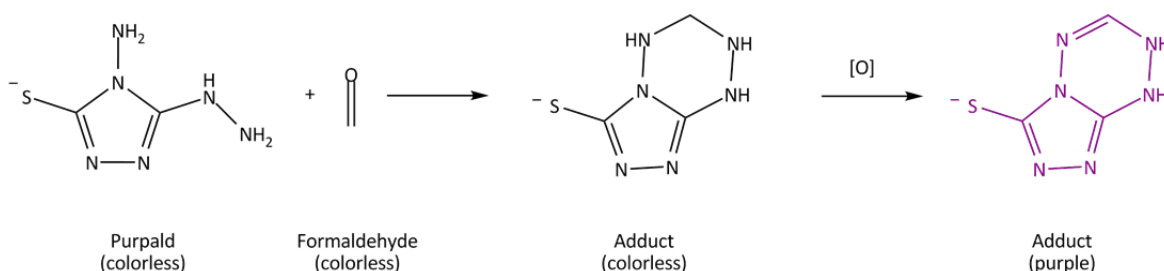


Figure 4.3. Reaction scheme for detection of formaldehyde, a product of P450 methyl group hydroxylation. The purple adduct is easily detected by eye and absorbs at 550 nm.

Aromatic alcohols are common products of P450 hydroxylation that can be detected using the scheme shown in Figure 4.4⁶. For holoenzyme reactions, 60 μL lysate was added to 110 μL 0.1 M phosphate buffer, pH 8.0, 10 μL 5 M substrate, and 20 μL 20 mM NADPH. For peroxygenase reactions, 50 μL lysate was added to 100 μL 0.1 M EPPS buffer, pH 8.2, 10 μL 5 M substrate, and 40 μL 5 mM hydrogen peroxide. The reactions were mixed and left to react for 2 hr at room temperature before 60 μL 8 M urea in 200 mM NaOH was added to quench the reaction. Next, 18 μL 1.2 wt% 4-AAP in water was added and the solutions were pre-read at 510 nm, before 18 μL 1.2 wt% $\text{K}_2\text{S}_2\text{O}_8$ was added. The samples were left to react for 20 min and then read at 510 nm.

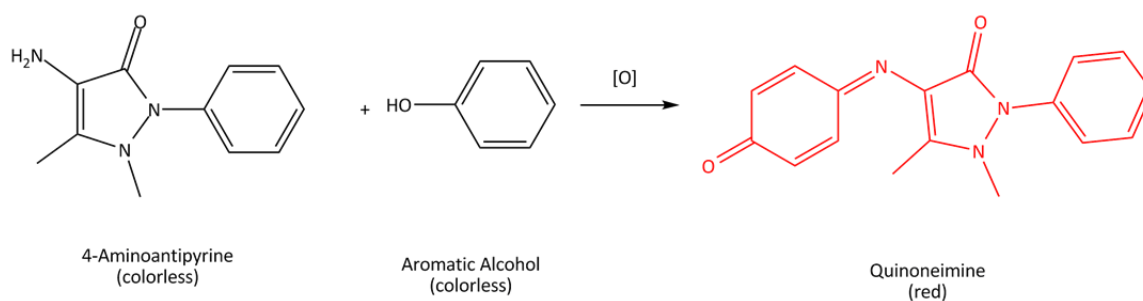


Figure 4.4. Reaction scheme for detection of aromatic alcohols. The red quinoneimine product is easily detected by eye and absorbs at 510 nm.

N. Chromatography

Reaction conditions were the same as those listed in section M, except after the 2 hr reaction time, 200 μL acetonitrile was added to quench the reactions. They were then spun down (for reactions with astemizole, verapamil, and LY294002) or concentrated and resuspended in water (for reactions with tramadol) before 25 μL was used on the HPLC or LCMS.

Analysis of the reactions of drug compounds verapamil, astemizole, LY294002, and tramadol with P450 variants was performed using a Supelco Discovery C18 column (2.1 x 150 mm, 3 μ m) from Sigma Aldrich (St. Louis, MO, USA) on a 2690 Separation module in conjunction with a 996 PDA detector from Waters (Milford, MA, USA). For verapamil, astemizole, LY294002 reactions, 25 μ L clarified reaction mixtures were analyzed with 0.2% formic acid (solvent A) and acetonitrile (solvent B) at the following conditions: 0-3 min, A:B 90:10; 3-25 min, linear gradient to A:B 30:70; 25-30 min, linear gradient to A:B 10:90. For tramadol reactions, 25 μ L concentrated reaction mixtures resuspended in water were analyzed with 15 mM phosphate buffer (solvent A) and acetonitrile (solvent B) at the following conditions: 0-3 min, A:B 100:0; 3-25 min, linear gradient to A:B 80:20; 25-30 min, linear gradient to A:B 10:90.

LCMS and MS/MS spectra were obtained with an LCQ Classic from ThermoFinnigan (San Jose, CA, USA) using identical conditions to the HPLC method detailed above for the LC conditions. The MS was operated in positive ESI mode, and the MS/MS spectra were acquired for the most abundant ions.

O. P450 Purification

For thermostability measurements, frozen culture pellets were thawed at room temperature for 1 hr, resuspended in 25 mM tris/HCl, pH 8.0, and sonicated. Crude extracts were filtered using 0.45 μ m syringe filters from Whatman (Kent, UK) and loaded onto HiTrap Q HP sepharose columns from GE Healthcare (Waukesha, WI, USA) which were equilibrated with 25 mM Tris/HCl, pH 8.0. Samples were eluted via step gradient

using an Akta purifier FPLC system (GE Healthcare) into 340 mM NaCl, 25 mM tris/HCl, pH 8.0 buffer. Samples were then desalted into 0.1 M phosphate buffer, pH 8.0 using regenerated cellulose 10,000 MWCO centrifugal filter units from Millipore (Billerica, MA, USA). Purified protein was flash frozen with liquid nitrogen and stored at -80 °C.

For crystallography experiments, frozen culture pellets were thawed at room temperature for 1 hr, resuspended in 100 mM NaCl, 20 mM imidazole, 20 mM tris/HCl, pH 8.0 buffer, and lysed by sonication with a Sonicator Heat Systems from Ultrasonic Systems, Inc (Haverhill, MA, USA) using 2 x 30 s, output control 5, 50% duty cycle. Crude extracts were centrifuged at 20,000 rpm for 30 min and filtered using 0.45 µm syringe filters (Whatman) and loaded onto Ni-NTA affinity columns (GE Healthcare) which were equilibrated with 100 mM NaCl, 20 mM imidazole, 20 mM tris/HCl, pH 8.0 buffer. After washing with five column volumes of 100 mM NaCl, 20 mM imidazole, 20 mM tris/HCl, pH 8.0 buffer, samples were eluted via step gradient using an Akta purifier FPLC system (GE Healthcare) into 100 mM NaCl, 300 mM imidazole, 20 mM tris/HCl, pH 8.0. Samples were then desalted into 25 mM tris/HCl, pH 8.0 using regenerated cellulose 10,000 MWCO centrifugal filter units (Millipore). Samples were loaded onto HiTrap Q HP sepharose columns (GE Healthcare) which were equilibrated with 25 mM Tris/HCl, pH 8.0 buffer. After washing with five column volumes of 25 mM Tris/HCl, pH 8.0 buffer, samples were eluted via step gradient using an Akta purifier FPLC system (GE Healthcare) into 340 mM NaCl, 25 mM tris/HCl, pH 8.0 buffer. Samples were then desalted into 25 mM HEPES, pH 7.0 buffer using regenerated cellulose 10,000 MWCO centrifugal filter units (Millipore) and concentrated to less than 1 mL. Samples were

loaded on a HiPrep 16/60 Sephacryl S-200 HR column (GE Healthcare) equilibrated with 25 mM HEPES, pH 7.0 buffer and separated by size exclusion. Purified protein was flash frozen with liquid nitrogen and stored at -80 °C. SDS-PAGE gel electrophoresis was used for quality control (Figure 4.5) and protein concentration was determined by CO binding as described in section M. Final yields were 10-30 mg protein/L culture.

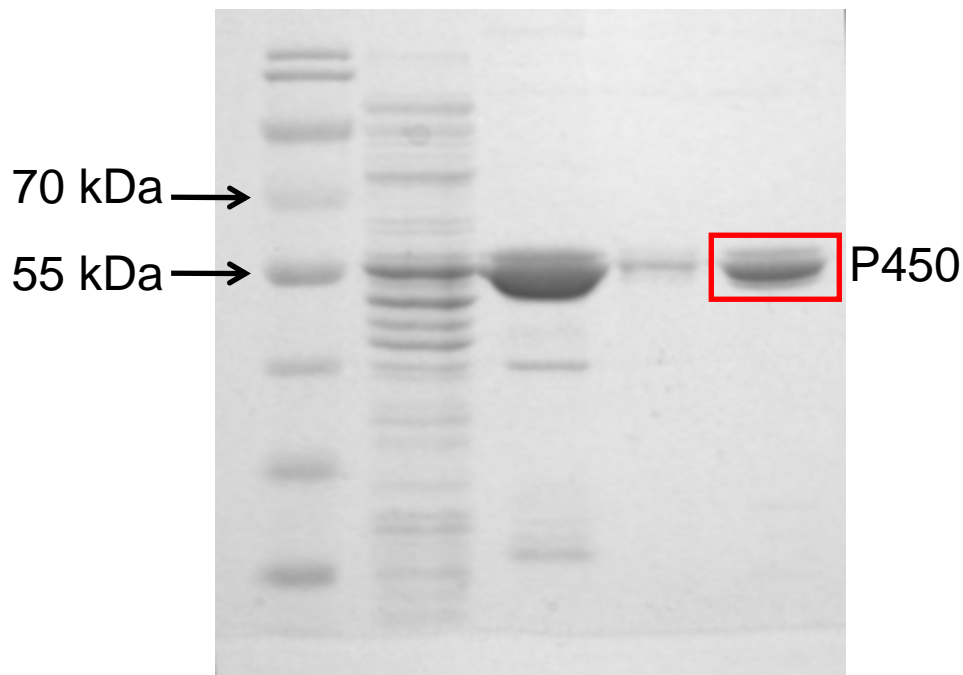


Figure 4.5. SDS-PAGE gel showing P450 sample after each purification step. The first column is the ladder, the second is the crude extract before purification, the third is the sample after affinity chromatography, the fourth is the sample after affinity chromatography and ion exchange, and the fifth is the final sample after affinity chromatography, ion exchange, and size exclusion.

P. Thermostability Measurements

The T_{50} value is defined as the temperature at which half of the protein unfolds after a ten-minute incubation in absence of substrate. For T_{50} measurements, purified protein was diluted to 1 μM in 0.1 M phosphate buffer and incubated at -4, 25, 75 $^{\circ}\text{C}$ and in a water bath across a range of temperatures bracketing the proteins T_{50} value. Water bath temperatures were measured using two different alcohol thermometers and observed to be consistent within 0.1 $^{\circ}\text{C}$. Protein concentration after incubation was determined by CO binding as described in section M and the T_{50} value was calculated by linear interpolation of data using Microsoft Excel.

At a given temperature, a half-life is defined as the incubation time in the absence of substrate that causes half of the protein to unfold. For half-life measurements, purified protein was diluted to 1 μM in 0.1 M phosphate buffer incubated in a water bath in triplicate. At time points between 0 and 3 hr, samples were removed and stored on ice. Protein concentration after incubation was determined by CO binding as described in section M and the half-lives were calculated with Microsoft Excel by fitting a straight line to a graph of the natural logarithm of the residual protein concentration vs. time.

Q. Protein Crystallization

Initial screening was done using the Macromolecular Crystallography Facility located in the Beckman Institute on the Caltech campus. Purified protein (see section O) at 30 mg/mL was screened against 480 crystallization conditions from the following

screens: Crystal Screen 1 and 2, Index, MembFac, and Peglon from Hampton Research (Aliso Viejo, CA, USA), Wizard 1 and 2 from Emerald Biosystems (Seattle, Washington, USA), and JCSG+ Suite (Qiagen).

In-house crystallization experiments were performed using the sitting drop vapor diffusion method. A 1:1 mixture of 15-30 mg/mL purified protein stock in 20 mM tris buffer, pH 8.0 and mother liquor was combined in 24 well sitting drop plates (Hampton Research). Mother liquor consisted of 0.1 M tris buffer, with pH 6-8, 1.5-3.0 M ammonium sulfate, and 0.1-0.3 M lithium sulfate. Additive Screen HT (Hampton Research) was added at 10% to the best conditions. Under normal conditions, crystal growth occurs over 28-35 days; however, growth times were significantly shortened to 3-4 days by microseeding with shards taken from these crystals. In addition, crystals grown under microseeding conditions grew larger and diffracted with improved resolution.

R. X-ray Data Collection and Protein Structure Determination

X-ray diffraction data were collected at the Stanford Synchrotron Radiation Lightsource beamline 12.2 on a Dectris Pilatus 6M detector. Data was collected at 100K and 1.033 Å. Diffraction datasets were integrated with XDS⁷ and scaled using SCALA⁸. Initial phases were determined using molecular replacement against WT structure taken from PDB 2IJ2, chain A⁹. Molecular replacement was accomplished using MOLREP software¹⁰, a component of the CCP4 crystallography software suite¹¹. Refinement was accomplished with iterative cycles of manual model building within COOT¹² and

automated refinement using REFMAC¹³ within CCP4. Final cycles of REFMAC refinement included refining TLS parameters. Noncrystallographic symmetry constraints were not utilized during refinement. Statistics for data collection and the final protein structure model are given in Table 3.5. Model quality was assessed using the “complete validation” tool included in the PHENIX software suite for automated structure determination¹⁴. Ramachandran outliers typically constituted ~1.0% of all residues, with favored Ramachandran representing >88.5% of all residues. All protein structure figures were generated using PyMOL software (The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.).

S. References

1. Higuchi, R., Krummel, B. & Saiki, R. K. (1988). A General-Method of Invitro Preparation and Specific Mutagenesis of DNA Fragments - Study of Protein and DNA Interactions. *Nucleic Acids Research* **16**, 7351-7367.
2. Nelson, N. (1944). A photometric adaptation of the Somogyi method for the determination of glucose. *Journal of Biological Chemistry* **153**, 375-380.
3. Omura, T. & Sato, R. (1964). Carbon Monoxide-Binding Pigment of Liver Microsomes .I. Evidence for Its Hemoprotein Nature. *Journal of Biological Chemistry* **239**, 2370-&.
4. Otey, C. R. (2003). High-Throughput Carbon Monoxide Binding Assay. In *Directed Enzyme Evolution: Screening and Selection Methods* (Arnold, F. H. & Georgiou, F., eds.), pp. 137-139. Humana Press, Totowa, New Jersey.
5. Hopps, H. B. (2000). Purpald (R): A reagent that turns aldehydes purple! *Aldrichimica Acta* **33**, 28-30.
6. Otey, C. R. & Joern, J. M. (2003). High-Throughput Screen for Aromatic Hydroxylation. In *Directed Enzyme Evolution: Screening and Selection Methods* (Arnold, F. H. & Georgiou, G., eds.), pp. 141-148. Humana Press, Totowa, New Jersey.
7. Kabsch, W. (2010). Xds. *Acta Crystallographica Section D-Biological Crystallography* **66**, 125-132.
8. Evans, P. (2006). Scaling and assessment of data quality. *Acta Crystallographica Section D-Biological Crystallography* **62**, 72-82.
9. Girvan, H. M., Seward, H. E., Toogood, H. S., Cheesman, M. R., Leys, D. & Munro, A. W. (2007). Structural and spectroscopic characterization of P450BM3 mutants with unprecedented P450 heme iron ligand sets - New heme ligation states influence conformational equilibria in P450BM3. *Journal of Biological Chemistry* **282**, 564-572.

10. Vagin, A. & Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. *Journal of Applied Crystallography* **30**, 1022-1025.
11. Bailey, S. (1994). The Ccp4 Suite - Programs for Protein Crystallography. *Acta Crystallographica Section D-Biological Crystallography* **50**, 760-763.
12. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D-Biological Crystallography* **60**, 2126-2132.
13. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D-Biological Crystallography* **53**, 240-255.
14. Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D-Biological Crystallography* **66**, 213-221.