Non-contiguous Protein Recombination

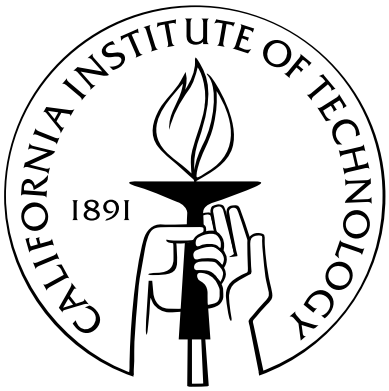
Thesis by

Matthew Alexander Smith

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2013

(Defended March 7th, 2013)

Acknowledgements

I have thoroughly enjoyed my time as a graduate student at Caltech. I am grateful to all the talented people who make up this stimulating and immersive community and I feel very fortunate to have been a part of it. The past four and a half years have been filled with the desire to learn and experience new things.

I would like to begin by thanking my advisor, Frances Arnold. Frances has given me the time and space to explore my interests in protein science and search for creative solutions to interesting problems. At the same time, Frances has always been available to offer constructive advice and she has instilled in me a practical perspective to research which will be invaluable in many aspects of my future career.

I thank my thesis committee members, Rudy Marcus, Rob Phillips, and Dave Tirrell for their advice and support. I thank Niles Pierce and Linda Scott for running a fantastic bioengineering program and for all their help. I would also like to thank the Resnick Institute for supporting me financially and in particular the director of the Resnick Institute, Neil Fromer, for his support over the last couple of years.

I have been fortunate to work with many exceptional people in the Arnold lab and I appreciate all the help I have been given over the last four and a half years. I specifically thank Andrea Rentmeister who mentored me when I started in the lab and taught me most of the experimental techniques in molecular biology that I know today. Tim Wu has assisted

me with experimental work for the past four years as an undergraduate researcher and I am grateful for his willingness to learn and help out. I also thank Claire Bedbrook who has been a great help working with me on the cellobiohydrolase I project.

I am grateful to both Phil Romero and Chris Snow for their help with the computational aspects of my research and for many stimulating discussions on protein engineering. I thank Eric Brustad for teaching me protein crystallography and his assistance with the β-glucosidase project. I also thank Indira Wu and Florence Mingardon for guidance on cellulase engineering and Sabine Brinkmann-Chen for her help as lab manager. I would also like to thank JD Bagert for many coffee breaks to bounce ideas off one another and for his close friendship over the past four and a half years.

Finally, I would like to thank my family - my parents Ian and Barbara, my brother Dugal, and my fiancée Mary for their love, support, and guidance.

Abstract

Swapping sequence elements among related proteins can produce chimeric proteins with novel behaviors and improved properties such as enhanced stability. Although homologous mutations are much more conservative than random mutations, chimeras of distantly-related proteins have a low probability of retaining fold and function. Here, I introduce a new tool for protein recombination that identifies structural blocks that can be swapped among homologous proteins with minimal disruption. This non-contiguous recombination approach enables design of chimeras and libraries of chimeras with less disruption than can be achieved by swapping blocks of sequence. Less disruption means that one can generate libraries with higher fractions of functional enzymes and enables recombination of more distant homologs.

Using this new tool I design and construct many functional chimeric cellulases. I illustrate the structurally conservative nature of this recombination by creating a functional prokaryotic-eukaryotic chimera and solving its structure. I also show how non-contiguous recombination can be used to efficiently identify stabilizing mutations that have been incorporated into homologs in nature.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction: Site-directed protein recombination

1.1 Introduction

The large diversity of natural proteins provides many highly optimized sequences that encode specific functions. How can we use nature's solutions to engineer proteins with improved properties or novel functions? One approach is to transfer sequence elements and their functions from one protein to another. This chapter focuses on methods and applications of site-directed recombination that generate new functions or enhance particular characteristics in a protein of interest.

Early work on DNA shuffling of homologous genes enabled the quick construction of libraries of novel sequences that are combinations of the parental genes [1]. This is a simple, effective way to produce diverse libraries of proteins for directed evolution [2, 3, 4, 5]. Many of these recombined sequences are functional because mutations from homologous sequences are conservative [6, 7]. More recently there has been substantial work recombining specific sequence elements between more distantly-related proteins (Figure 1.1), producing chimeras with hybrid properties. This site-directed recombination presents an opportunity to rapidly engineer proteins when a desired property is known to exist in another sequence.

The first part of this chapter outlines available computational tools for identifying swappable protein fragments that will generate folded, functional proteins. I follow this with several studies that have recovered function and stability in chimeric proteins by mutating residues at fragment interfaces. The second part of the chapter provides recent examples that have used site-directed recombination to introduce or enhance specific protein functions. Finally, I look at a few examples that recombine distantly-related or non-homologous proteins and I discuss the potential of this strategy for engineering proteins with new functions.

1.2 Swapping sequence elements while preserving protein function

It is easy to shuffle homologous gene sequences at defined points, and there are a number of published methods to do so [8, 9, 10, 11, 12, 13]. However for homologs with a DNA sequence identity below 70% it is difficult to swap sequence elements and retain protein function and stability. Domains and small structural pieces such as loops can often be easily identified from a protein structure as suitable fragments to recombine. For swapping subdomain protein fragments, several labs have developed scoring functions that try to quantify the likelihood a given chimera will be functional. Residues can then be grouped into swappable blocks based on optimizing one of these metrics.

One method pioneered by Voigt *et al.* uses structural information to break up homologous sequences at specific crossover points [14]. SCHEMA scores a chimera based on how many native residue-residue contacts are disrupted. Less disruption increases the probability a chimera will fold and function [15]. Optimal crossover points can be found that minimize the average disruption to a library of chimeras [16], enriching a library in functional sequences. Rather than swapping contiguous elements of sequence, a recent approach identifies fragments of structure that minimize SCHEMA disruption upon recombination [17].

Bailey-Kellogg and colleagues have published a similar method that accounts for higher-order multi-residue interactions within chimeras [18]. If structural information is not available, an alternative scoring function developed by the Maranas lab uses information from a multiple sequence alignment to count conserved residue-residue properties that deviate in a chimera [19].

1.3 Swapping sequence elements and recovering protein function

Site-directed recombination often does not produce functional proteins. While computational methods for choosing crossovers (see previous section) can improve the probability the progeny proteins are functional, it is still a challenge to swap sequence elements between distantly-related proteins and retain high levels of protein activity. Several labs have recently explored mutating the interfaces between fragments of protein structure to recover chimera function.

Work from the Koide lab improved the binding affinity of an Erbin PDZ binding domain fused to a fibronectin type III domain more than 500-fold towards a specific peptide sequence by mutating residues on the interface [20]. Inspired by this, Zhou *et al.* mutated interface residues of their insoluble acylpeptide hydrolase / carboxylesterase chimera [21]. Changing seven newly exposed hydrophobic residues to hydrophilic residues resurrected the chimera's solubility and activity. Similarly, Geitner *et al.* introduced disulfide bonds at linker regions between domains to rescue the activity, solubility, and stability of their isomerase-chaperone chimeras [22].

Hoi *et al.* improved the brightness and Ca^{2+} response of a chimeric Ca^{2+}-dependent fluorescent protein by mutating residues between the chimera's Ca^{2+} indicator domain and its photoconvertible fluorescent protein domain [23]. Additional random mutagenesis of the chimeric gene and screening produced a final protein that exhibits a 4.6-fold increase in fluorescence upon binding Ca^{2+} which could be valuable for studying Ca^{2+} signaling.

In a follow-up experiment to building a chimera from different protein folds [24], Eisenbeis et al. used computationally guided mutations (Rosetta) to improve the stability and

adjust the fold of their chimera [25]. Interestingly, the five most favorable mutations selected by Rosetta were all localized to the interface between the parental fragments (Figure 1.2). Introducing these mutations stabilized the chimera, improved solubility and corrected the $(\beta\alpha)_8$-barrel fold. Furthermore, two additional mutations improved the chimera's phosphate binding 10-fold.

Considering the importance of residues on the interface between swapped sequence elements, it may be beneficial to design recombination experiments to allow amino acid variability at these regions. Ochoa-Leyva *et al.* explored varying residues either side of inserted loops [26]. Seven loops (3 homologous, 3 non-homologous and 1 computationally designed) were each tested as a replacement for loop 6 of N-(5'-phosphoribosyl)anthranilate isomerase from *E. coli* (ecTrpF). As one might expect, homologous loop insertions were much more likely to result in functional chimeras. However, functional ecTrpF variants were found for all seven loop replacements, demonstrating the value of recombination designs that accommodate mutations at interface regions.

1.4 Swapping sequence elements to transfer protein function

Many recent examples have emerged that use site-directed recombination to transfer specific properties, including stability, substrate specificity, and allostery between two or more proteins. This strategy offers a fast way to engineer desirable properties into a protein of interest by borrowing solutions from other proteins in nature (Figure 1.3).

Thermostable enzymes are desirable for many industrial applications and there have been many examples in which site-directed recombination has led to stabilized enzymes. Clusters of residues are known to contribute additively to thermal stability [7, 27] and in SCHEMA site-directed recombination libraries, the stabilities of chimeric proteins can

be accurately predicted by simply summing the stability contributions from each block of sequence [28]. This approach can be used to engineer highly stable proteins by piecing together stabilizing protein fragments from homologous parents. In addition, testing the individual mutations within a stable block of sequence can reveal highly stabilizing amino acid substitutions [29]. More recently, Heinzelman *et al.* presented a useful strategy to identify stabilizing pieces of sequence from proteins with poor heterologous expression by substituting one piece at a time into a well-expressed homolog [30].

An alternative approach to generate stable enzymes with a specific activity is to engineer the desired activity in a highly stable homolog. Campbell *et al.* replaced three substrate-binding loops from *P. furiosus* alcohol dehydrogenase D (AdhD) with those from a human aldose reductase (hAR) homolog [31]. This chimera retains the extreme thermostability of AdhD (elevated activity at 100°C) and was able to reduce DL-glyceraldehyde, a model substrate for hAR, albeit with a three orders of magnitude lower catalytic efficiency. Interestingly, while AdhD primarily used NAD(H) as a cofactor, the chimera, like hAR, had a strong preference for NADP(H), indicating that the enzyme cofactor was also switched upon recombination.

Similarly, van Beek *et al.* recombined a thermostable phenylacetone monooxygenase (PAMO) and two homologs with broader substrate specificities, a cyclohexanone monooxygenase (CHMO) and a steroid monooxygenase (STMO) [32]. These Baeyer-Villiger monooxygenases (BVMOs) are potential industrial biocatalysts. However, PAMO is the only known stable BVMO and it accepts a narrow range of substrates. Guided by the structure, the authors replaced a subdomain from PAMO with the corresponding elements from CHMO and STMO. These two chimeras had higher stabilities than the parents CHMO and STMO and exhibited broad substrate specificities. Not all parental activities were present in

the chimeras, but on some substrates chimeras had improved catalytic properties over the parental enzymes. In addition, PAMO was recombined with a putative BVMO gene from a metagenomic sample. This third chimera exhibited effective oxidation of substrates that are poorly converted by the other chimeric and parental monooxygenases, and is a good example of using protein recombination to explore metagenomic sequences.

Jones shuffled six specific loop regions between seven serine proteases from the subtilisin family [33]. Regions of sequence were selected for their known functional importance in substrate binding, metal ion binding and catalysis, and the chimeric proteases displayed novel specificities for a range of peptide substrates. Chen *et al.* altered the product specificity of cytochrome P450 CYP102A1 on farnesol to produce 12-hydroxyfarnesol by exchanging small regions of sequence involved in substrate recognition with P450 CYP4C7, which naturally produces 12-hydroxyfarnesol [34].

There are several examples of engineering allosteric interactions through chimeragenesis. Most notably, the Lim lab generated chimeric regulatory proteins that were activated by different input ligands [35]. Ostermeier and colleagues created a chimera from a maltose binding protein and a beta-lactamase that binds Zn^{2+}, and this binding switches off enzymatic beta-lactam activity [36]. More recently, Duret *et al.* transferred allostery between distantly-related proteins by combining the extracellular domain (ECD) of a bacterial ligand-gated ion channel (LGIC) with the transmembrane domain (TMD) of the human α1 glycine receptor (α1GlyR) [37]. The chimera, like the bacterial LGIC, was activated by protons but exhibited a similar allosteric regulation to the human α1GlyR. Additionally, Cross *et al.* recombined two distant homologs of the first shikimate pathway enzyme (DAH7PS) [38]. By transposing the regulatory domain of a tyrosine-regulated bacterial DAH7PS onto the catalytic domain of an unregulated archaeal DAH7PS, the authors created a chimera

that was inhibited by tyrosine.

1.5 Swapping sequence elements can probe sequence-function relationships

Swapping defined sequence fragments can improve our understanding of sequence-function relationships and guide engineering efforts. Romero *et al.* recently introduced a general strategy for modeling protein characteristics that uses Gaussian processes to predict the protein fitness landscape from experimental data (Figure 1.4) [39]. A Gaussian process landscape identified cytochrome P450s that are more thermostable than any previously engineered, demonstrating that such models can accelerate searches through sequence space for desirable proteins.

Protein recombination has also been used to probe protein biochemistry. Patel *et al.* explored melanocortin receptor selectivity with the antagonistic ligands Agouti-related protein (AgRP) and agouti signaling protein (ASIP) by methodically swapping sequence elements between these two homologs and measuring binding to three melanocortin receptors [40]. Recombination revealed that binding to one of these receptors was dependent on six amino acids present in ASIP but not in AgRP. Ohtomo *et al.* investigated β-lactoglobulin dimerization through chimeragenesis of the monomeric equine β-lactoglobulin (ELG) with the dimeric bovine β-lactoglobulin (BLG) [41]. While swapping nine mutations of the BLG dimer interface did not make ELG dimerize, a chimera with BLG secondary structure did dimerize indicating that secondary structure is important for β-lactoglobulin dimerization. Other recent biochemical relationships elucidated by protein recombination include a correlation between a protein's isoelectric point and its stability [42] and a higher tolerance to

thermal denaturation leads to catalysis at elevated temperatures [43].

1.6 Pushing the limits of protein recombination

Distantly-related homologs are more likely to have divergent properties than closely-related sequences, and recombination of these diverse proteins can provide quick access to new properties. Proteins can be engineered by transferring desirable properties from distant homologs provided one can preserve function when exchanging elements of sequence.

Recent work has produced chimeras constructed from proteins in different kingdoms of life by swapping domains [37, 44], shuffling substrate binding loops [31], and through exchanging structural elements within a protein domain [17]. These chimeras acquire properties from their distantly-related parents: switched cofactor specificities [31], new allostery [38], and altered backbone conformations [17]. Interestingly, in the latter case the fragments of structure that make up the chimera maintain the backbone conformations found in their respective parental structures.

Pushing the limits of protein recombination further, there has been substantial work on generating chimeras of unrelated proteins. Several recent examples are discussed here. For older examples we refer the reader to an excellent review on recombining non-homologous domains [45].

Edwards *et al.* fused a heme-binding cytochrome b562 to the middle of a β-lactamase and produced a number of active chimeras with a novel allosteric property [46]. Cytochrome b562 undergoes a significant structural rearrangement upon binding heme and Edwards *et al.* use this property to disrupt the structure of the β-lactamase chimeras. Several of the chimeric β-lactamases exhibited over a 100-fold decrease in activity upon the addition of heme.

Rather than combining whole domains, Shanmugaratnam *et al.* built a highly stable chimera from two proteins with different folds and no obvious substructure [24, 47]. Part of a $(\beta\alpha)_8$-barrel imidazole glycerol phosphate synthase (HisF) was swapped with three proteins that have a flavodoxin-like fold: a chemotaxis response regulator (CheY), a nitrogen response regulator (NarL) and a methylmalonyl CoA mutase (MMCoA). While the MMCoA-HisF chimera was insoluble, both the CheY-HisF and NarL-HisF chimeras are soluble and stable suggesting that this recombination is somewhat generalizable. Although the chimeras have no known catalytic activity, the ability to produce a novel protein fold using recombination opens up a new avenue of exploration for protein engineers.

1.7 Summary

Site-directed recombination can transfer properties between proteins and generate chimeras with novel functions and enhanced characteristics, including increased thermostability, altered substrate specificity, switched cofactor specificity, and new allosteric interactions. Computational tools are available to guide recombination; scoring functions and design algorithms reduce structural disruption to chimeras and sequence-based models help identify protein fragments that encode desirable protein properties. Recombination will continue to expand the variety of chimeric proteins with novel properties and directed evolution will be increasingly used to optimize these new functions [48].

1.8 This work

In this thesis I study site-directed recombination by constructing many functional chimeric cellulases. Chapters 2 and 3 describe swapping contiguous elements of protein sequence and

how this can be used to identify desirable protein fragments and probe protein biochemistry. In chapters 4, 5, and 6, I present a new computational tool for identifying structural blocks that can be swapped among homologous proteins with minimal disruption and I offer several examples of enzyme engineering using this approach.

1.9 Figures

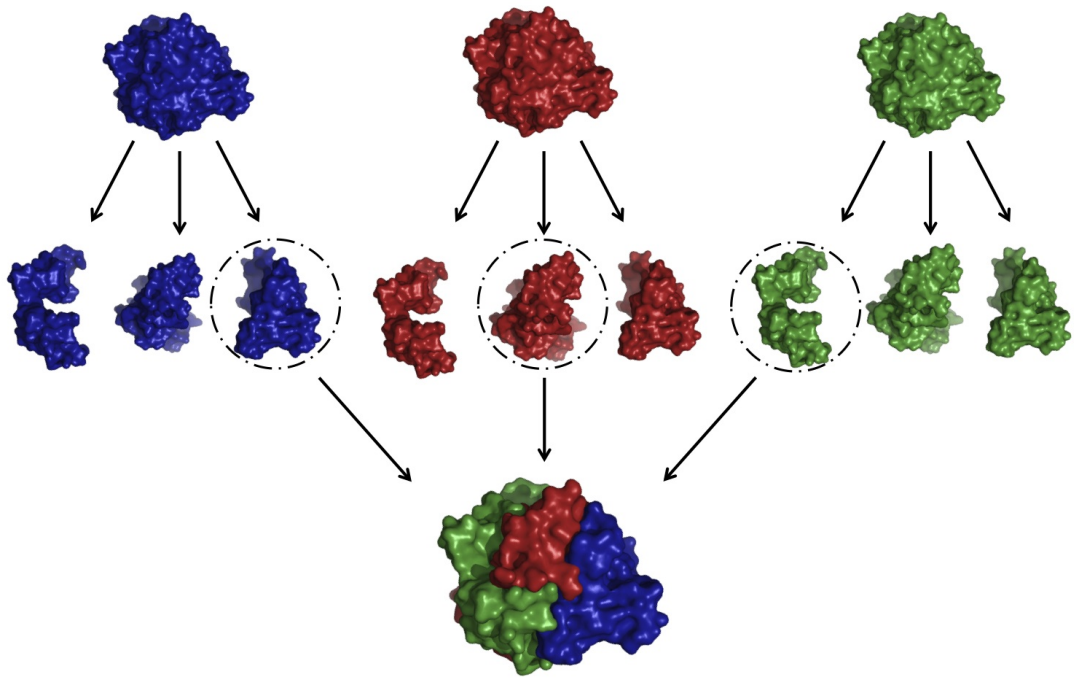


Figure 1.1: Site-directed homologous recombination. Two or more proteins are fragmented into well-defined pieces. These protein fragments are recombined to form chimeric proteins.

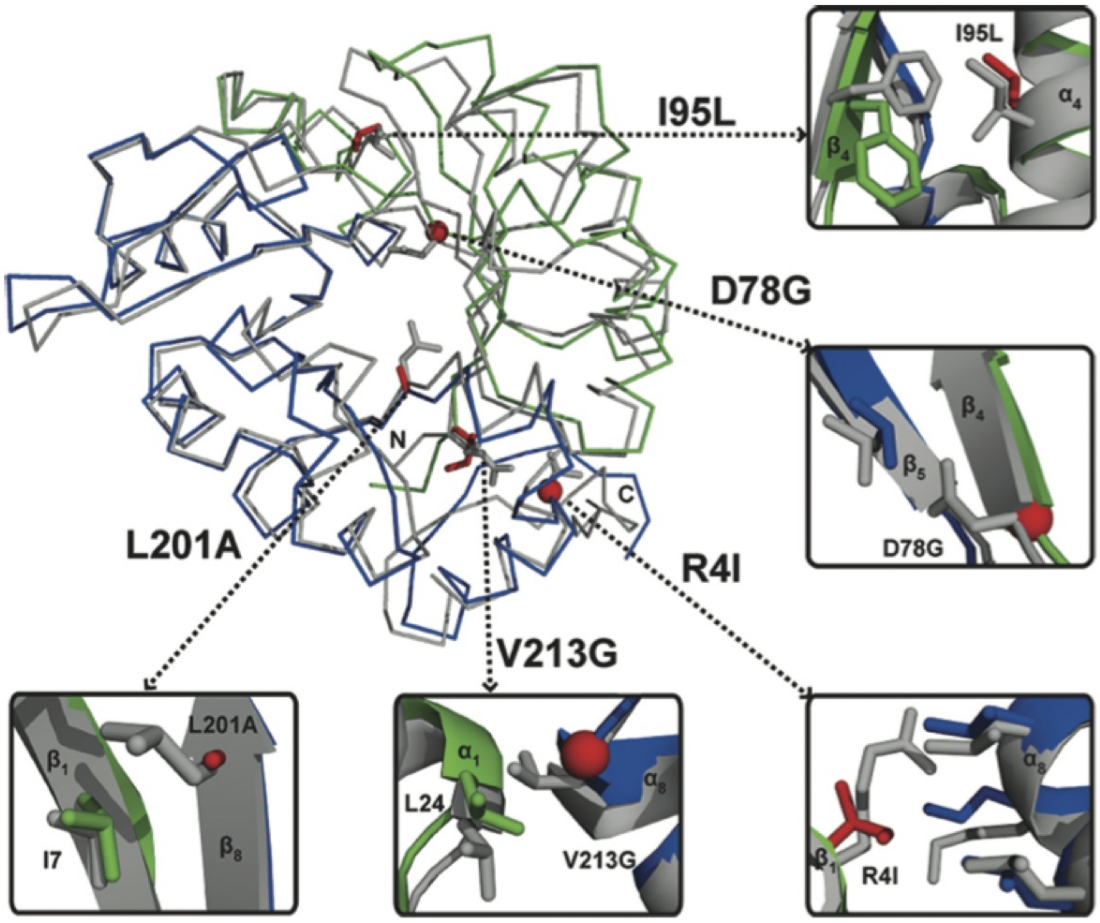


Figure 1.2: Interfacial mutations stabilize a chimera made from fragments of different folds. The chimera is made from fragments of a response regulator CheY (green) and an imidazole glycerol phosphate synthase HisF (blue). Five stabilizing mutations predicted by Rosetta are highlighted in red and a model of the structure is in gray. Reproduced with permission from reference [25].

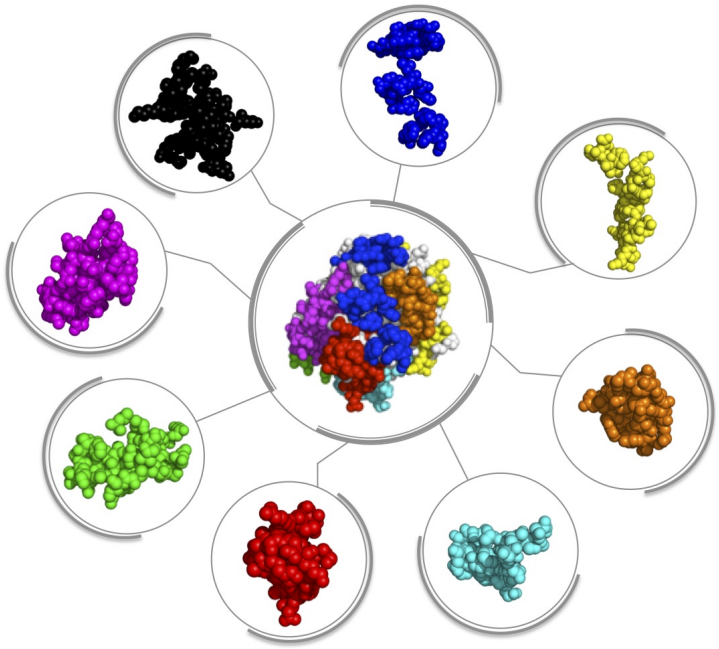


Figure 1.3: Different protein fragments can be responsible for different protein functions. The ability to identify and successfully recombine these fragments enables the transfer of function between proteins. This approach can rapidly engineer improved protein properties and novel functions.

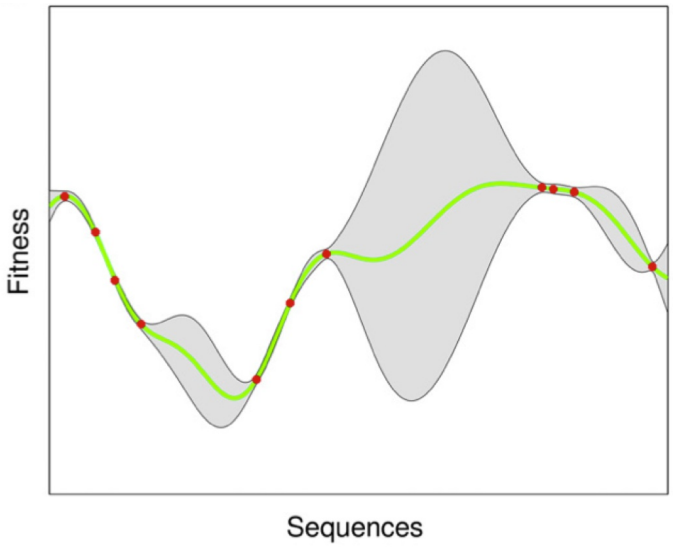


Figure 1.4: A 1-dimensional example of a protein fitness landscape, predicted from experimental data using Gaussian processes. Experimental measurements are represented by red dots, the green line illustrates the Gaussian process models mean predictions, and the model's 95% confidence intervals are in gray. Predictions close to experimental data have much lower uncertainties. Figure reproduced from reference [40]. Copyright 2012 National Academy of Sciences, U.S.A.

1.10 References

1. Crameri A., Raillard S. A., Bermudez E., and Stemmer W. P. C. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288-291.

2. Goldsmith M., and Tawfik D. S. (2012) Directed enzyme evolution: beyond the low-hanging fruit. *Curr. Opin. Struc. Biol.* **22**, 406-412.

3. Dalby P. A. (2011) Strategy and success for the directed evolution of enzymes. *Curr. Opin. Struc. Biol.* **21**, 473-480.

4. Tracewell C. A., and Arnold F. H. (2009) Directed enzyme evolution: climbing fitness peaks one amino acid at a time. *Curr. Opin. Chem. Biol.* **13**, 3-9.

5. Romero P. A., and Arnold F. H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866-876.

6. Drummond D. A. (2005) On the conservative nature of intragenic recombination. *Proc. Natl. Acad. Sci. USA* **102**, 5380-5385.

7. Romero P. A., and Arnold F. H. (2012) Random field model reveals structure of the protein recombinational landscape. *PLoS Comput. Biol.* **8**, e1002713.

8. Hiraga K., and Arnold F. H. (2003) General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.* **330**, 287-296.

9. Farrow M. F., and Arnold F. H. (2010) Combinatorial recombination of gene fragments to construct a library of chimeras. *Curr. Protoc. Protein Sci.* **62**, 26.2.1-26.2.20.

10. Engler C., and Marillonnet S. (2011) Generation of families of construct variants using golden gate shuffling. In: *Methods Mol. Biol.*, Humana Press, pp. 167-181.

11. Villiers B. R. M., Stein V., and Hollfelder F. (2009) USER friendly DNA recombination (USERec): a simple and flexible near homology-independent method for gene library construction. *Protein Eng. Des. Sel.* **23**, 1-8.

12. O'Maille P. E., Bakhtina M., and Tsai M.-D. (2002) Structure-based combinatorial protein engineering (SCOPE). *J. Mol. Biol.* **321**, 677-691.

13. Dokarry M., Laurendon C., and O'Maille P. E. (2012) Automating gene library synthesis by structure-based combinatorial protein engineering: examples from plant sesquiterpene synthases. *Meth. Enzymol.* **515**, 21-42.

14. Voigt C. A., Martinez C., Wang Z.-G., Mayo S. L., and Arnold F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553-558.

15. Meyer M., Hochrein L., and Arnold F. H. (2006) Structure-guided SCHEMA recombination of distantly related β-lactamases. *Protein Eng. Des. Sel.* **19**, 563-570.

16. Endelman J., Silberg J., Wang Z.-G., and Arnold F. H. (2004) Site-directed protein recombination as a shortest-path problem. *Protein Eng. Des. Sel.* **17**, 589-594.

17. Smith M. A., Romero P. A., Wu T., Brustad E. M., and Arnold F. H. (2013) Chimeragenesis of distantly-related proteins by noncontiguous recombination. *Protein Sci.* **22**, 231-238.

18. Ye X., Friedman A. M., and Bailey-Kellogg C. (2007) Hypergraph model of multi-residue interactions in proteins: sequentially-constrained partitioning algorithms for optimization of site-directed protein recombination. *J. Comput. Biol.* **14**, 777-790.

19. Saraf M. C., Horswill A. R., Benkovic S. J., and Maranas C. D. (2004) FamClash: a method for ranking the activity of engineered enzymes. *Proc. Natl. Acad. Sci. USA* **101**, 4142-4147.

20. Huang J., Koide A., Makabe K., and Koide S. (2008) Design of protein function leaps by directed domain interface evolution. *Proc. Natl. Acad. Sci. USA* **105**, 6578-6583.

21. Zhou X., Wang H., Zhang Y., Gao L., and Feng Y. (2012) Alteration of substrate specificities of thermophilic α/β hydrolases through domain swapping and domain interface optimization. *Acta Biochim. Biophys. Sin. (Shanghai)* **44**, 965-973.

22. Geitner A.-J., and Schmid F. X. (2012) Combination of the human prolyl isomerase FKBP12 with unrelated chaperone domains leads to chimeric folding enzymes with high activity. *J. Mol. Biol.* **420**, 335-349.

23. Hoi H., Matsuda T., Nagai T., and Campbell R. E. (2013) Highlightable Ca(2+) indicators for live cell imaging. *J. Am. Chem. Soc.* **135**, 46-49.

24. Bharat T. A. M., Eisenbeis S., Zeth K., and Hocker B. (2008) A $\beta\alpha$-barrel built by the combination of fragments from different folds. *Proc. Natl. Acad. Sci. USA* **105**, 9942-9947.

25. Eisenbeis S., Proffitt W., Coles M., Truffault V., Shanmugaratnam S., Meiler J., and Hocker B. (2012) Potential of fragment recombination for rational design of proteins. *J. Am. Chem. Soc.* **134**, 4019-4022.

26. Ochoa-Leyva A., Barona-Gomez F., Saab-Rincon G., Verdel-Aranda K., Sanchez F., and Soberon X. (2011) Exploring the structure-function loop adaptability of a $(\beta/\alpha)_8$-barrel enzyme through loop swapping and hinge variability. *J. Mol. Biol.* **411**,

143-157.

27. LeMaster D. M., and Hernandez G. (2005) Additivity in both thermodynamic stability and thermal transition temperature for rubredoxin chimeras via hybrid native partitioning. *Structure* **13**, 1153-1163.

28. Li Y., Drummond D. A., Sawayama A. M., Snow C. D., Bloom J. D., and Arnold F. H. (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25**, 1051-1056.

29. Heinzelman P. , Snow C. D., Smith M. A., Yu X., Kannan A., Boulware K., Villalobos A., Govindarajan S., Minshull J., and Arnold F. H. (2009) SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *J. Biol. Chem.* **284**, 26229-26233.

30. Heinzelman P., Komor R., Kanaan A., Romero P. A., Yu X., Mohler S., Snow C. D., Arnold F. H. (2010) Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Eng. Des. Sel.* **23**, 871-880.

31. Campbell E., Chuang S., and Banta S. (2012) Modular exchange of substrate-binding loops alters both substrate and cofactor specificity in a member of the aldo-keto reductase superfamily. *Protein Eng. Des. Sel.*, doi:10.1093/protein/gzs095.

32. van Beek H. L., Gonzalo G., Fraaije M. W. (2012) Blending Baeyer-Villiger monooxygenases: using a robust BVMO as a scaffold for creating chimeric enzymes with novel catalytic properties. *Chem. Commun.* **48**, 3288-3290.

33. Jones D. D. Recombining low homology, functionally rich regions of bacterial subtilisins by combinatorial fragment exchange. *PLoS ONE* **6**, e24319.

34. Chen C.-K. J., Berry R. E., Shokhireva T. K., Murataliev M. B., Zhang H., and Walker F. A. (2010) Scanning chimeragenesis: the approach used to change the substrate selectivity of fatty acid monooxygenase CYP102A1 to that of terpene omega-hydroxylase CYP4C7. *J. Biol. Inorg. Chem.* **15**, 159-174.

35. Dueber J. E., Yeh B. J., Chak K., and Lim W. A. (2003) Reprogramming control of an allosteric signaling switch through modular recombination. *Science* **301**, 1904-1908.

36. Liang J., Kim J. R., Boock J. T., Mansell T. J., and Ostermeier M. (2007) Ligand binding and allostery can emerge simultaneously. *Protein Sci.* **16**, 929-937.

37. Duret G., Van Renterghem C., Weng Y., Prevost M., Moraga-Cid G., Huon C., Sonner J. M., and Corringer P.-J. (2011) Functional prokaryotic-eukaryotic chimera from the pentameric ligand-gated ion channel family. *Proc. Natl. Acad. Sci. USA* **108**, 12143-12148.

38. Cross P. J., Allison T. M., Dobson R. C. J., Jameson G. B., and Parker E. J. (2013) Engineering allosteric control to an unregulated enzyme by transfer of a regulatory domain. *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.1217923110.

39. Romero P. A., Krause A., and Arnold F. H. (2012) Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.1215251110.

40. Patel M. P., Cribb Fabersunne C. S., Yang Y.-K., Kaelin C. B., Barsh G. S., and Millhauser G. L. (2010) Loop-swapped chimeras of the agouti-related protein and

the agouti signaling protein identify contacts required for melanocortin 1 receptor selectivity and antagonism. *J. Mol. Biol.* **404**, 45-55.

41. Ohtomo H., Konuma T., Utsunoiya H., Tsuge H., and Ikeguchi M. (2011) Structure and stability of Gyuba, a β-lactoglobulin chimera. *Protein Sci.* **20**, 1867-1875.

42. Romero P. A., Stone E., Lamb C., Chantranupong L., Krause A., Miklos A., Hughes R., Fechtel B., Ellington A. D., Arnold F. H., and Georgiou G. (2012) SCHEMA-designed variants of human arginase I and II reveal sequence elements important to stability and catalysis. *ACS Synth. Biol.* **1**, 221-228.

43. Smith M. A., Rentmeister A., Snow C. D., Wu T., Farrow M. F., Mingardon F., and Arnold F. H. (2012) A diverse set of family 48 bacterial glycoside hydrolase cellulases created by structure-guided recombination. *FEBS J.* **279**, 4453-4465.

44. Goyal R., Salahudeen A. A., and Jansen M. (2011) Engineering a prokaryotic cys-loop receptor with a third functional domain. *J. Biol. Chem.* **286**, 34635-34642.

45. Koide S. (2009) Generation of new protein functions by nonhomologous combinations and rearrangements of domains and modules. *Curr. Opin. Biotechnol.* **20**, 398-404.

46. Edwards W. R., Williams A. J., Morris J. L., Baldwin A. J., Allemann R. K., and Jones D. D. (2010) Regulation of beta-lactamase activity by remote binding of heme: functional coupling of unrelated proteins through domain insertion. *Biochemistry* **49**, 6541-6549.

47. Shanmugaratnam S., Eisenbeis S., and Hocker B. (2012) A highly stable protein chimera built from fragments of different folds. *Protein Eng. Des. Sel.* **25**, 699-703.

48. Brustad E. M., and Arnold F. H. (2011) Optimizing non-natural protein function with directed evolution. *Curr. Opin. Chem. Biol.* **15**, 201-210.

Chapter 2

Designing libraries of chimeric proteins using SCHEMA recombination and RASPP

2.1 Summary

SCHEMA is a method for designing libraries of novel proteins by recombination of homologous sequences. The goal is to maximize the number of folded proteins, while simultaneously generating significant sequence diversity. Here, we use the RASPP algorithm to identify optimal SCHEMA designs for shuffling contiguous elements of sequence. Our design recombines 5 fungal cellobiohydrolases (CBH1s) to produce a library of more than 390,000 novel CBH1 sequences.

2.2 Introduction

SCHEMA recombination shuffles sequence elements (blocks) defined by a set of crossover locations in homologous proteins to generate novel chimeric proteins [1] (see Figure 2.1). Despite that fact that homologous mutations are more conservative than random mutations, a chimera containing many mutations is less likely to be functional than one closer in sequence to one of its parent proteins. SCHEMA recombination seeks to maximize the probability that a library of chimeric proteins will be functional by using structural information to pick crossover locations that minimize disruption of the folded structure. Our metric for disruption is the number of non-native residue-residue contacts, which we refer to as a chimeras SCHEMA energy (E). Minimizing the average SCHEMA energy ($< E >$) of all the chimeras in a library increases the fraction of functional chimeras [2]. For sequence elements that are contiguous along the polypeptide chain, we developed the RASPP [3] computational tool to identify crossovers that minimize $< E >$.

Because chimeric proteins retain sequence elements (e.g. catalytic residues) that are shared among the parents, properly folded chimeras usually retain the overall function of

the parents. The new combinations of amino acids in other parts of the protein, however, can lead to significant changes in key properties such as stability [4, 5], expression level [6], or substrate specificity [7]. By analyzing a subset of the possible chimera sequences we can build predictive models and identify the chimeras having useful changes in those properties [8].

In this chapter, we design a SCHEMA library that recombines 5 fungal cellobiohydrolases (CBH1s). We use RASPP to identify optimal libraries having 7 crossover sites (8 blocks). Shuffling these blocks among the 5 homologs generates a recombination library of $5^8 = 390,625$ possible sequences. We previously designed a very similar library [6], and analysis of a subset of chimeras led us to identify chimeric CBH1s that are more stable than any of the 5 parents.

2.3 Materials

1. A Unix-based computer that can run python scripts (*see* **Note 1**). Python can be downloaded from: http://www.python.org/download/

2. Download and unpack the RASPP toolbox. This is available from: http://cheme.che.caltech.edu/groups/fha/media/schema-tools.zip

3. A multiple sequence alignment of the parental sequences that are to be recombined (*see* **Note 2**). This alignment should be in ALN format (such as that produced by ClustalW), without a header (*see* **Note 3**). As recombination parents, we picked the CBH1 sequences from *C. thermophilum*, *T. aurantiacus*, *H. jecorina*, *A. thermophilum*, and *T. emersonii*, which share approximately 60% sequence identity. These CBH1s have a catalytic domain, a linker and a cellulose-binding domain. The available crys-

tal structures are for the catalytic domain, thus we only considered this domain for recombination (*see* **Note 4**). To eliminate the possibility of generating unpaired disulfide bonds, we mutated two residues in the *T. emersonii* and *T. aurantiacus* CBH1 sequences to cysteine (*see* **Note 5**). We used ClustalW2 [9] to align the parental sequences and we named our alignment file 'CBH1-msa.txt'.

4. A PDB structure file of one of the parental sequences (*see* **Note 6**). We used the *T. emersonii* structure, '1Q9H.pdb'.

5. A sequence alignment of one of the parental sequences with the sequence from the PDB structure file (see Note 7). We used ClustalW2 to align the parental sequences and we named our alignment file 'Temer-1Q9H.txt'.

2.4 Methods

1. Place the parent sequence alignment file (CBH1-msa.txt), the PDB structure file (1Q9H.pdb) and the PDB alignment file (Temer-1Q9H.txt) in the 'schema-tools' folder.

2. Run the following command (*see* **Note 8**) in the 'schema-tools' directory:

   ```
   python schemacontacts.py -pdb 1Q9H.pdb -msa CBH1-msa.txt -pdbal
   Temer-1Q9H.txt -o contacts.txt
   ```

   This generates a file containing the SCHEMA contacts called 'contacts.txt' (*see* **Note 9**).

3. Run the following command (*see* **Note 10**) in the 'schema-tools' directory:

   ```
   python rasppcurve.py -msa CBH1-msa.txt -con contacts.txt -xo 7 -o
   ```

```
opt.txt -min 15
```

This RASPP script identifies a set of 8-block candidate libraries with low $< E >$ (*see* **Note 11**). Each block is required to have at least 15 mutations. These libraries are saved to the file 'opt.txt' (*see* **Note 12**) (Figure 2.2).

4. Pick a library from the results file 'opt.txt' (*see* **Note 13**). In this case, we pick the library with crossover points [33 73 107 175 264 366 415], $< E >= 21.2$ and $< m >= 74.7$ (Figure 2.3).

5. Create a text file called 'CBH1-xo.txt' that contains the crossover points of the chosen library each separated by a space (*see* **Note 14**). The contents of the text file should be the following:

```
33 73 107 175 264 366 415
```

6. Run the following command (*see* **Note 15**) in the 'schema-tools' directory:

```
python schemaenergy.py -msa CBH1-msa.txt -con contacts.txt -xo
CBH1-xo.txt -E -m -o energies.txt
```

This generates a list of all the chimeras in the chosen library along with their SCHEMA energies and number of mutations (*see* **Note 16**). This list is saved to the file 'energies.txt'.

7. At this point we constructed a small chimera test set by substituting each block from each parent into the parental sequence from *T. emersonii*; the corresponding genes were synthesized (*see* **Note 17**). We could also have synthesized the genes encoding a different subset of the library (*see* **Note 18**) or even constructed the entire library (*see* **Note 19**). Before expressing the CBH1 chimeras, we add a linker and cellulose-binding domain to the recombined catalytic domains.

2.5 Notes

1. The RASPP toolbox 'schema-tools' is written for python 2.6 on a Unix-based system. We recommend using this python release for the RASPP toolbox.

2. As a general rule, when picking sequences for SCHEMA recombination we try to ensure the sequence identity between the homologs is not lower than ∼55% if individual genes are to be synthesized. In our experience, recombining sequences with much lower identities results in libraries with a high proportion of non-functional chimeras, even using SCHEMA. (This may not be a problem if the whole library is constructed and screened for functional chimeras.) The parental sequences are assumed to share the same fold; homologs with >55% identity are likely to have very similar structures. If a structure is available for multiple parental sequences, we confirm they have the same fold by aligning the parental structures.

3. Lines starting with '#' are ignored in the multiple sequence alignment file. Sequence similarity symbols and trailing numbers are also ignored.

4. SCHEMA library designs require a protein structure. If no structural information is available for a parent sequence, but there are structures of homologs, we can use MODELLER to build a structure model [10]. An inaccurate homology model hinders SCHEMA library design; an actual structure is preferred.

5. We assumed but did not verify that broken disulfide bonds are destabilizing. In this case, *C. thermophilum*, *H. jecorina*, and *A. thermophilum* CBH1s have 10 disulfide bonds while *T. aurantiacus* and *T. emersonii* have 9 disulfide bonds. If the cysteines from the missing disulfide bond are in separate sequence blocks, chimeras with un-

paired cysteines can result. We avoided this by modifying the parental sequences of *T. aurantiacus* and *T. emersonii* to include the remaining cysteine pair.

6. A structure is necessary to identify the residue-residue contacts. When possible, we pick a high-resolution structure (< 2.0 Å).

7. The sequence of the PDB file can be extracted with the following (run from the 'schema-tools' directory):

   ```
   python -c "import pdb; pdb.get('1Q9H.pdb')"
   ```

   We aligned this PDB sequence with the corresponding parent sequence (*T. emersonii* CBH1) from the parental alignment. The parent sequence must have the same identifier in both alignment files ('Temer') and the identifier of the PDB sequence must be the name of the PDB structure ('1Q9H'). The PDB sequence can be identical to the parent sequence, but this is not always the case; often the PDB sequence will be truncated or contain several point mutations. In our case we have mutated several of the residues in *T. emersonii* CBH1 to cysteine (*see* **Note 5**).

8. The python script 'schemacontacts.py' calculates all of the SCHEMA contacts. Several arguments need to be provided when running this script:

   - '-pdb 1Q9H.pdb': name of the PDB structure
   - '-msa CBH1-msa.txt': name of the parental sequence alignment
   - '-pdbal Temer-1Q9H.txt': name of the PDB sequence alignment
   - '-o contacts.txt': name of an output file to store the contacts

9. Each contact is represented as a pair of residue numbers in 'contacts.txt'. Numbering is given in terms of both the parental sequence alignment and the PDB sequence

alignment.

10. The python script 'rasppcurve.py' finds crossover points that minimize the average SCHEMA energy for a library. Several arguments need to be provided when running this script:

    - '-msa CBH1-msa.txt': name of the parental sequence alignment
    - '-con contacts.txt': name of the contacts file
    - '-xo 7': number of crossovers
    - '-min 15': minimum number of non-identical residues in a block (prevents trivial solutions)
    - '-o opt.txt': name of an output file for the results

    This script may take several hours to complete, depending on protein size and computer specifications. Increasing the number of crossovers in a library increases library size and reduces the average number of mutations in a block. The user may want smaller blocks if searching for properties from single point mutations. However, it is harder to find desirable chimeras in larger libraries and increasing the number of blocks increases a library's $< E >$. We chose to split our 5 parent proteins into 8 blocks.

11. There is a trade-off between the average SCHEMA energy of a library ($< E >$) and the average number of mutations from the closest parent ($< m >$), which depends on the relative block sizes (see Figure 2.2b). If all the blocks are evenly sized, $< m >$ is very high but the solution space of possible libraries is very small and so $< E >$ is large. As block sizes become uneven, the solution space of possible libraries increases.

This enables RASPP to find libraries with lower $< E >$ but these libraries have lower $< m >$. RASPP is designed to find low $< E >$ libraries for a range of $< m >$.

12. Each library is defined by 7 crossover points. The crossover points are given by the first residue of each new fragment (excluding the first fragment, which is always 1) based on the numbering of the parental sequence alignment. The results file 'opt.txt' also gives $< E >$ and the average number of mutations from the closest parent ($< m >$) for each library.

13. RASPP returns a set of candidate libraries with a range of $< m >$ values. A lower $< E >$ implies more functional chimeras in the library. For moderately sized proteins (250-500 amino acids) we try to pick SCHEMA libraries with $< E >$ less than 30. Protein-specific biochemical and structural knowledge may help users pick from the candidate libraries.

14. Lines starting with '#'are ignored in the crossover file.

15. The python script 'schemaenergy.py' lists the chimeras in a library. Several arguments need to be provided when running this script:

    - '-msa CBH1-msa.txt': name of the parental sequence alignment
    - '-con contacts.txt': name of the contacts file
    - '-xo CBH1-xo.txt': name of the crossover file that defines the library
    - '-E -m': specifies that the chimeras should be listed with their E and m values
    - '-o energies.txt': name of an output file for the results

16. Chimeras are numbered according to the parental sequence of each block with the numbers ordered from the first block to the last block. Parents are numbered based

on the order they appear in the parental sequence alignment. For example, chimera '14221313' has parent 1 as the sequence of its first block, parent 4 as its second block, etc.

17. The fungal CBH1 enzymes have poor heterologous expression in *S. cerevisiae*. Because *T. emersonii* CBH1 expresses much better than the other parents, we analyzed the blocks one at a time in the background of *T. emersonii* CBH1. These chimeras tend to have low SCHEMA energies and they can be easily constructed via overlap extension PCR. Using this 'monomera' approach, we identified stable CBH1 chimeras in a SCHEMA library similar to the one presented here [6].

18. We pick a subset of the library to analyze. We ensure every block from every parent is represented independently of one another in this subset. This enables us to model the effect blocks have on biochemical properties such as stability [5].

19. It is possible to construct an entire SCHEMA library in the laboratory by assembling blocks of sequence with specific overhangs [11, 12]. This approach is appropriate for searching for chimeras with specific properties that cannot be predicted from a small library sample.

2.6 Figures

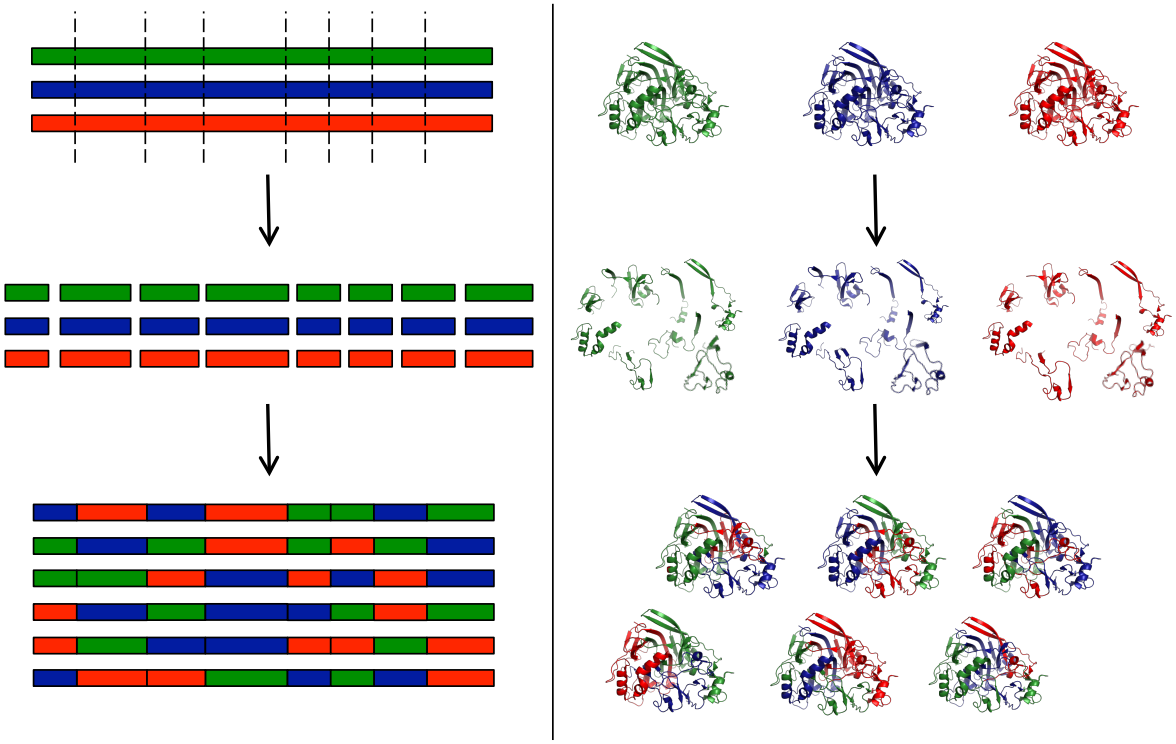


Figure 2.1: SCHEMA recombination. Homologous protein sequences are split into blocks at fixed crossover locations. These blocks are shuffled to generate novel chimeric proteins.

A

# <E>	<m>	crossover points
16.6800	63.5313	33 73 105 152 206 374 415
16.0000	64.9914	33 73 107 149 204 366 415
16.2800	65.6751	33 73 107 149 206 366 415
16.6400	66.6670	33 73 107 149 212 366 415
17.4000	67.9348	33 73 107 149 204 359 415
17.6800	68.5816	33 73 107 149 206 359 415
18.8000	70.0666	33 73 107 149 212 356 415
19.6000	70.8609	33 73 107 175 221 356 415
20.4000	72.3239	77 107 149 212 351 374 415
20.8800	73.1818	33 73 107 175 250 366 415
20.9600	73.8879	33 73 107 149 254 366 415
→ 21.1600	74.6592	33 73 107 175 264 366 415
22.2800	75.6126	33 73 107 175 271 366 415
22.5600	76.6614	33 73 107 175 264 359 415
22.5600	78.2912	33 73 107 149 264 346 366
22.6400	78.6867	33 73 107 149 264 344 366
23.5600	79.6392	33 73 107 175 264 338 366
25.2000	80.7276	73 107 175 271 329 366 415
26.6800	81.5644	43 107 175 264 317 356 406

B

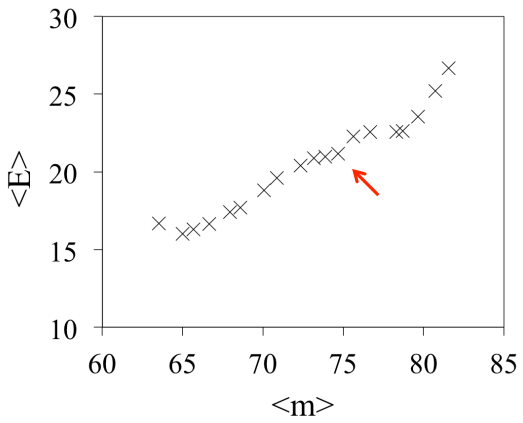


Figure 2.2: Libraries returned by RASPP. (a) The contents of 'opt.txt', which lists the crossover locations of candidate libraries identified by RASPP. (b) A graph of the possible libraries plotting average SCHEMA energy ($< E >$) of each library against the average number of mutations ($< m >$). The trade-off between $< E >$ and $< m >$ is apparent. The chosen library is highlighted with an arrow.

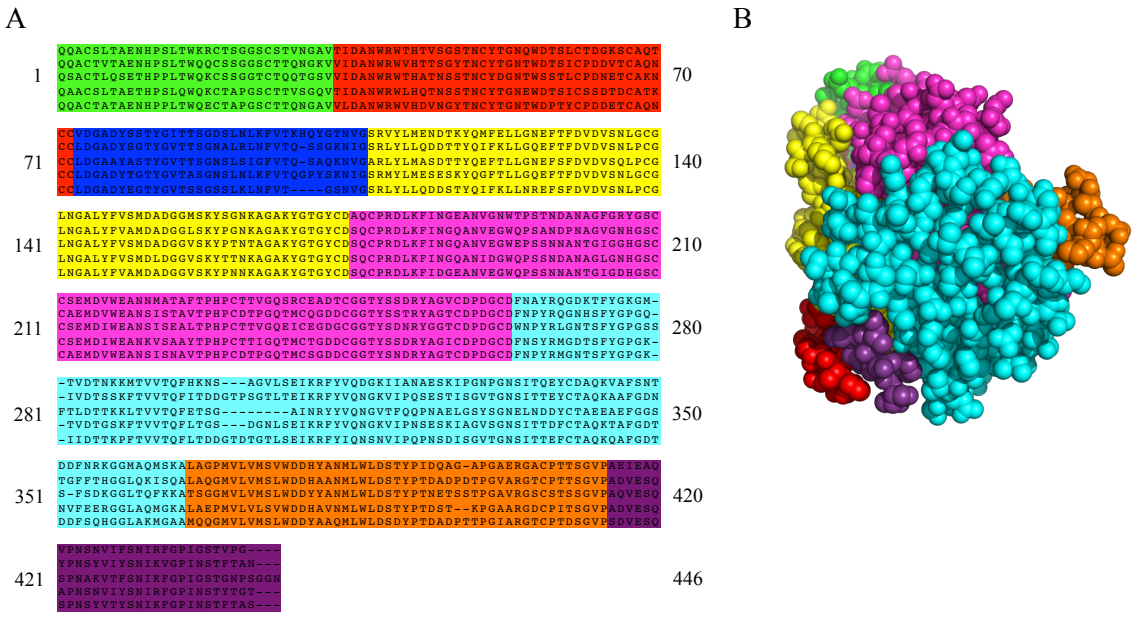


Figure 2.3: Visualizing the chosen RASPP design. (a) The multiple sequence alignment of the parent CBH1s with each of the 8 blocks highlighted in a different color. (b) The blocks highlighted on the CBH1 structure '1Q9H.pdb'.

2.7 References

1. Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L., and Arnold, F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553-558.

2. Meyer, M., Hochrein, L., and Arnold, F. H. (2006) Structure-guided SCHEMA recombination of distantly related β-lactamases. *Protein Eng. Des. Sel.* **19**, 563-570.

3. Endelman, J., Silberg, J., Wang, Z., and Arnold, F. H. (2004) Site-directed protein recombination as a shortest-path problem. *Protein Eng. Des. Sel.* **17**, 589-594.

4. Romero, P., Stone, E., Lamb, C., Chantranupong, L., Krause, A., Miklos, A., Hughes, R., Fechtel, B., Ellington, A. D., Arnold, F. H., and Georgiou, G. (2012) SCHEMA-designed variants of human arginase I and II reveal sequence elements important to stability and catalysis. *ACS Synth. Biol.* **1**, 221-228.

5. Li, Y., Drummond, D. A., Sawayama, A. M., Snow, C. D., Bloom, J. D., and Arnold, F. H. (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25**, 1051-1056.

6. Heinzelman, P., Komor, R., Kanaan, A., Romero, P. A., Yu, X., Mohler, S., Snow, C., and Arnold, F. H. (2010) Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Eng. Des. Sel.* **23**, 871-880.

7. Otey, C. R., Landwehr, M., Endelman, J. B., Hiraga, K., Bloom, J. D., and Arnold, F. H. (2006) Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol.* **4**, e112.

8. Heinzelman, P., Romero, P. A., and Arnold, F. H. (2013) Efficient sampling of SCHEMA Chimera Families for Identification of Useful Sequence Elements. In: Keasling, A (ed) *Methods in Enzymology: Methods in Protein Design*, Elsevier Ltd, Oxford, U.K.

9. 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., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

10. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U., and Sali, A. (2007) Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Protein Sci.* **50**, 2.9.1-2.9.31.

11. Hiraga, K., and Arnold, F. (2003) General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.* **330**, 287-296.

12. Farrow, M. F., and Arnold, F. H. (2010) Combinatorial Recombination of Gene Fragments to Construct a Library of Chimeras. *Curr. Protoc. Protein Sci.* **62**, 26.2.1-26.2.20.

Chapter 3

A diverse set of family 48 bacterial cellulases created by structure-guided recombination

A modified version of this chapter appears in: Smith M. A., Rentmeister A., Snow C. D., Wu T., Farrow M. F., Mingardon F., and Arnold F. H. (2012) A diverse set of family 48 bacterial cellulases by structure-guided recombination, *FEBS J.* **279**, 4453-4465, and is reprinted with permission from Wiley-VCH.

3.1 Abstract

Sequence diversity within a family of functional enzymes provides a platform for elucidating structure-function relationships and for protein engineering to improve properties important for applications. Access to natures vast sequence diversity is often limited by the fact that only a few enzymes have been characterized in a given family. Here, we recombine the catalytic domains of three family 48 bacterial cellulases (Cel48, **EC 3.2.1.176**), *Clostridium cellulolyticum* Cel48F, *Clostridium stercorarium* Cel48Y, and *Clostridium thermocellum* Cel48S, to create a diverse library of family 48 cellulases having an average of 106 mutations from the closest native enzyme. Within this set we find large variations in properties such as the functional temperature range, stability, and specific activity on crystalline cellulose. We show that functional status and stability are predictable from simple linear models of the sequence-property data: recombined protein fragments contribute additively to these properties in a given chimera. Using this, we correctly predict sequences that are as stable as any of the native Cel48 enzymes described to date. Characterizing 60 active Cel48 chimeras expands the number of characterized family 48 cellulases from 13 to 73. Our work illustrates the role structure-guided recombination can play in helping to identify sequence-function relationships within a family of enzymes by supplementing natural diversity with synthetic diversity.

3.2 Introduction

Cellulolytic anaerobic bacteria use macromolecular structures known as cellulosomes to hydrolyze recalcitrant cellulosic substrates [1]. Within the cellulosome, cellulases and other glycoside hydrolases [2, 3] are assembled onto multidomain scaffoldin proteins for efficient

degradation of cellulosic substrates [4]. Cellulosome assembly is achieved by binding dockerin domains from enzymes with cohesin domains in scaffoldin, while interaction with the substrate is mediated by one or more carbohydrate binding modules (CBMs) on the scaffoldin [1, 5].

The modularity of cellulosomes has spurred interest in 'designer cellulosomes' [4,6], where different cellulases are synthetically combined for a specific application. Within a given glycoside hydrolase family, a diverse pool of potential cellulases would be beneficial for designer cellulosomes by providing a suite of enzymes with differing properties and an extensive platform for further enzyme engineering. Family 48 cellulases (Cel48) are ideal candidates for designer cellulosomes. As one of the most important families of bacterial cellulases [7, 8], they are usually a major constituent of bacterial cellulosomes [9, 10]. Of the 116 bacterial Cel48 genes currently predicted in the CAZy database (http://www.cazy.org/) [11], only 13 have been characterized.

Here, we use SCHEMA recombination to synthesize a diverse set of new family 48 sequences. SCHEMA [12] is a structure-guided, site-directed protein recombination method that has been used to generate thousands of novel P450s [13], β-lactamases [14], and fungal cellulases [15, 16]. SCHEMA identifies optimal crossover locations for shuffling homologous genes, based on minimizing structural disruption in the resulting chimeric proteins. The chimeric proteins that are made by recombining natural sequences differ from the parent sequences at many amino acid positions and provide a convenient platform for structure-function studies. The new Cel48 enzymes described here are chimeras of the catalytic domains of three native Cel48 enzymes from mesophilic and thermophilic *Clostridia*. Sequence-function analysis of this synthetic enzyme library demonstrates a high degree of additivity in the sequence-stability relationship, as observed in previous studies [15, 17]. This simple

relationship between the sequence block identity and its contribution to chimera stability has allowed us to predict highly stable, highly active Cel48 enzymes. We have also investigated the relationship between thermostability and optimal catalytic temperature in this enzyme family.

3.3 Results

3.3.1 Cel48 parental enzymes

Three extensively characterized Cel48 cellulases were chosen as parents for construction of the SCHEMA recombination library: Cel48F [10] from the mesophile *Clostridium cellulolyticum* ATCC 35319, Cel48Y [18] from the thermophile *Clostridium stercorarium*, and Cel48S (also known as Cel48A) [19] from the thermophile *Clostridium thermocellum* ATCC 27405. All three enzymes are known to act on crystalline cellulose in a processive manner [10, 11]. Crystal structures of CelF and CelS show that the family 48 catalytic domain is an $(\alpha/\alpha)_6$ barrel fold. The sequence and structural similarities of the catalytic domains (Supplementary Figure 3.10) suggest that these enzymes can be recombined to make functional catalytic domain chimeras.

Outside of the catalytic domain, however, the parent enzymes exhibit significant structural variations. *C. cellulolyticum* CelF and *C. thermocellum* CelS consist of a 70kDa catalytic domain connected to their organisms' respective dockerin domains, whereas *C. stercorarium* CelY is a non-cellulosomal 103 kDa protein with its N-terminal catalytic domain attached via a 10 kDa domain of unknown function (DUF) to a 17 kDa cellulose binding domain (CBM3) [18]. Thus CelY can directly bind cellulose, whereas CelF and CelS bind their respective scaffoldins.

Since the non-catalytic domains (dockerin, scaffoldin, CBM) differ among the parent enzymes, we chose to construct the library using the *C. thermocellum* architecture. Having a single architecture for the cellulases enables fair comparison of the chimeric cellulase catalytic domains. A miniscaffoldin consisting of a *C. thermocellum* cohesin and a cellulose binding module (CBM) was constructed as described [20], and the *C. thermocellum* CelS dockerin was fused to the C-terminus of the catalytic domains of CelF, CelS, and CelY (see Materials and Methods). The parental constructs, with the added *C. thermocellum* dockerin domain, are referred to as CelF-1, CelS, and CelY-2 and are highlighted (boxed) in Figure 3.1. These constructs can attach to the miniscaffoldin to produce minicellulosomes. Another CelY construct was created with the addition of its domain of unknown function (DUF). Because the presence or absence of this DUF did not affect activity of the CelY constructs (Supplementary Figure 3.11B), the DUF was excluded in constructing the recombination library.

We first characterized and compared the activities on crystalline cellulose of the parental enzymes with and without the miniscaffoldin. For all cellulases having a dockerin, activity was substantially higher in the presence of miniscaffoldin than without it (Figure 3.2A-C). Thus cohesin-dockerin binding occurs, and CBM-mediated attachment to cellulose enhances the rate of sugar release from crystalline cellulose, as observed previously [21]. Figure 3.2D directly compares the activity profiles for the dockerin-containing cellulases in the presence of *C. thermocellum* miniscaffoldin. Under these conditions, CelY and CelS displayed the highest activity at 70 - 80°C and very low activity below 50°C. In contrast, CelF is most active at ∼50°C, but quickly loses activity at higher temperatures. In a previous study we compared the activities of three homologous bacterial glycoside hydrolase family 9 CBM3c cellulases from mesophilic and thermophilic organisms over a range of temperatures. They

all displayed similar activities at lower temperatures, and that activity increased with temperature until the enzyme was no longer stable [20]. Here, in contrast, the Cel48 cellulase from the mesophilic organism is significantly more active than its two thermophilic homologs at the lower temperature.

3.3.2 SCHEMA recombination library design

A structure-guided computational approach to designing a library of chimeric genes, SCHEMA identifies crossover sites for recombination of homologous proteins that maximize the likelihood that proteins in the resulting library will retain their folded structure [12]. Contacts (residues that are less than 4.5 $\AA$ from one another) are identified from one or more of the crystal structures, and SCHEMA energy E for a given chimera is calculated by counting the number of residue-residue contacts that are disrupted by recombination. Recombination sites are chosen to minimize the average SCHEMA energy, $< E >$, of all possible sequences made by recombining those sequence fragments.

We designed the recombination library of glycoside hydrolase family 48 catalytic domains using the RASPP algorithm [22] to identify crossover sites that minimized $< E >$ [12]. RASPP returned a set of candidate library designs (Supplementary Figure 3.12). The chosen library has crossovers located before residues Pro122, Ala260, Asp292, His348, Gly396, Asn437, Leu556, based on the numbering of CelS (pdb **1L2A**). This library has an average SCHEMA energy $< E >$ of 31 and an average number of mutations from the closest parent $< m >$ of 106. The individual structural elements ('blocks') for this design, shown in Figure 3.3A, are not obvious based on secondary or domain structure. Crossovers between blocks B-C, C-D, and G-H, for example, lie within α-helices. This design, however, sequesters as many residue-residue contacts as it can within blocks, given limitations on block size

(Figure 3.3B).

Chimeric genes were assembled from 24 gene fragments, representing the 8 blocks from each of the 3 parents, using the Sequence-Independent Site-Directed Chimeragenesis (SISDC) method [23] to generate a gene library of $3^8 = 6,561$ different sequences (Supplementary Table 3.1 and Supplementary Figure 3.13). A *C. thermocellum* dockerin was attached to the C-terminus of each chimeric sequence during reassembly. Methods used to express, purify and identify functional chimeras are described in detail in the Materials and Methods.

3.3.3 Characterization of chimeric family 48 cellulales

Upon screening 4,872 library members using a 96-well plate cellulase activity assay (see Materials and Methods), we identified the functional enzymes, from which we purified and characterized 50 unique, novel family 48 cellulases. As shown in Figure 3.4, these enzymes have, on average, more than 80 mutations from the closest parent cellulase. Their SCHEMA E values range from 8 to 36, and they have 12 to 142 mutations from the closest parent cellulase. Sequences from all three parental enzymes are well represented at each block in the functional chimeras, except for CelF, which is underrepresented in blocks E, G and H.

We measured the thermostabilities (T_{50}) and optimal catalytic temperatures (T_{opt}) of the 50 Cel48 chimeras and their three parents; these values are reported in Figure 4. T_{50} is the temperature at which an enzyme loses 50% of its activity after a 10-minute incubation (see Materials and Methods) and is a measure of its ability to resist temperature-induced irreversible inactivation. T_{opt} is the temperature at which a cellulase is most active over a 2-hour assay (see Materials and Methods) and is a measure of its ability to remain active at elevated temperature. Thermostability, the ability to withstand denaturation, is necessary

but not sufficient for increasing an enzymes optimal catalytic temperature. In the chimeras, both these measured properties extend beyond the range of the parents. Many of the chimeras are very stable: indeed, this experiment has added 35 new Cel48 enzymes with a $T_{opt} > 60°C$ to the 6 natural thermostable cellulases that have been characterized to date: *Clostridium thermocellum* ATCC 27405 CelS [24], *Clostridium thermocellum* F7 CelS [25], *Clostridium thermocellum* ATCC 27405 CelY [26], *Thermobifida fusca* YX CelF [27], *Clostridium stercorarium* CelY [28], and *Anaerocellum thermophilum* DSM 6725 CelA [29].

We also measured the specific activities of all the Cel48 chimeras at their respective optimal catalytic temperatures (Figure 3.4 and Figure 3.5A). The chimeras tend to have specific activities that are similar to or slightly less than the parent enzymes. We did not observe a correlation between T_{opt} and specific activity at that temperature for all of the sampled chimeras (Figure 3.5B). However, recombination may have compromised the activities of many of the chimeras. If only the most active enzymes are considered, there does appear to be a correlation between T_{opt} and specific activity (Figure 3.5B, dotted line), where increasing temperature leads to higher specific activity.

3.3.4 Modeling and predicting function of chimeric cellulases

As previously demonstrated for fungal CBHI and CBHII cellulases [15, 16], we can use information from a small number of sequences to predict properties of all the chimeras in the recombination library. To demonstrate this for Cel48, we built predictive models of T_{50} and T_{opt} based on the sequences and SCHEMA E values of the 50 functional chimeric cellulases and the 3 parental enzymes. We modified the simple sequence-stability linear regression model first used by Li et al. [17] to include an additional parameter for second-order SCHEMA contacts in the chimeras (Supplementary Equation 3.1). As shown in Figure

3.6A, the thermostability model fits the T_{50} measurements of all 53 enzymes well ($r^2 = 0.88$) and is an improvement over the simpler model that does not include the SCHEMA E parameter ($r^2 = 0.82$), as illustrated in Supplementary Figure 3.14.

With this model we were able to identify the contribution that each sequence block makes to stability (Figure 3.6B). When trained on T_{opt} measurements, the same block-additive model also accurately predicts the measured values (Figure 3.6C), and the block contributions to optimal catalytic temperature are very similar to their contributions to thermostability (Figure 3.6D). These models trained on data from the sample set can be used to predict the T_{50} and T_{opt} of all the remaining chimeras in the library.

We wished to construct and test the chimeric cellulases that are predicted to be the most thermostable. Not every chimeric cellulase, however, is functional. To investigate how recombination leads to nonfunctional sequences, we analyzed 28 unique inactive chimeras identified during the activity screen. A chimera was defined as nonfunctional if upon a five-fold increase in enzyme concentration, from 0.2 μM to 1 μM, no detectable activity was measured between 45°C and 80°C. These nonfunctional cellulases are all soluble proteins of the correct length on an SDS page gel (data not shown). Using circular dichroism, we analyzed 17 of the 28 non-functional chimeras at 25°C and found that all gave a similar signal to the parent enzymes (Supplementary Figure 3.15), suggesting that nonfunctional chimeras are folded and have a similar secondary structure to functional ones.

Inspired by the success of the additive block models for thermostability and thermoactivity, we took a similar approach to modeling and predicting chimera functional status. We constructed a linear model where each block contributes independently to whether a chimera is functional or not. As with thermostability, we also included the SCHEMA E value as a parameter. The output from the model should be a value between 0 and 1 to

represent the probability that a chimera is active. To do this we augmented the output of the linear model using a linking function, f_{link}, which scales outputs of the model to the required range (Supplementary Equation 3.2). The coefficients for this model can be found by linear regression (Supplementary Table 3.2), although, unlike the thermostability model, the block contributions are only additive under the linking function.

We trained the activity model on 81 cellulases (53 active, 28 inactive) and assessed its predictive ability by cross-validating the predictions of functional chimeras to the measurements of functional chimeras. The model successfully predicted the functional status of 88% of the chimeras (Supplementary Table 3.3). A low SCHEMA E value is known to increase the likelihood of a chimera being active [14], but E alone correctly predicted the functional status of only 77% of these chimeras under the same cross-validated conditions. Running the functionality model on all block combinations, we predict that the library contains more than 3,000 unique active Cel48 cellulases.

Using the T_{50} model trained on the 53 experimentally active sequences in combination with the functionality model, we predicted the 13 most stable enzymes that are also expected to be catalytically active. These were constructed and characterized. Ten of the 13 were active; these sequences and their stabilities are reported in Figure 3.4. As shown in Figure 3.7A, their stabilities closely matched the predictions. Five of these variants were slightly more stable than the most stable parental enzymes. Interestingly, two of the highly stable chimeras also hydrolyze more cellulose than the most active parental enzyme, CelY-2 both in a 1-hour assay (Figure 3.5, Figure 3.7C and D) and in a 48-hour assay (Figure 3.7B), demonstrating the potential utility of these chimeric enzymes for designer cellulosomes.

3.3.5 Probing biochemistry with synthetic diversity

With 60 active cellulase chimeras in hand, we next examined the relationship between the optimal temperature for catalytic activity (T_{opt}) and resistance to temperature-induced denaturation (T_{50}) over a broad range of temperatures. These two properties are closely correlated (Figure 3.8), indicating that engineering Cel48 enzymes for greater thermostability increases their optimal catalytic temperatures. Some of the chimeric cellulases have a T_{opt} value higher than their T_{50}. We believe this reflects the stabilizing effect of cellulose substrate, because the substrate is present in the T_{opt} assays but not in the denaturation step of the T_{50} assays. This effect can be seen in Supplementary Figure 3.16, where T_{50} values in the presence of cellulose are $\sim 2°$C higher than in its absence.

3.4 Discussion

The dearth of characterized family 48 cellulases with different properties is an impediment to their use in designer cellulosomes for specific engineering applications, and inhibits the discovery of sequence-function relationships for this important enzyme. We have used structure-guided protein recombination to expand the diversity of characterized family 48 bacterial cellulases. Using SCHEMA to identify suitable crossover locations for shuffling sequence blocks among the three parent Cel48 catalytic domains, we have generated a large set of novel, active cellulases which have the same architecture and express under the same conditions in the same *E. coli* host, where they are straightforward to characterize and compare. As expected, we find that properties such as T_{opt} (the ability to remain active at elevated temperature), T_{50} (the ability to withstand denaturation at high temperature), and the specific activity at T_{opt} vary greatly among these novel enzymes. We also find

that functional status, T_{50}, and T_{opt} can be predicted from simple linear models built from sequence-function data from a small sample of the library. This has enabled us to efficiently identify stable chimeras, some of which have high cellulolytic activities.

This set of related enzymes can contribute to our understanding of how sequence affects family 48 cellulase properties. The thermostability model illuminates stabilizing blocks of amino acids, whether they exist in the most stable proteins or not. Two of the most stabilizing blocks are predicted to be from parent CelS at positions F and G. These blocks are located in the C-terminus of the catalytic domain, close to where the dockerin attaches, which suggests an important stabilizing interaction between these blocks and the *C. thermocellum* dockerin. When the dockerin binds the cohesin, the linker between catalytic domain and dockerin is pleated, and this brings the dockerin in close contact with the catalytic domain [30]. A CelS-dockerin-cohesin crystal structure would be valuable for identifying specific stabilizing interactions between these two domains.

With this work we also address another biochemical question with important engineering implications. Using this accessible set of related enzymes, we investigated the correlation between the temperature at which an enzyme is most active and the temperature at which it denatures irreversibly. We find that Cel48 chimeras with greater thermostability also have their activity optima at higher temperatures, and that these temperatures are closely related. In other words, the ability to withstand temperature-induced denaturation at ever-higher temperatures leads to increases in the optimum temperature for activity. It is not necessarily the case that increased structural stability and resistance to denaturation and irreversible inactivation will result in the ability to catalyze the reaction efficiently at higher temperature, particularly if local instability or dynamics influence catalysis [31]. Among the Cel48 chimeras, however, there is sufficient structural stability in key catalytic regions

to render T_{50} a good surrogate for T_{opt}.

We found two of the predicted thermostable chimeras had higher specific activities at T_{opt} than the most active parental enzyme, CelY-2. When assayed over a 48-hour period, they hydrolyzed twice as much cellulose as CelY-2. These chimeric enzymes, which we have analyzed in a cellulosomal construct, may find potential uses in designer cellulosomes. An important next step will be to determine if they provide an enhanced cellulolytic capability to a system such as the *C. thermocellum* cellulosome.

3.5 Materials and methods

3.5.1 Parental enzyme constructs

Cel48 genes from CelF and CelS were PCR-amplified using Phusion-polymerase from genomic DNA using primers CTHE312.40, CTHE2453.40 for CelS and CCEL786.41 and CCEL2864.41 for CelF, introducing HindIII and SacI sites at the 5'-end as well as a NotI site at the 3'-end (Supplementary Table 3.5). Taq polymerase was used to add A-overhangs for TA-cloning into pGEM-T Easy (Promega). The resulting plasmids were called pGEMT-CTHEwt and pGEMT-CCELwt. The CelS dockerin was added to the CelF catalytic domain to create the plasmid pGEMT-CCELmut1. These constructs were cloned into pET-22(+) using NdeI and NotI sites.

We designed a synthetic gene for CelY from *C. stercorarium* based on available sequence information but removed restriction sites NdeI, HindIII, BsaXI, PstI, SapI. The gene was codon-optimized for expression in *E. coli* by DNA 2.0. The CelY gene was cloned into pET-22(+) using NdeI and NotI restriction sites. The resulting construct was termed pET22b+CSTEwt and contains the catalytic domain, the domain of unknown function, and

the CBM. Two more constructs were made from the CelY gene: CelY-1 containing only the catalytic domain and *C. thermocellum* dockerin and CelY-2 containing the catalytic domain and the domain of unknown function (DUF) and the *C. thermocellum* dockerin. Products were cloned into pET-22(+) using NdeI and NotI restriction sites.

An XbaI site was introduced by overlap extension PCR into all parental constructs between the catalytic domain and the dockerin. Introducing an XbaI restriction site between the catalytic domain and the dockerin allowed swapping catalytic domains and dockerins. The XbaI site did not affect activity (Supplementary Figure 3.11A).

3.5.2 Recombination library design

The SCHEMA library was designed using the tools available on the Arnold group homepage (http://www.che.caltech.edu/groups/fha/). The catalytic domains of CelF, CelY, and CelS were aligned using ClustalW from Tyr40Phe661, based on numbering of CelS. We analyzed all available structures without point mutations of the catalytic domains of CelS and CelF (CelF pdbs: 1F9O, 1FAE, 1FBO, 1FCE, 1G9G; CelS pdbs: 1L1Y (6 chains), 1L2A (6 chains); a total of 17 chains). Of the 3035 unique residue-residue contacts in all 17 structures, on average 73% are conserved between any CelF structure and CelS structure. This compares to an average of 80% of contacts conserved between any two CelF structures and 80% of contacts conserved between any two CelS structures. Since contacts between structures of the same enzyme vary almost as much as contacts between structures of CelF and CelS, we made use of all 17 available structures in designing the library. The average SCHEMA energy for a library ($< E >$) was calculated for each structure and libraries were evaluated based on the average $< E >$ from all 17 structures. Seven crossover sites were chosen using the RASPP algorithm [22] with a minimum fragment size of 30 residues.

RASPP returned a set of candidate libraries characterized by $< E >$ (the average number of contacts broken within a library for a given structure), $<< E >>$ (the average of $< E >$ for a given library across all 17 different structures), and $< m >$ (the average number of amino acid substitutions from the closest parent within a library). Supplementary Figure 3.12A shows $<< E >>$ as a function of $< m >$. We removed solutions without a conserved amino acid at the designated crossover sites (Supplementary Figure 3.12B). To obtain libraries with mutations more evenly distributed into blocks, we also calculated the standard deviation of the average number of mutations per block for each library. Lower numbers indicate more evenly distributed blocks. Supplementary Figure 3.12C shows $<< E >>$ as a function of the standard deviation of block mutations. From this set we picked a library that would contain a large number of active enzymes with high sequence diversity: the chosen library has an $<< E >>$ of 31.3 and $< m >$ of 106. Calculated for each of the 17 structures, $< E >$ for the library varies from 28 to 34.

3.5.3 Construction of chimeras

Chimeric genes were assembled from 24 gene fragments, representing the 8 blocks from each of the 3 parents, using the Sequence-Independent Site-Directed Chimeragenesis (SISDC) method [23]. The following consensus sites were used for the crossover sites: 1) CCG, 2) GCC, 3) GAC, 4) CAT, 5) GGT, 6) AAC, 7) TTA (Supplementary Table 3.6). Mini-libraries were cloned into pGEMT using SpeI and Sac II sites. Full libraries were made by isolating large amounts of DNA from plasmids digested with SpeI and SacII, not by PCR amplification. Instead of SapI, the isochizomer LguI was used. A *C. thermocellum* dockerin was attached to the C-terminus of each chimeric sequence during reassembly. The genes were expressed in pET-22(+) under the control of an IPTG-inducible T7 promoter

in E. coli BL21(DE3). A similar approach was taken for constructing the specific chimeras predicted to be thermostable, but with the difference that only the specific blocks for the desired chimera were used in the ligation steps.

3.5.4 Quality of library

We completely sequenced 61 randomly-chosen chimeras in order to assess the frequency of library construction artifacts, including point mutations, deletions, and insertions. 89% of the library (54 out of 61) contained no amino acid mutations, no insertions and no deletions. We found one single insertion, and two sequences were missing one-half of the library. Two sequences were back-to-front in the vector, and two sequences contained one remaining tag. Every block from every parent was found in the randomly sequenced chimeras, but CelF block E appears to be underrepresented in the library. The distribution of each block is displayed in Supplementary Table 3.7.

3.5.5 Protein expression in 96-well plates

In 96-well shallow-well plates, 300 μL of LB medium (10 g tryptone, 5 g yeast extract, 10 g NaCl) containing 100 mg/L ampicillin were inoculated with a single colony of *E. coli* BL21(DE3) having the cellulase gene on a pET-22(+) plasmid. Plates were grown overnight in an orbital shaker at 37°C, 250 rpm. In a 96-well deep-well plate, 900 μL of TB medium (12 g tryptone, 24 g yeast extract, 4 mL glycerol, in 1L H_2O with 17 mM KH_2PO_4 and 72 mM K_2HPO_4) containing 100 mg/L ampicillin were inoculated with 50 μL and grown in an orbital shaker at 37°C until the OD600 reached 1.6-1.8. Plates were cooled to < 17°C, induced with a final concentration of 50 μM IPTG and grown at 17°C for 16 hours. Cultures were harvested by centrifugation and stored at -20°C.

3.5.6 Cellulase activity assay in 96-well plates

Cells were resuspended in 300 μL lysis buffer (10 mM Tris, pH 8.0, 10 mM $MgCl_2$, 0.7 mg/mL lysozyme, 4 U/mL DNase) per well and incubated for 60 min at 37°C. Plates were centrifuged for 5 min at 5,000 g at 4°C. From the supernatant 100 μL were transferred to a 96-well PCR plate with 50 μL of a 10 g/L Avicel suspension in reaction buffer (50 mM succinate, pH 6.0, 1 mM $CaCl_2$) and 0.2 μM purified miniscaffoldin (Supplementary Figure 3.17). Hydrolysis proceeded overnight at both 50°C and 75°C. Plates were centrifuged for 3 min at 200 g at 4°C, and from each well 50 μL of supernatant were transferred to a new plate. The amount of reducing ends was determined using the Park-Johnson assay.

3.5.7 Park-Johnson activity assay [32]

Reagent A: 0.5 g/L $K_3Fe(CN)_6$, 0.2 M K_2HPO_4, pH 10.6. Reagent B: 5.3 g/L Na_2CO_3, 0.65 g/L KCN. Reagent C: 2.5 g/L $FeCl_3$, 10 g/L polyvinylpyrrolidone, 1 M H_2SO_4. In a 96-well PCR plate, 50 μL of test sample was mixed with 150 μL of a 2:1 A/B mixture (i.e. 100 μL A and 50 μL B). The plate was sealed and heated to 95°C for 15 min, then cooled to 4°C. Out of this plate 180 μL were transferred to a transparent flat-bottom screening plate containing 90 μL reagent C. The plate was incubated in the dark for 1-3 min before the OD at 520 nm was measured in a TECAN plate reader. If glucose equivalents were determined, a calibration curve made from solutions of defined glucose concentrations was included on each plate.

3.5.8 Enzymatic glucose activity assay

BG: 0.25 g/L almond beta-glucosidase in 50 mM sodium acetate, pH 5.0. TMB: 0.8 g/L tetramethylbenzidine in ddH_2O. HRP: 0.15 g/L horseradish peroxidase in 50 mM sodium

acetate, pH 5.0. GOX: 0.1 g/L glucose oxidase in 50 mM sodium acetate, pH 5.0. In a transparent flat-bottom screening plate, 100 μL of test sample was mixed with 50 μL of BG. If glucose equivalents were determined, a calibration curve made from solutions of defined glucose concentrations was included on each plate. The plate was sealed and incubated for 16 h at 37°C. For development, 50 μL of TMB, and 20 μL each of HRP and GOX were added to the plate. After 5 minutes, the OD at 650 nm was measured in a TECAN plate reader.

3.5.9 Protein purification

Each cellulase was purified from *E. coli* BL21(DE3) which contains the cellulase gene with a C-terminal his-tag on a pET-22(+) plasmid under the control of an IPTG- inducible promoter. The cells were grown in TB medium (12 g tryptone, 24 g yeast extract, 4 mL glycerol, in 1 L H_2O with 17 mM KH_2PO_4 and 72 mM K_2HPO_4) at 37°C with 100 mg/L ampicillin. Cells were induced with a final concentration of 50 μM IPTG, grown for 16 hours at 17°C and harvested by centrifugation for 10 min at 5000 g. Pellets were resuspended in buffer A (20 mM Tris, pH 7.4). The solution was lysed by sonication and centrifuged at 75,000 g for 30 min to sediment cell debris. The supernatant was loaded onto a 1 mL Ni-NTA His-trap column (GE Healthcare) and purified by washing with 1% buffer B (20 mM Tris, pH 7.4, 100 mM NaCl, 300 mM imidazole) for 15 column volumes (CV), followed by a gradient elution (increase to 80% buffer B in 10 CV). Cellulase-containing fractions were pooled and concentrated using protein concentrators with cellulose-free membranes (Vivaspin). Buffer was exchanged to 10 mM Tris, pH 8.0 by repeated refills. Purified proteins were flash frozen and stored at -20°C for up to 3 months. Protein concentration was determined using the Bradford assay and bovine serum albumin as protein standard.

Protein purity was determined from SDS-polyacrylamide gels. Isolated protein was 1560 mg/L for dockerin-containing constructs and 120 mg/L for CelY.

3.5.10 Thermostability assay (T_{50} measurements)

For each well of a 96-well PCR plate, 50 μL of a 20 g/L Avicel suspension in reaction buffer (50 mM succinate, pH 6.0, 1 mM $CaCl_2$) was mixed with 25 μL of 0.8 μM miniscaffoldin and spun down for 10 min at 5,000 g. In a different PCR plate, 30 μL of 0.8 μM cellulase in reaction buffer were pipetted per well. Plates were incubated for 10 min in a gradient PCR cycler at indicated temperatures, and then placed on ice. Heat-treated cellulases were transferred (25 μL per well) to the Avicel-containing PCR plate and the reaction was run for 60 min at the indicated temperature. Plates were spun down for 3 min at 200 x g. Then, 50 μL of supernatant were transferred to a new 96-well PCR plate and tested with either the Park- Johnson assay or the enzymatic glucose assay.

3.5.11 Temperature profiles (T_{opt} measurements)

A final concentration of 0.2 μM enzyme or 0.2 μM enzyme plus 0.2 μM miniscaffoldin was added to a preheated suspension of 10 g/L Avicel in reaction buffer (50 mM succinate, pH 6.0, 1 mM $CaCl_2$). The hydrolysis was performed at a range of temperatures for 2 hours in duplicate. Samples were spun down for 1 min at 200 x g at 4°C. From each well, 50 μL of the supernatant were transferred to a 96-well PCR plate and analyzed using either the Park-Johnson assay or the enzymatic glucose assay. The T_{opt} was determined from the temperature profiles of the chimeras.

3.5.12 Forty-eight hour activity assay

A final concentration of 0.2 μM enzyme plus 0.2 μM miniscaffoldin was added to a preheated suspension of 10 g/L Avicel in reaction buffer (50 mM succinate, pH 6.0, 1 mM $CaCl_2$) at 75°C. At regular intervals, the Avicel was resuspended and a sample of the reaction mixture was removed and cooled to 4°C. Samples were spun for 1 min at 200 x g and 50 μL of a 1:10 dilution of the supernatant were analyzed using the Park-Johnson assay. The measurements were performed in triplicate.

3.5.13 Circular dichroism

Circular dichroism measurements were carried out using an Aviv Model 62DS spectrometer with 6 μM protein sample concentration. Wavelength scans to determine the ellipticity were carried out at 25 °C.

3.5.14 Linear regression

Regression models for T_{50} and T_{opt} were trained using Matlab's 'regress' function. The regression model for functionality was trained using L1 regularized logistic regression from the toolbox glmnet for Matlab [33, 34].

3.6 Figures

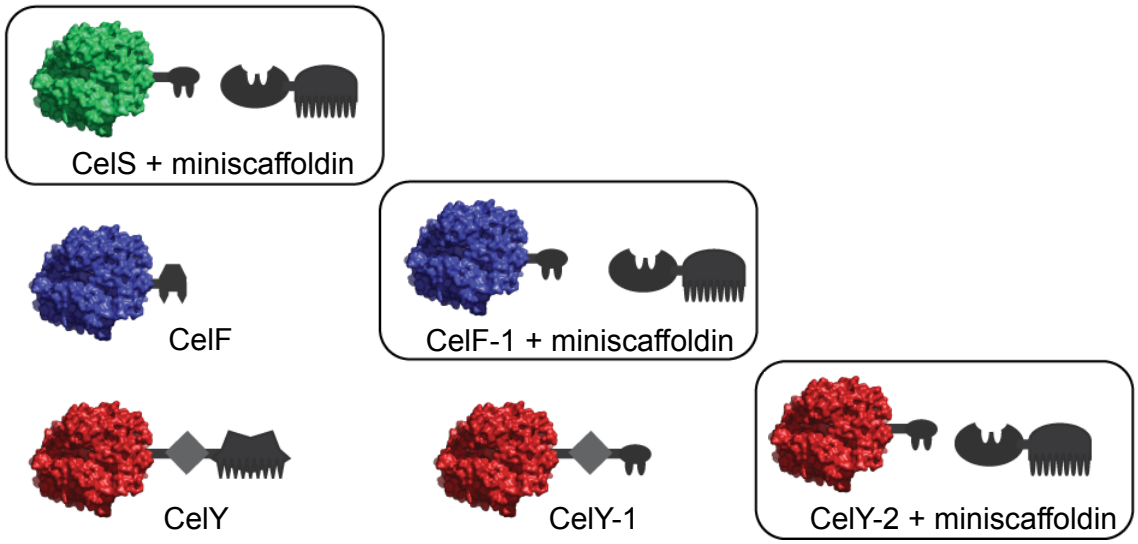


Figure 3.1: Architectures of parent family 48 glycosyl hydrolases and derived constructs. Wild-type CelS and CelF consist of an N-terminal catalytic domain and a C-terminal dockerin that binds specifically to its cohesin. Miniscaffoldin (black) consists of a *C. thermocellum* cohesin and CBM. Construct CelF-1 contains a C-terminal *C. thermocellum* dockerin and binds to the miniscaffoldin. CelY from *C. stercorarium* consists of an N-terminal catalytic domain, a domain of unknown function (DUF) and a CBM. CelY constructs CelY-1 and CelY-2 contain the CelY catalytic domain and a C-terminal dockerin from *C. thermocellum*. CelY-1 also contains the DUF. CelY-1 and CelY-2 bind to the miniscaffoldin. All constructs used to prepare the chimera library (boxes) have the *C. thermocellum* dockerin and bind the *C. thermocellum* miniscaffoldin.

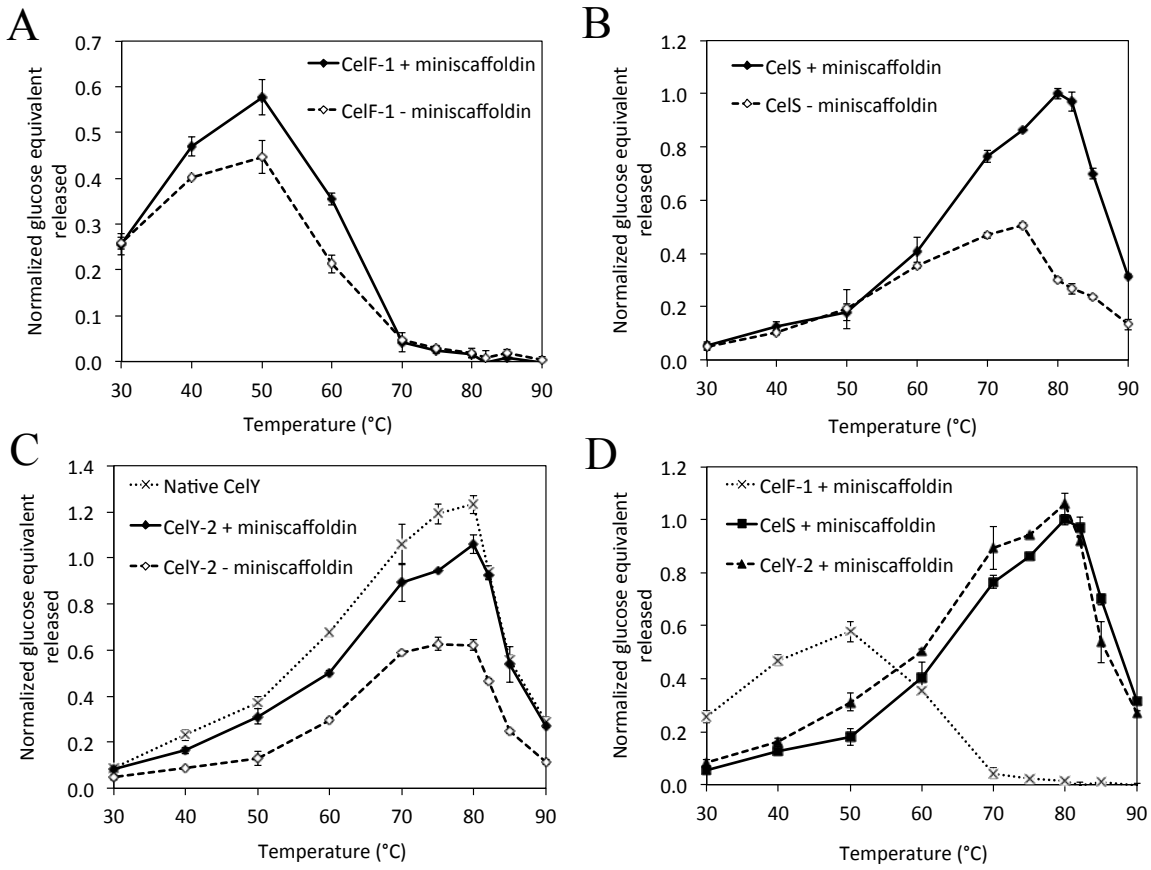


Figure 3.2: Activities of purified family 48 cellulases as a function of temperature, in the presence and absence of equimolar amounts of miniscaffoldin. Activities were determined from the total glucose equivalent released, using the enzymatic glucose assay (see Materials and Methods), in a 1-hour reaction with 0.2 μM enzyme and 10 g/L Avicel. All activities are normalized to the activity of CelS at its maximum, at 80°C. A) CelF-1, B) CelS, C) CelY-2, along with the native CelY enzyme. D) Temperature profiles of CelF-1, CelS, and CelY-2 constructs with miniscaffoldin. CelS and CelY-2 are most active at 75 - 80°C, whereas CelF-1 is most active at 50°C.

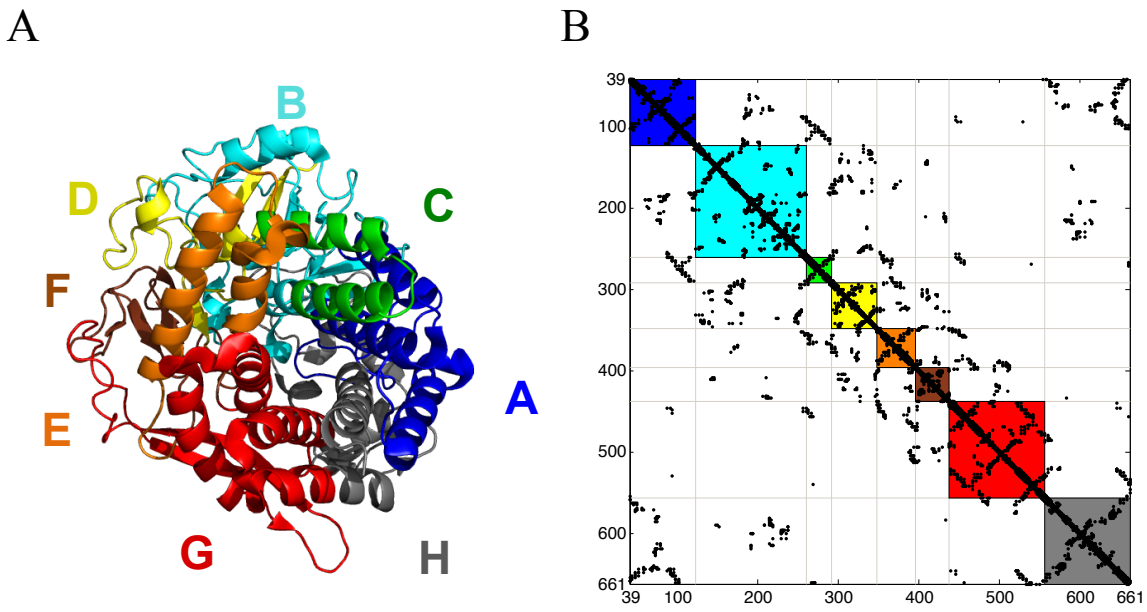
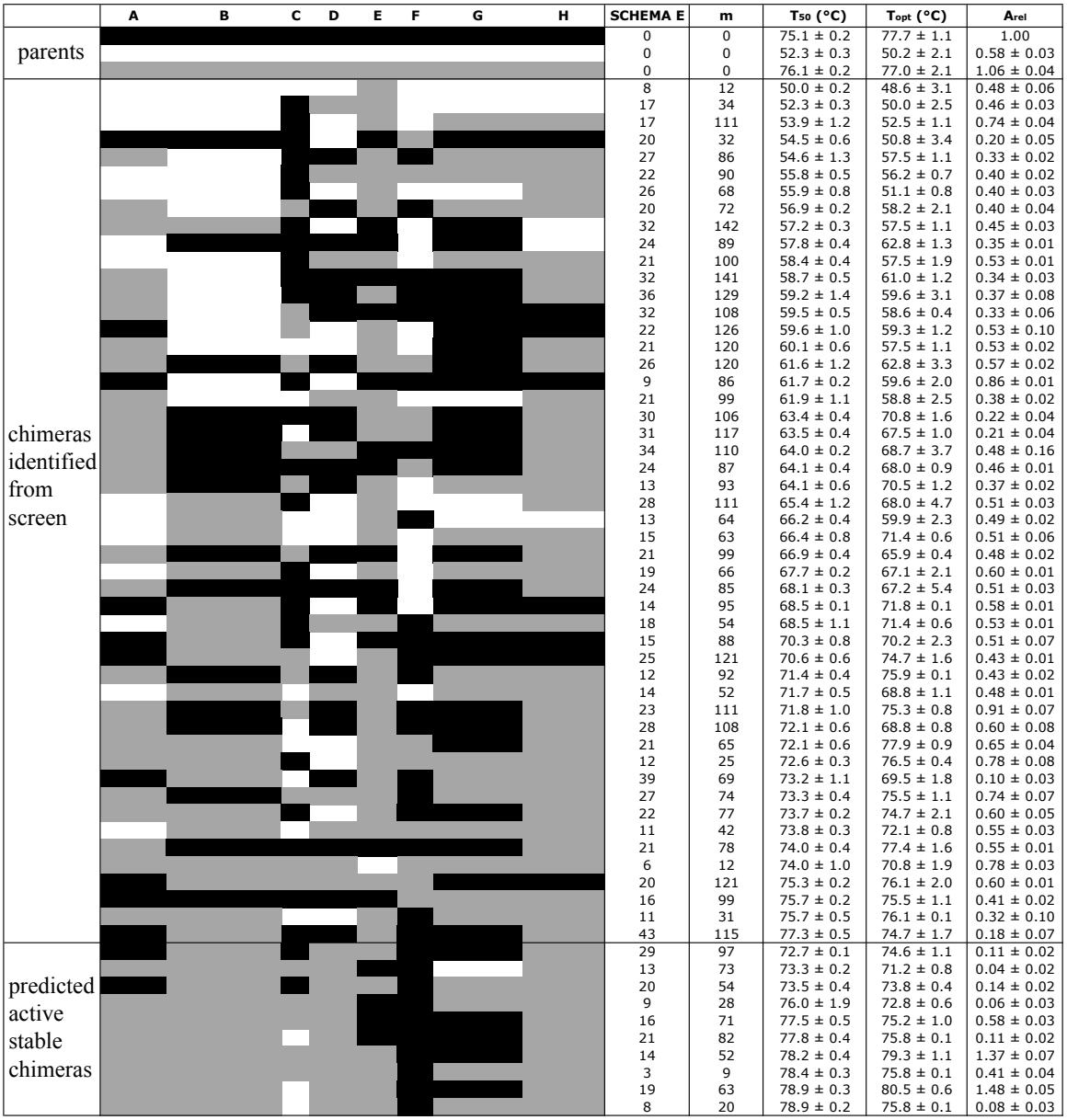


Figure 3.3: Sequence blocks in family 48 chimeras designed by SCHEMA. A) Structure of CelS with color-coded blocks A-H. B) Residue-residue contact map showing the combined contacts from 17 Cel48 structures. The positions of the blocks are indicated with colored squares. Most contacts are sequestered within the blocks and cannot be broken upon recombination.



	SCHEMA E	m	T_{50} (°C)	T_{opt} (°C)	A_{rel}
parents	0	0	75.1 ± 0.2	77.7 ± 1.1	1.00
	0	0	52.3 ± 0.3	50.2 ± 2.1	0.58 ± 0.03
	0	0	76.1 ± 0.2	77.0 ± 2.1	1.06 ± 0.04
chimeras identified from screen	8	12	50.0 ± 0.2	48.6 ± 3.1	0.48 ± 0.06
	17	34	52.3 ± 0.3	50.0 ± 2.5	0.46 ± 0.03
	17	111	53.9 ± 1.2	52.5 ± 1.1	0.74 ± 0.04
	20	32	54.5 ± 0.6	50.8 ± 3.4	0.20 ± 0.05
	27	86	54.6 ± 1.3	57.5 ± 1.1	0.33 ± 0.02
	22	90	55.8 ± 0.5	56.2 ± 0.7	0.40 ± 0.02
	26	68	55.9 ± 0.8	51.1 ± 0.8	0.40 ± 0.03
	20	72	56.9 ± 0.2	58.2 ± 2.1	0.40 ± 0.04
	32	142	57.2 ± 0.3	57.5 ± 1.1	0.45 ± 0.03
	24	89	57.8 ± 0.4	62.8 ± 1.3	0.35 ± 0.01
	21	100	58.4 ± 0.4	57.5 ± 1.9	0.53 ± 0.01
	32	141	58.7 ± 0.5	61.0 ± 1.2	0.34 ± 0.03
	36	129	59.2 ± 1.4	59.6 ± 3.1	0.37 ± 0.08
	32	108	59.5 ± 0.5	58.6 ± 0.4	0.33 ± 0.06
	22	126	59.6 ± 1.0	59.3 ± 1.2	0.53 ± 0.10
	21	120	60.1 ± 0.6	57.5 ± 1.1	0.53 ± 0.02
	26	120	61.6 ± 1.2	62.8 ± 3.3	0.57 ± 0.02
	9	86	61.7 ± 0.2	59.6 ± 2.0	0.86 ± 0.01
	21	99	61.9 ± 1.1	58.8 ± 2.5	0.38 ± 0.02
	30	106	63.4 ± 0.4	70.8 ± 1.6	0.22 ± 0.04
	31	117	63.5 ± 0.4	67.5 ± 1.0	0.21 ± 0.04
	34	110	64.0 ± 0.2	68.7 ± 3.7	0.48 ± 0.16
	24	87	64.1 ± 0.4	68.0 ± 0.9	0.46 ± 0.01
	13	93	64.1 ± 0.6	70.5 ± 1.2	0.37 ± 0.02
	28	111	65.4 ± 1.2	68.0 ± 4.7	0.51 ± 0.03
	13	64	66.2 ± 0.4	59.9 ± 2.3	0.49 ± 0.02
	15	63	66.4 ± 0.8	71.4 ± 0.6	0.51 ± 0.06
	21	99	66.9 ± 0.4	65.9 ± 0.4	0.48 ± 0.02
	19	66	67.7 ± 0.2	67.1 ± 2.1	0.60 ± 0.01
	24	85	68.1 ± 0.3	67.2 ± 5.4	0.51 ± 0.03
	14	95	68.5 ± 0.1	71.8 ± 0.1	0.58 ± 0.01
	18	54	68.5 ± 1.1	71.4 ± 0.6	0.53 ± 0.01
	15	88	70.3 ± 0.8	70.2 ± 2.3	0.51 ± 0.07
	25	121	70.6 ± 0.6	74.7 ± 1.6	0.43 ± 0.01
	12	92	71.4 ± 0.4	75.9 ± 0.1	0.43 ± 0.02
	14	52	71.7 ± 0.5	68.8 ± 1.1	0.48 ± 0.01
	23	111	71.8 ± 1.0	75.3 ± 0.8	0.91 ± 0.07
	28	108	72.1 ± 0.6	68.8 ± 0.8	0.60 ± 0.08
	21	65	72.1 ± 0.6	77.9 ± 0.9	0.65 ± 0.04
	12	25	72.6 ± 0.3	76.5 ± 0.4	0.78 ± 0.08
	39	69	73.2 ± 1.1	69.5 ± 1.8	0.10 ± 0.03
	27	74	73.3 ± 0.4	75.5 ± 1.1	0.74 ± 0.07
	22	77	73.7 ± 0.2	74.7 ± 2.1	0.60 ± 0.05
	11	42	73.8 ± 0.3	72.1 ± 0.8	0.55 ± 0.03
	21	78	74.0 ± 0.4	77.4 ± 1.6	0.55 ± 0.01
	6	12	74.0 ± 1.0	70.8 ± 1.9	0.78 ± 0.03
	20	121	75.3 ± 0.2	76.1 ± 2.0	0.60 ± 0.01
	16	99	75.7 ± 0.2	75.5 ± 1.1	0.41 ± 0.02
	11	31	75.7 ± 0.5	76.1 ± 0.1	0.32 ± 0.10
	43	115	77.3 ± 0.5	74.7 ± 1.7	0.18 ± 0.07
predicted active stable chimeras	29	97	72.7 ± 0.1	74.6 ± 1.1	0.11 ± 0.02
	13	73	73.3 ± 0.2	71.2 ± 0.8	0.04 ± 0.02
	20	54	73.5 ± 0.4	73.8 ± 0.4	0.14 ± 0.02
	9	28	76.0 ± 1.9	72.8 ± 0.6	0.06 ± 0.03
	16	71	77.5 ± 0.5	75.2 ± 1.0	0.58 ± 0.03
	21	82	77.8 ± 0.4	75.8 ± 0.1	0.11 ± 0.02
	14	52	78.2 ± 0.4	79.3 ± 1.1	1.37 ± 0.07
	3	9	78.4 ± 0.3	75.8 ± 0.1	0.41 ± 0.04
	19	63	78.9 ± 0.3	80.5 ± 0.6	1.48 ± 0.05
	8	20	78.9 ± 0.2	75.8 ± 0.1	0.08 ± 0.03

Figure 3.4: Representation of three Cel48 parents and 60 active chimeras, with CelF in white, CelY in gray, and CelS in black. SCHEMA E values, number of mutations from closest parent (m), T_{50}, T_{opt}, and A_{rel} are also provided. T_{50} is the temperature at which an enzyme loses 50% of its activity in a 10-minute incubation. T_{opt} is the temperature at which a cellulase liberates the most glucose from crystalline cellulose in a 2-hour hydrolysis assay. A_{rel} is the cellulases specific activity at its respective optimal temperature measured in a 1-hour assay with 0.2 μM enzyme and 0.2 μM miniscaffoldin in 10 g/L Avicel. Values are normalized relative to the specific activity of CelS.

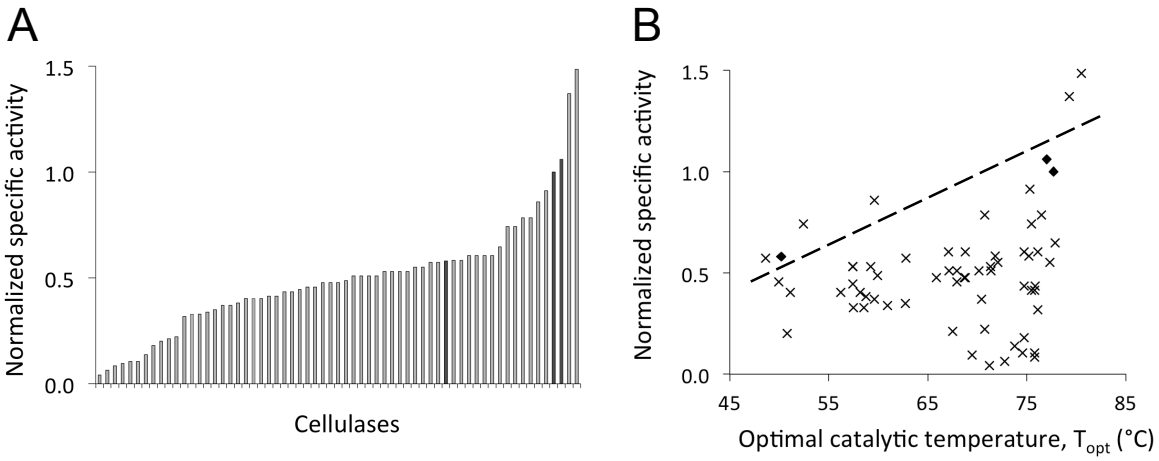


Figure 3.5: Specific activities of chimeric cellulases. A) Specific activities of the Cel48 enzymes at their respective optimal catalytic temperatures (tabulated in Figure 4). The activities are measured in a 1-hour assay, with 0.2 μM enzyme and 0.2 μM miniscaffoldin in 10 g/L Avicel at the respective optimal catalytic temperature. The activities are normalized to the maximum specific activity of CelS (T_{opt} = 77.7°C). The parent enzymes are highlighted in bold. B) The normalized specific activities versus the optimal catalytic temperatures of the cellulases. The parent enzymes are highlighted as black diamonds and the possible correlation among the most active cellulases is indicated with a dotted line.

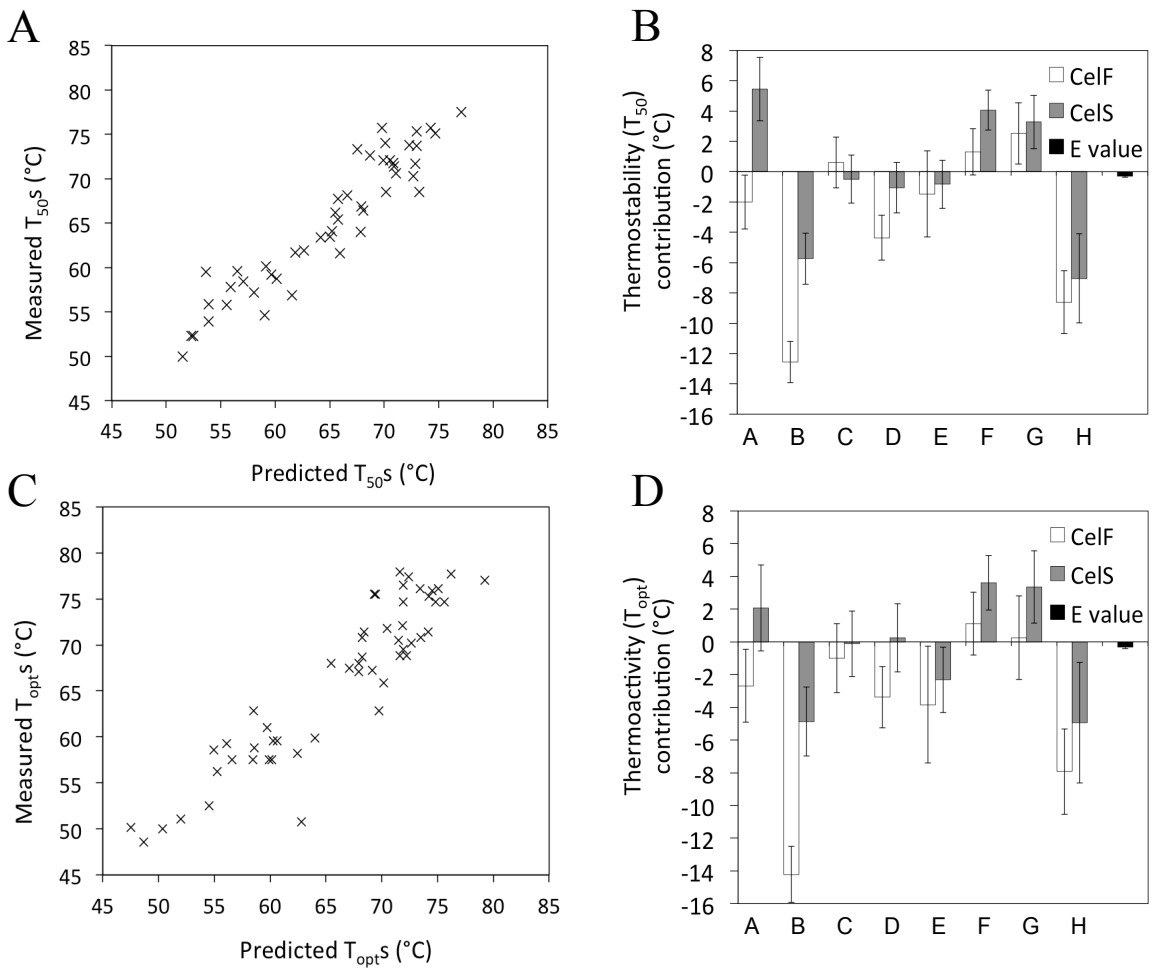


Figure 3.6: Modeling thermostability and thermoactivity. A) Predicted T_{50} values from a simple linear model closely correlate with the measured T_{50} for 53 bacterial family 48 cellulases over a range of almost 30°C. B) Stabilizing or destabilizing effects of each sequence block, for CelF (gray) and CelS (white), relative to CelY for the T_{50} model. Most blocks are destabilizing with respect to the most thermostable parent, CelY. Blocks A, F and G from CelS and, to a lesser extent, blocks C, F and G from CelF are predicted to be stabilizing. Effect of the SCHEMA E value on the T_{50} predictions is -0.29°C per disrupted structural contact (black). C) Predicted T_{opt} values from the same linear model also correlate with the measured T_{opt} over a similar range. D) Stabilizing or destabilizing effects of each block, for CelF (gray) and CelS (white), relative to CelY for the T_{opt} model. Block contributions are similar in magnitude to those in the T_{50} model.

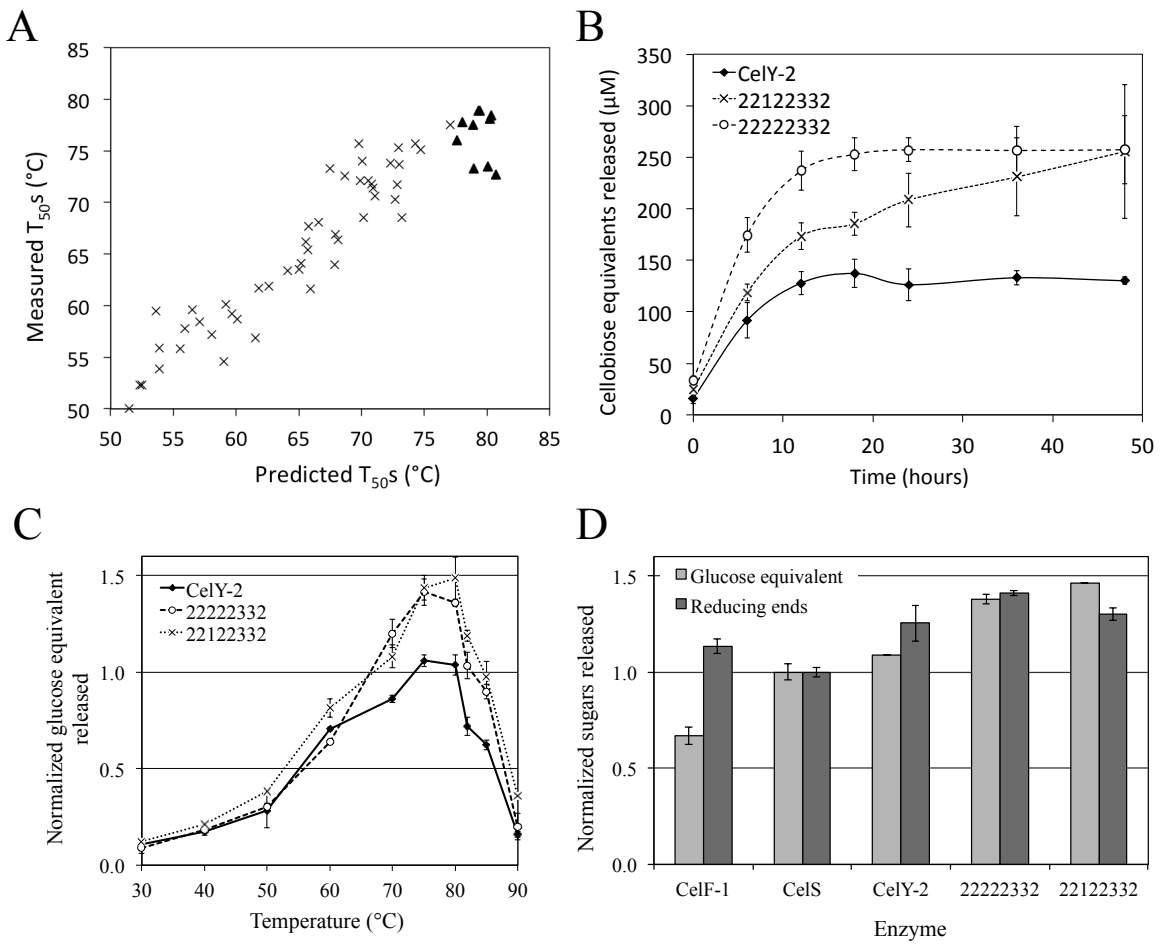


Figure 3.7: Predicting the most stable Cel48 chimeras. A) T_{50} model trained on all 53 active parent and chimeric test cellulases (crosses) was used to predict 10 very stable chimeras that were subsequently constructed. All 10 are very stable (triangles). B) Activities of the two most stable, most active chimeras and the most stable, most active cellulosomal parent sequence, CelY-2. Activities were measured in the form of reducing-end sugars released (reported as cellobiose equivalents released) over a 48-hour period, with 0.2 μM enzyme and 0.2 μM miniscaffoldin in 10 g/L Avicel at 75°C. All measurements were carried out in triplicate. C) Temperature-activity profiles for the two most stable, most active chimeras and the most stable, most active cellulosomal parent sequence, CelY-2. Activities were measured in a 1-hour assay, with 0.2 μM enzyme and 0.2 μM miniscaffoldin in 10 g/L Avicel. The activities are normalized to the maximum activity of CelS. D) The maximum activities of the three parent constructs and two of the most stable, most active chimeras. The activities are measured for a 1-hour assay, with 0.2 μM enzyme and 0.2 μM miniscaffoldin in 10 g/L Avicel. The activities are normalized to the maximum activity of CelS. Activities are measured both by the number of reducing-end sugars released and the total glucose released.

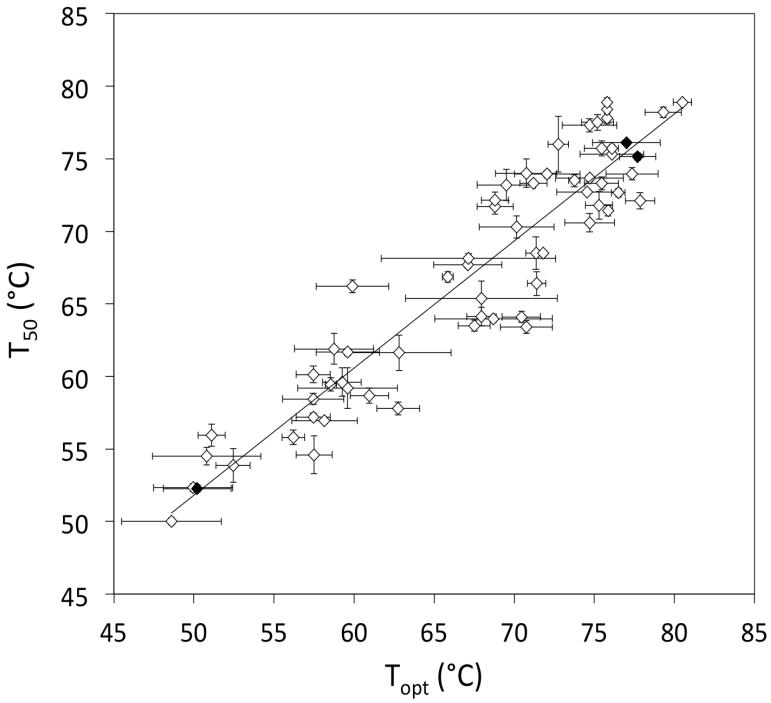


Figure 3.8: The correlation between optimum operating temperature for a 2-hour assay (T_{opt}) and thermostability (T_{50}) for all 63 chimeric and parent Cel48 cellulases in this study. There is a strong correlation ($r^2 = 0.83$): chimeras with greater stability tend to be most active at higher temperatures. The parents are highlighted in black.

3.7 References

1. Lynd L. R., Weimer P. J., van Zyl W. H., and Pretorius I. S. (2008) Microbial cellulose utilization: Fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 506-577.

2. Blum D. L., Kataeva I. A., Li L., and Ljungdahl L. G. (2000) Feruloyl esterase activity of the Clostridium thermocellum cellulosome can be attributed to previously unknown domains of XynY and XynZ. *J. Bacteriol.* **182**, 1346-1351.

3. Tamaru Y., and Doi R. H. (2001) Pectate lyase A, an enzymatic subunit of the Clostridium cellulovorans cellulosome. *Proc. Natl. Acad. Sci. USA* **98**, 4125-4129.

4. Fierobe H. P., Bayer E. A., Tardif C., Czjzek M., Mechaly A., Belaich A., Lamed R., Shoham Y., and Belaich J. P. (2002) Degradation of cellulose substrates by cellulosome chimeras. *J. Biol. Chem.* **277**, 49621-49630.

5. Boraston A. B., Bolam D. N., Gilbert H. J., and Davies G. J. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **382**, 769-781.

6. Mitsuzawa S., Kagawa H., Li Y., Chan S. L., Paavola C. D., and Trent J. D. (2009) The rosettazyme: a synthetic cellulosome. *J. Biotechnol.* **143**, 139-144.

7. Olsen D. G., Tripathi S. A., Giannone R. J., Lo J., Caiazza N. C., Hogsett D. A., Hettich R. L., Guss A. M., Dubrovsky G., and Lynd L. R. (2010) Deletion of the Cel48S cellulase from Clostridium thermocellum. *Proc. Natl. Acad. Sci. USA* **107**, 17727-17732.

8. Vazana Y., Morais S., Barak Y., Lamed R., and Bayer E. A. (2010) Interplay between

Clostridium thermocellum family 48 and family 9 cellulases in cellulosomal versus noncellulosomal states. *Appl. Environ. Microbiol.* **76**, 3236-3243.

9. Kruus K., Wang W. K., Chiu P. C., Ching J. T., Wang T. Y., and Wu J. H. D. (1994) CelS - A major exoglucanase component of Clostridium thermocellum cellulosome. In: Himmel M. E., Baker J. O., and Overend R. P. (eds.) *Enzymatic conversion of biomass for fuels production*, American Chemical Society, Washington, D.C.

10. Reverbel-Leroy C., Pages S., Belaich A., Belaich J. P., and Tardif C. (1997) The processive endocellulase CelF, a major component of the Clostridum cellulolyticum cellulosome: purification and characterization of the recombinant form. *J. Bacteriol.* **179**, 46-52.

11. Cantarel B. L., Coutinho P. M., Rancurel C., Bernard T., Lombard V., and Henrissat B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, 233-238.

12. Voigt C. A., Martinez C., Wang Z. G., Mayo S. L., and Arnold F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553-558.

13. Otey C. R., Landwehr M., Endelman J. B., Hiraga K., Bloom J. D., and Arnold F. H. (2006) Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol.* **4**, 789-798.

14. Meyer M. M., Silberg J. J., Voigt C. A., Endelman J. B., Mayo S. L., Wang Z. G., and Arnold F. H. (2003) Library analysis of SCHEMA-guided protein recombination. *Protein Sci.* **12**, 1686-1693.

15. Heinzelman P., Snow C. D., Wu I., Nguyen C., Villalobos A., Govindarajan S., Min-

shull J., and Arnold F. H. (2009) A family of thermostable fungal cellulases created by structure-guided recombination. *Proc. Natl. Acad. Sci. USA* **106**, 5610-5615.

16. Heinzelman P., Komor R., Kanaan A., Romero P., Yu X., Mohler S., Snow C. D., and Arnold F. H. (2010) Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Eng. Des. Sel.* **23**, 871-880.

17. Li Y., Drummond D. A., Sawayama A. M., Snow C. D., Bloom J. D., and Arnold F. H. (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25**, 1051-1056.

18. Bronnenmeier K., Kundt K., Riedel K., Schwarz W. H., and Staudenbauer W. L. (1997) Structure of the Clostridium stercorarium gene celY encoding the exo-1,4-β-glucanase Avicelase II. *Microbiol.* **143**, 891-898.

19. Wang W. K., Kruus K., and Wu J. H. D. (1993) Cloning and DNA sequence of the gene coding for Clostridium thermocellum cellulase Ss (CelS), a major cellulosome component. *J. Bacteriol.* **175**, 1293-1302.

20. Mingardon F., Bagert J. D., Maisonnier C., Trudeau D. L., and Arnold F. H. (2011) Comparison of family 9 cellulases from mesophilic and thermophilic bacteria. *Appl. Environ. Microbiol.* **77**, 1436-1442.

21. Fierobe H. P., Mingardon F., Mechaly A., Belaich A., Rincon M. T., Pages S., Lamed R., Tardif C., Belaich J. P., and Bayer E. A. (2005) Action of designer cellulosomes on homogeneous versus complex substrates. *J. Biol. Chem.* **280**, 16325-16334.

22. Endelman J. B., Silberg J. J., Wang Z. G., and Arnold F. H. (2004) Site-directed

protein recombination as a shortest path problem. *Protein Eng. Des. Sel.* **17**, 589-594.

23. Hiraga K. and Arnold F. H. (2003) General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.* **330**, 287-296.

24. Kruus K., Wang W., Ching J., and Wu J. (1995) Exoglucanase activities of the recombinant Clostridium thermocellum CelS, a major cellulosome component. *J. Bacteriol.* **177**, 1641-1644.

25. Tuka K., Zverlov V. V., Bumazkin B. K., Velikodvorskaya G. A., and Ya Strongin A. (1990) Cloning and expression of Clostridium thermocellum genes coding for thermostable exoglucanases (cellobiohydrolases) in Escherichia coli cells. *Biochem. Biophys. Res. Comm.* **169**, 1055-1060.

26. Berger E., Zhang D., Zverlov V. V., and Schwarz W. H. (2007) Two noncellulosomal cellulases of Clostridium thermocellum, Cel9I and Cel48Y, hydrolyse crystalline cellulose synergistically. *FEMS Microbiol. Lett.* **268**, 194-201.

27. Irwin D. C., Zhang S., and Wilson D. B. (2000) Cloning, expression and characterization of a family 48 exocellulase, Cel48A, from Thermobifida fusca. *Eur. J. Biochem.* **267**, 4988-4997.

28. Bronnenmeier K., Rucknagel K. P., and Staudenbauer W. L. (1991) Purification and properties of a novel type of exo-1,4-β-glucanase (Avicelase II) from the cellulolytic thermophile Clostridium stercorarium. *Eur. J. Biochem.* **200**, 379-385.

29. Zverlov V.V., Mahr S., Riedel K., and Bronnenmeier K. (1998) Properties and gene structure of a bifunctional cellulolytic enzyme (CelA) from the extreme thermophile

'Anaerocellum thermophilum' with separate glycosyl hydrolase family 9 and 48 catalytic domains. *Microbiol.* **144**, 457-465.

30. Hammel M., Fierobe H. P., Czjzek M., Finet S., and Receveur-Brechot V. (2004) Structural insights into the mechanism of formation of cellulosomes probed by small angle X-ray scattering. *J. Biol. Chem.* **279**, 55985-55994.

31. Daniel R. M. and Danson M. J. (2010) A new understanding of how temperature affects the catalytic activity of enzymes. *Trends Biochem. Sci.* **35**, 584-591.

32. Park J. T. and Johnson M. J. (1949) A submicro determination of glucose. *J. Biol. Chem.* **181**, 149-151.

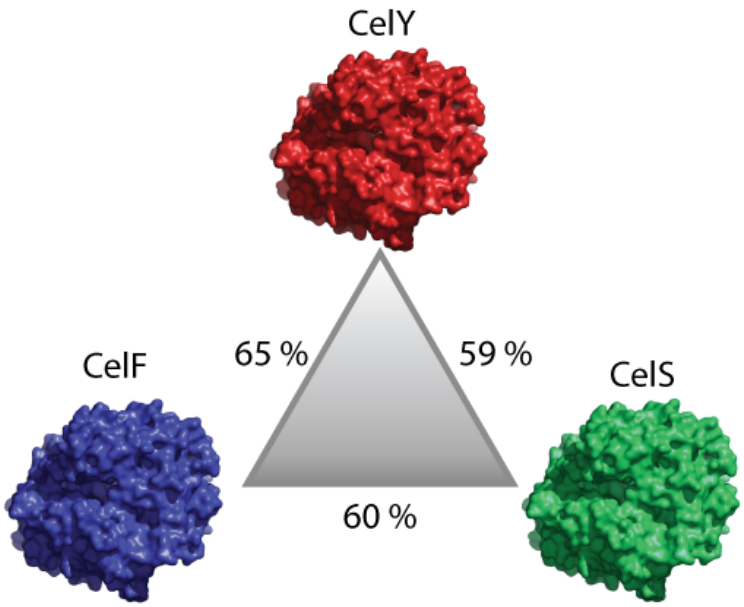
33. Friedman J., Hastie T., Hofling H., and Tibshirani R. (2007) Pathwise coordinate optimization. *Ann. Appl. Stat.* **1**, 302-332.

34. Friedman J., Hastie T., and Tibshirani R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1-22.

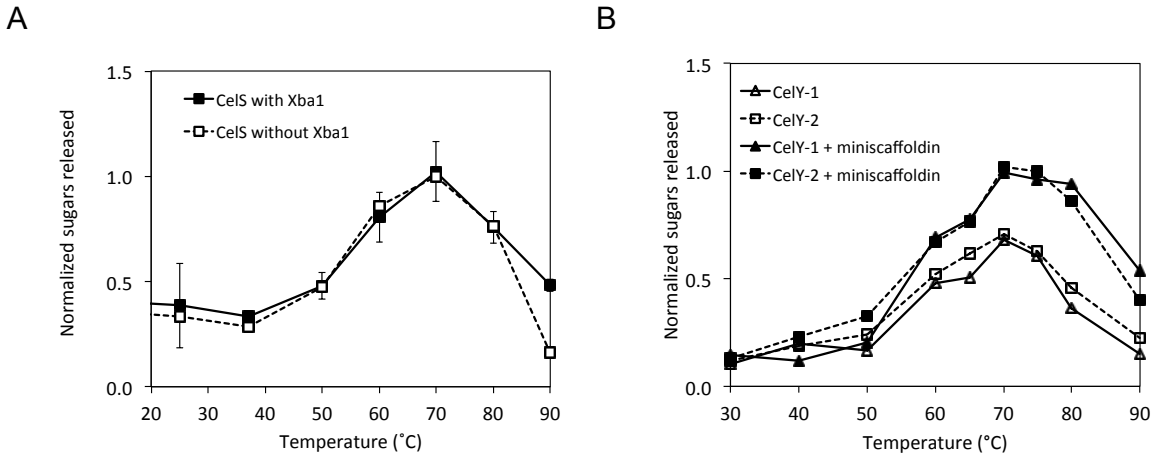
3.8 Supplementary information

```
CATATGGCCAGCAGCGATGATCCGTATAAGCAACGTTTCTTGGAACTGTGGGAAGAGTTGCACGATCCGAGCAACGGTTATTTCAG
CTCCCATGGTATTCCGTACCACGCGGTCGAGACGCTGATCGTTGAGGCACCTGATTATGGCCACCTGACCACCAGCGAAGCGATGT
CTTACTATCTGTGGCTGGAAGCGCTGTACGGCAAATTTACGGGTGATTTTAGCTATTTCATGAAGGCCTGGGAAACCATTGAGAAG
TACATGATTCCGACCGAGCAGGATCAACCGAACCGCTCCATGGCTGGTTACAATCCGGCTAAACCAGCGACCTATGCCCCTGAATG
GGAAGAACCGAGCATGTATCCGTCTCAGCTGGACTTCAGCGCACCGGTGGGCATTGACCCGATTTACAATGAGCTGGTGTCCACCT
ATGGTACCAATACGATTTACGGTATGCACTGGCTGCTGGATGTGGATAACTGGTACGGCTTTGGCCGTCGTGCGGACCGTATCAGC
AGCCCAGCCTATATCAACACCTTCCAACGTGGCAGCCAAGAGTCCGTGTGGGAGACGATCCCGCAACCGTGCTGGGATGATCTGAC
CATCGGTGGCCGTAACGGTTTTCTGGACCTGTTTGTCGGCGATAGCCAGTACTCGGCACAATTTAAGTACACGAATGCACCGGACG
CGGATGCGCGTGCCATCCAGGCGACGTACTGGGCGAACCAGTGGGCGAAAGAGCACGGCGTGAATTTGAGCCAGTATGTTAAGAAG
GCAAGCCGCATGGGCGACTACCTGCGCTATGCAATGTTCGACAAATACTTTCGTAAAATTGGTGATTCCAAACAAGCAGGTACCGG
CTACGACGCAGCCCATTACCTGCTGTCCTGGTACTATGCGTGGGGTGGTGGCATCACGGCTGATTGGGCATGGATTATTGGCTGTT
CCCACGTTCATGCAGGCTACCAGAATCCGATGACGGCGTGGATTCTGGCCAACGATCCGGAGTTTAAACCGGAAAGCCCGAACGGT
GCTAATGATTGGGCGAAAAGCCTGGAGCGCCAGCTGGAGTTCTATCAATGGCTGCAGAGCGCTGAGGGTGCAATCGCAGGTGGTGC
GACGAATAGCTACAAAGGTCGCTACGAAACCCTGCCAGCAGGTATCAGCACGTTCTATGGCATGGCGTATGAAGAACATCCGGTGT
ACCTGGATCCGGGTAGCAACACGTGGTTTGGCTTTCAGGCGTGGACGATGCAGCGCGTGGCGGAATACTACTATCTGACCGGTGAT
ACGCGTGCAGAGCAACTGTTGGACAAATGGGTCGATTGGATCAAGTCCGTTGTTCGTCTGAACAGCGACGGCACCTTCGAGATTCC
GGGTAACCTGGAGTGGTCGGGTCAACCGGACACCTGGACCGGTACTTACACGGGTAATCCGAACCTGCATGTCAGCGTTGTTTCTT
ATCGTACGGACTTGGGTGCAGCGGGTTCTCTGGCAAATGCTCTGCTGTACTATGCCAAAACCAGCGGTGACGACGAAGCACGTAAT
CTGGCGAAAGAATTGCTGGACCGTATGTGGAACCTGTACCGTGACGACAAAGGTTTGTCCGCACCGGAGACTCGCGAAGATTACGT
CCGCTTTTTCGAACAAGAGGTTTACGTTCCACAGGGTTGGTCTGGTACGATGCCTAACGGCGATCGTATCGAACCGGGTGTTACTT
TCCTGGACATCCGCTCGAAATACCTGAACGACCCGGACTACCCGAAGCTGCAGCAGGCGTATAACGAAGGCAAAGCGCCAGTGTTC
AACTATCACCGTTTCTGGGCTCAATGCGACATCGCTATCGCGAACGGCTTGTATAGCATTCTGTTTGGCAGCGAGCAAGCCAATGA
TAGCTTCATCACCCCGACCAGCGCGACGTTCGACAAGAATAACCAGGAAGACATTTCTGTTACGGTCACCTACAATGGTAATACCC
TGCTGGGCATCAAGAGCGGTAGCAGCTATCTGATTGAGGGTGTCGACTACATTGTGAACGGCGATGTGATTATCATTAAGAAAGAA
TTTCTGGCAGGCCAGGCTACCGGCAGCATTAGCCTGCTGTTCGATTTCAGCGCAGGTCTGGACCGCACCCTGACCATTGATATTAT
CGATACGGGTGGCGGTGAAGAACCTGTCGAGCCGGTGGAGCCTGTGGAGGGCGTCCTGATCATCCAAAGCTTCAATGCCAACACTC
AAGAGATTAGCAACTCGATCATGCCACGTTTCCGTATCTACAATAGCGGCAATACCAGCATTCCGTTGAGCGAGGTCAAGTTGCGC
TATTACTACACCGTGGACGGTGACAAGCCGCAGAACTTCTGGTGTGACTGGGCGAGCATTGGTAGCAGCAATGTGACTGGCACCTT
TGTTAAGATGGATGGTGCGACTACCGGTGCCGATTATTATCTGGAGATTGGCTTCACCCCACAGGCTGGTACGCTGGAACCGGGTG
CAAGCATCGAGGTCCAGGGTCGTTTTAGCAAGATTGACTGGACCGACTACACCCAAACCAATGACTACAGCTTTAATCCGACCGCG
TCTAGCTATGTTGACTTTAACAAGATCACCGCGTACATCAGCGGTAATCTGGTTTATGGTATCGAGCCGTGAGCGGCCGC
```

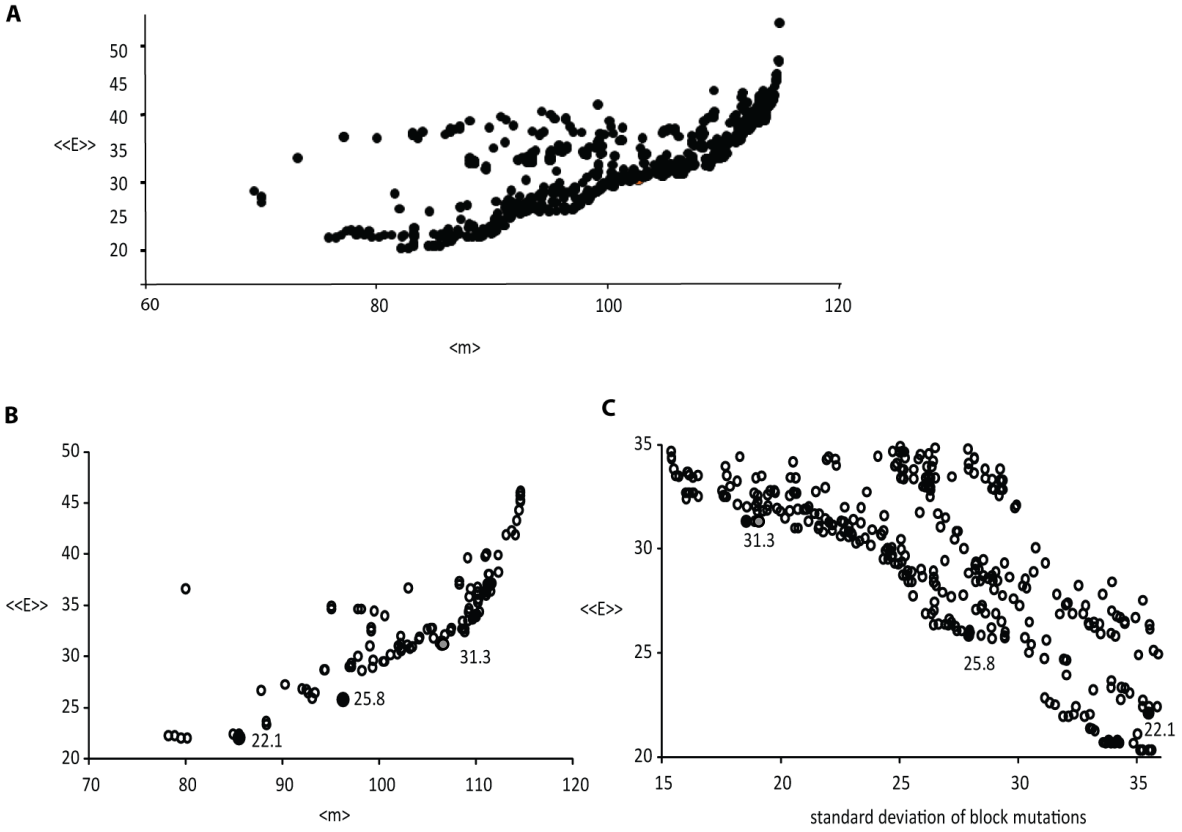
Supplementary Figure 3.9: CelY synthetic gene complete sequence (synthesized by DNA2.0, Menlo Park, CA, USA)



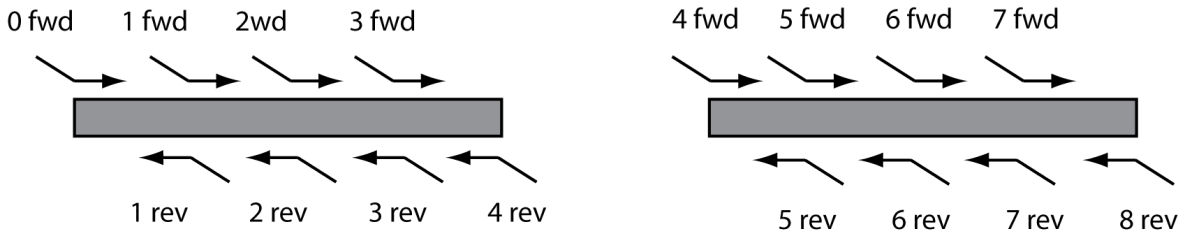
Supplementary Figure 3.10: Catalytic domains of CelY, CelF, and CelS and their sequence identities.



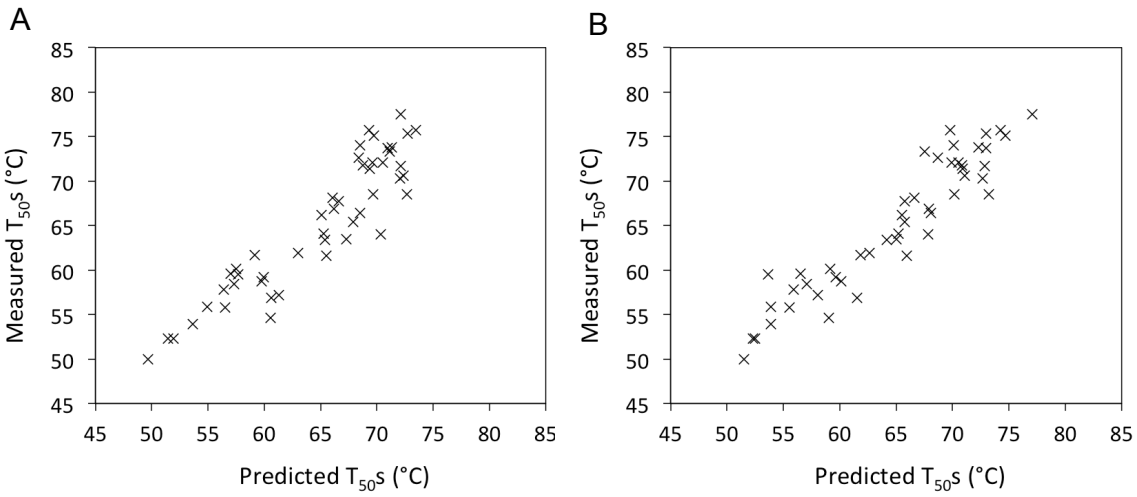
Supplementary Figure 3.11: CelS Xba1 site and CelY DUF do not affect cellulolytic activities. A) Comparison of CelS with (dashed line) and without (solid line) an additional XbaI site between the catalytic domain and the dockerin in the presence of miniscaffoldin. Measurements were taken in duplicate from a 2-hour end point assay. B) Comparison of cellulosomal constructs CelY-1 (triangles) and CelY-2 (squares) in the presence (filled) and absence (empty) of miniscaffoldin. Measurements were taken as duplicates from a 2-hour assay.



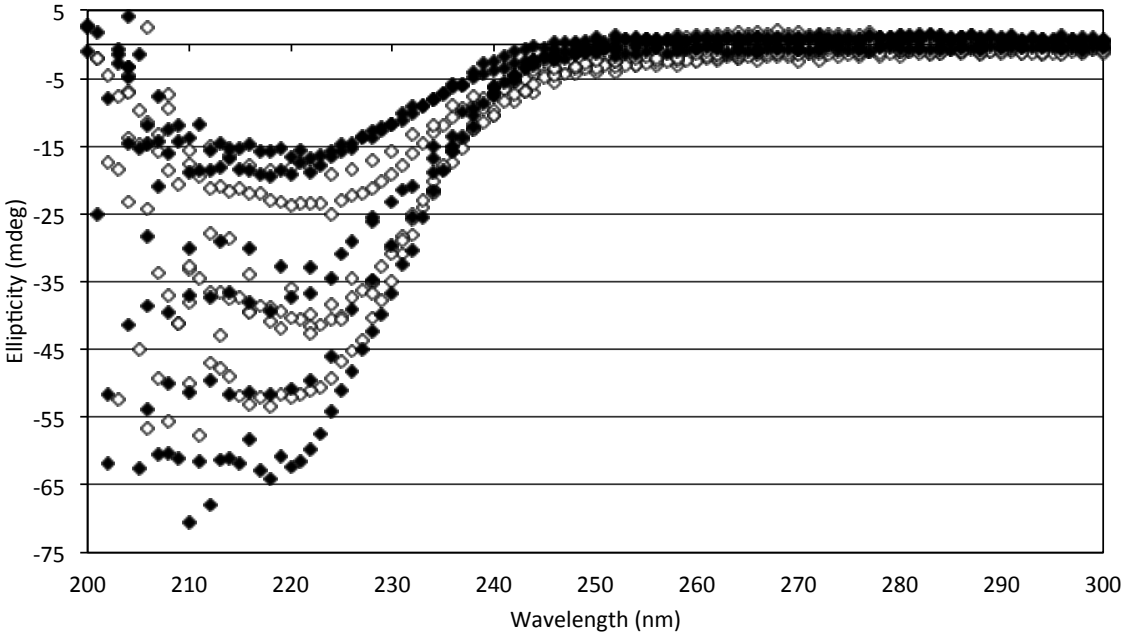
Supplementary Figure 3.12: SCHEMA library designs with eight blocks and a minimum block length of 30 amino acids. Multiple RASPP curves are calculated from contacts maps based on 17 available structures. $<< E >>$ is the average $< E >$ for a certain library based on different contact maps. A) Initially RASPP returned more than 600 libraries and there were several solutions with comparable characteristics. B) Solutions are removed that did not contain the same amino acid at the designated crossover sites. The chosen library has an $<< E >>$ of 31.3 (highlighted in gray). C) $<< E >>$ as a function of the standard deviation of block mutations. A low standard deviation of block mutations indicates a more homogenous distribution of mutations and more equal contribution of each block. The chosen library ($<< E >> = 31.3$) has a standard deviation of block mutations of 19 and is highlighted in gray.



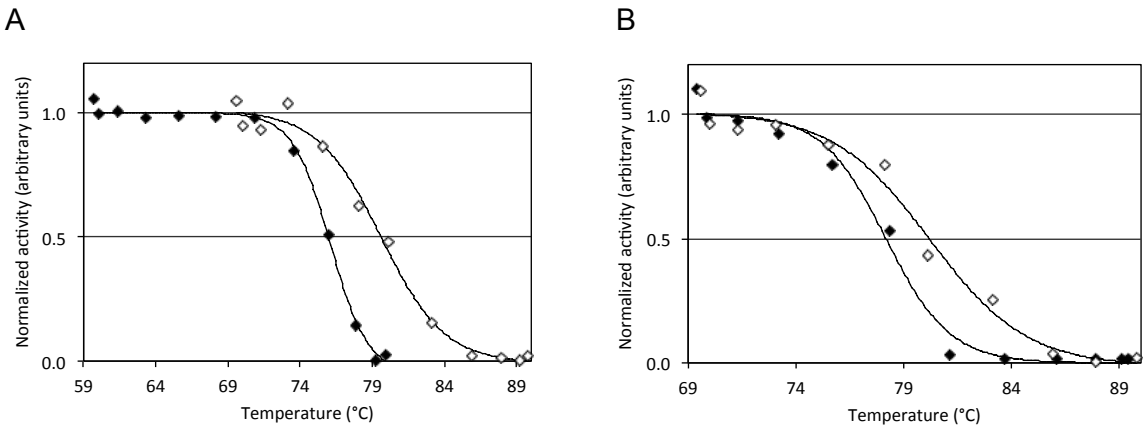
Supplementary Figure 3.13: Naming scheme for primers used for SCHEMA library construction. Each primer is named by the parent species (CCEL for CelF, CSTE for CelY, or CTHE for CelS), the crossover site, and direction, as indicated.



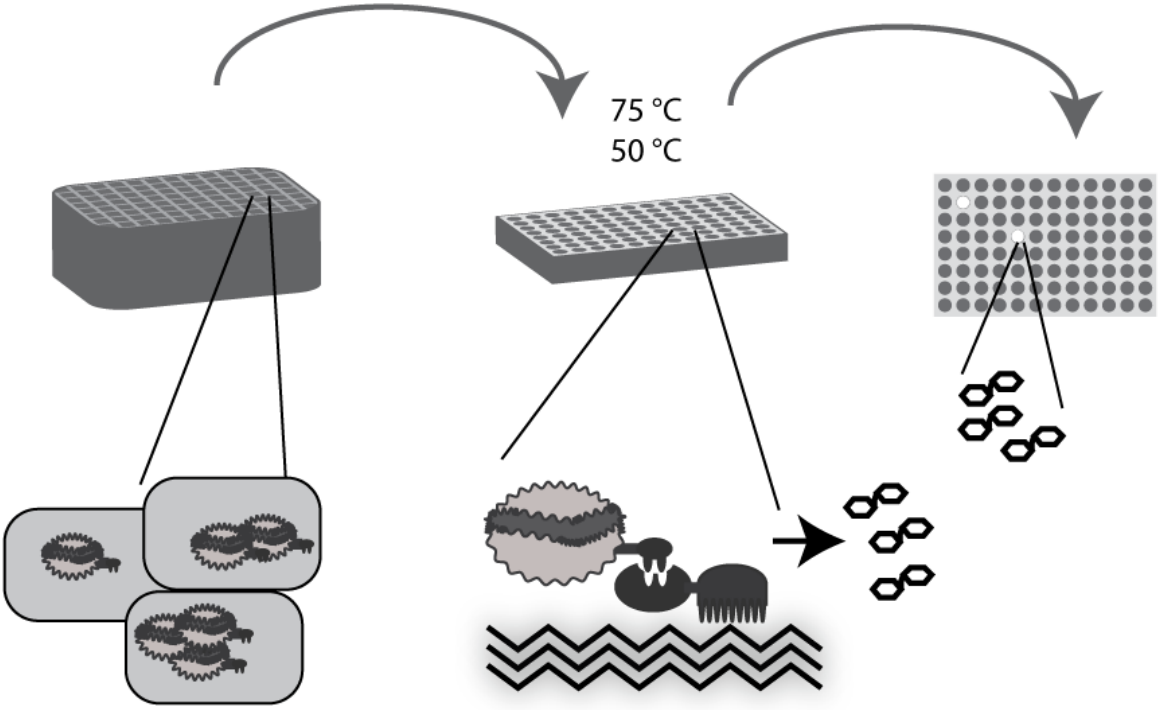
Supplementary Figure 3.14: Comparison of T_{50} thermostability models for the three parents and 50 chimeras. A) Simple thermostability model without SCHEMA E value as a parameter, $r^2 = 0.82$. B) Modified thermostability model with the SCHEMA E value as a parameter, $r^2 = 0.88$.



Supplementary Figure 3.15: CD measurements of several functional (filled diamonds) and non-functional (open diamonds) chimeras. The ellipticity profiles are similar for both the functional and non-functional chimeras.



Supplementary Figure 3.16: Examples of T_{50} measurements, with incubation in the absence (filled diamonds) and presence (open diamonds) of miniscaffoldin and substrate. Binding to the cellulose substrate increases the T_{50} of the cellulases by approx. 2°C. A) CelY-2. B) Chimera 22222332.



Supplementary Figure 3.17: High-throughput screen for activity of Cel48 chimeras on crystalline cellulose. Cellulase chimeras are overexpressed in *E. coli* in 96-well plates and harvested. Cell lysate is transferred to a solution of miniscaffoldin on crystalline cellulose. In our setup, cellulases bind to the cohesin of the miniscaffoldin via their dockerin. The whole complex attaches to the solid substrate via the CBM of the miniscaffoldin. Non-binding proteins are removed by washing thereby purifying cellulases bound to the solid substrate. Cellulases are incubated overnight at both 50°C and 75°C. Enzymatic hydrolysis of cellulose releases soluble sugars. These sugars are transferred to a new plate and the amount of reducing ends is determined using the Park-Johnson assay.

Block A	CelS	YKDLFLELYGKIKDPKNGYFSPDEGIPYHSIETLIVEAPDYGHVTTSEAFSYYVWLEAMY
		GNLTGNWSGVETAWKVMEDWII
	CelF	YQDRFESMYSKIKDPANGYFS-EQGIPYHSIETLMVEAPDYGHVTTSEAMSYYMWLEAMH
		GRFSGDFTGFDKSWSVTEQYLI
	CelY	YKQRFLELWEELHDPSNGYFS-SHGIPYHAVETLIVEAPDYGHLTTSEAMSYYLWLEALY
		GKFTGDFSYFMKAWETIEKYMI
Block B	CelS	PDSTEQP--GMSSYNPNSPATYADEYEDPSYYPSELKFDTVRVGSDPVHNDLVSAY-GPN
		MYLMHWLMDVDNWYGFG----TGTRATFINTFQRGEQESTWETIPHPSIEEFKYGGPNGF
		LDLFTKDR-SYAKQWRYTNAPDAEGR
	CelF	PTEKDQPNTSMSRYDANKPATYAPEFQDPSKYPSPLDT-SQPVGRDPINSQLTSAYGTSM
		LYGMHWILDVDNWYGFGARADGTSKPSYINTFQRGEQESTWETIPQPCWDEHKFGGQYGF
		LDLFTKDTGTPAKQFKYTNAPDADAR
	CelY	PTEQDQPNRSMAGYNPAKPATYAPEWEEPSMYPSQLDF-SAPVGIDPIYNELVSTYGTNT
		IYGMHWLLDVDNWYGFGRRADRISSPAYINTFQRGSQESVWETIPQPCWDDLTIGGRNGF
		LDLFVGDS-QYSAQFKYTNAPDADAR
Block C	CelS	AIQAVYWANKWAKEQGKGSAVASVVSKAAKMG
	CelF	AVQATYWADQWAKEQGK-S-VSTSVGKATKMG
	CelY	AIQATYWANQWAKEHGV-N-LSQYVKKASRMG
Block D	CelS	DFLRNDMFDKYFMKIGAQDKTPATGYDSAHYLMAWYTAWGGGIGASWAWKIGCSHA
	CelF	DYLRYSFFDKYFRKIGQPS-QAGTGYDAAHYLLSWYYAWGGGIDSTWSWIIGSSHN
	CelY	DYLRYAMFDKYFRKIG-DSKQAGTGYDAAHYLLSWYYAWGGGITADWAWIIGCSHV
Block E	CelS	HFGYQNPFQGWVSATQSDFAPKSSNGKRDWTTSYKRQLEFYQWLQSAE
	CelF	HFGYQNPFAAWVLSTDANFKPKSSNGASDWAKSLDRQLEFYQWLQSAE
	CelY	HAGYQNPMTAWILANDPEFKPESPNGANDWAKSLERQLEFYQWLQSAE
Block F	CelS	GGIAGGATNSWNGRYEKYPAGTSTFYGMAYVPHPVYADPGS
	CelF	GAIAGGATNSWNGRYEAVPSGTSTFYGMGYVENPVYADPGS
	CelY	GAIAGGATNSYKGRYETLPAGISTFYGMAYEEHPVYLDPGS
Block G	CelS	NQWFGFQAWSMQRVMEYYLETGDSSVKNLIKKWVDWVMSEIKLYDDGTFAIPSDLEWSGQ
		PDTWTG--TYTGNPNLHVRVTSYGTDLGVAGSLANALATYAAATERWEGKLDTKARDMAA
		E
	CelF	NTWFGMQVWSMQRVAELYYKTGDARAKKLLDKWAKWINGEIKFNADGTFQIPSTIDWEGQ
		PDTWNPTQGYTGNANLHVKVVNYGTDLGCASSLANTLTYYAAKS------GDETSRQNAQ
		K
	CelY	NTWFGFQAWTMQRVAEYYYLTGDTRAEQLLDKWVDWIKSVVRLNSDGTFEIPGNLEWSGQ
		PDTWTG--TYTGNPNLHVSVVSYRTDLGAAGSLANALLYYAKTS------GDDEARNLAK
		E
Block H	CelS	LVNRAWYNFYCSEGKGVVTEEARADYKRFFEQEVYVPAGWSGTMPNGDKIQPGIKFIDIR
		TKYRQDPYYDIVYQAYLRGEAPVLNYHRFWHEVDLAVAMGVLATYF
	CelF	LLDAMWNNYSD--SKGISTVEQRGDYHRFLDQEVFVPAGWTGKMPNGDVIKSGVKFIDIR
		SKYKQDPEWQTMVAALQAGQVPTQRLHRFWAQSEFAVANGVYAILF
	CelY	LLDRMWNLYRD--DKGLSAPETREDYVRFFEQEVYVPQGWSGTMPNGDRIEPGVTFLDIR
		SKYLNDPDYPKLQQAYNEGKAPVFNYHRFWAQCDIAIANGLYSILF

Supplementary Table 3.1: Library blocks of the Cel48 SCHEMA library. The blocks are from the catalytic domains of the parental enzymes: CelS, CelF, and CelY.

Constant contribution (a_0)	Blocks	CelF-1 block contributions (a_{i2})	CelS block contributions (a_{i3})
4.2	A	0.00	-1.40
	B	1.20	-0.62
	C	-1.20	0.21
E value contribution (a_E)	D	0.56	0.00
	E	0.00	-0.96
-0.11	F	0.00	0.00
	G	-0.52	1.80
	H	0.00	-0.44

Supplementary Table 3.2: Coefficients of the functionality model, as determined by regularized logistic regression on data from all chimeras.

Correctly predicted active chimeras	50/53
Correctly predicted inactive chimeras	21/28
False positives	3/53
False negatives	7/28
Percentage correctly predicted	71/81 = 88%

Supplementary Table 3.3: The functionality model. Using block-sequence and the SCHEMA E value to predict which chimeras are functional. The data presented use 'leave one out' cross-validation.

Construct	Predicted functionality	Measured functionality	Predicted T_{50}(°C)	Measured T_{50}(°C)
32322322	1	1	80.1	73.5
22223312	1	1	79.0	73.3
32322332	1	1	80.7	72.7
22223322	1	1	77.6	76.0
22223332	1	1	78.9	77.5
22123332	1	1	78.0	77.4
22222322	1	1	80.3	78.1
22122322	1	1	79.4	77.8
22222332	1	1	80.2	78.2
22122332	1	1	79.3	78.9
32222322	1	0	78.3	-
22123312	1	0	78.1	-
22123322	1	0	76.8	-
32222222	1	0	75.1	-

Supplementary Table 3.4: Cellulosomal chimeras constructed, along with the predicted functionality and T_{50}s of the chimeras and measured functionality and T_{50}s. The constructs are identified using a set of 8 numbers that represent the parental identity of each block (A to H) in the chimera: 1 for CelF-1, 2 for CelY-2, and 3 for CelS. Thus 32132322 is CelS block A, CelY-2 block B, etc.

Primer name	Sequence (5' to 3')
CTHE312.40	ACGAAGCTTGAGCTCATGGGTCCTACAAAGGCACCTACAA
CTHE2453.40	ACGCTGCGGCCGCGTTCTTGTACGGCAATGTATCTATTTC
CCEL786.41	ACGAAGCTTGAGCTCATGGCTTCAAGTCCTGCAAACAAGGT
CCEL2864.41	ACGCTGCGGCCGCTTGGATAGAAAGAAGTGCTTTCTTTAAA
XbaIinCTHEfwd	GCTACATACTTCCCGGATTCTAGAATGACATATAAAGTACCT
XbaIinCTHErev	AGGTACTTTATATGTCATTCTAGAATCCGGGAAGTATGTAGC
pGEMTfwd1	ATTGGGCCCGACGTCGCATGCTCC
pGEMTrev1	CTCTCCCATATGGTCGACCTGCA
CCELcdCTHEdoc+XbaIfwd	GCAATACTCTTCCCAGATTCTAGAATGACATATAAAGTACCTGGTA CTCCTTCTACT
CCELcdCTHEdoc+XbaIrev	AGGTACTTTATATGTCATTCTAGAATCTGGGAAGAGTATTGCATAA ACTCCATTTGC
CCELcdCTHEdocfwd	GCAATACTCTTCCCAGATATGACATATAAAGTACCTGGTACTCCTT CTACT
CCELcdCTHEdocrev	AGGTACTTTATATGTCATATCTGGGAAGAGTATTGCATAAACTCCA TTTGC
CCELfwdNdeI	ATCACGCTCATATGGCTTCAAGTCCTGCAAACAAGGT
CCELfwd470mut	ACAAACCGGCTACATACGCACCGGAATTTCAGGACCC
CCELrev470mut	GGGTCCTGAAATTCCGGTGCGTATGTAGCCGGTTTGT
CCELfwd515mut	TCTCCGTTGGATACCAGTCAACCTGTTGGT
CCELrev515mut	ACCAACAGGTTGACTGGTATCCAACGGAGA
CCELfwd788u818mut	AAAGGATACAGGTACACCGGCAAAGCAATTCAAATATACAAATGCA CCAGATGCTGATGC
CCELrev788u818mut	GCATCAGCATCTGGTGCATTTGTATATTTGAATTGCTTTGCCGGTG TACCTGTATCCTTT
CTHEfwdNdeI	ATCACGCTCATATGGGTCCTACAAAGGCACCTACAA
CTHEfwd470mut	CAAACAGCCCTGCCACGTATGCTGACGAATATG
CTHErev470mut	CATATTCGTCAGCATACGTGGCAGGGCTGTTTG
CTHEfwd779u809mut	TACAAAGGACAGATCCTATGCAAAACAGTGGCGTTATACAAACGCA CCTGACGCAGAAGG
CTHErev779u809mut	CCTTCTGCGTCAGGTGCGTTTGTATAACGCCACTGTTTTGCATAGG ATCTGTCCTTTGTA
CTHEfwd1322mut	CGTTCTATGGTATGGCCTATGTTCCGCATCCTG
CTHErev1322mut	CAGGATGCGGAACATAGGCCATACCATAGAACG
CSTEfwdNde1	ATCACGCTCATATGGCCAGCAGCGATGATCCGTATAA
CSTEcdCTHEdoc+XbaIrev	AGGTACTTTATATGTCATTCTAGAGCTGCCAAACAGAATGCTATAC AAG
CSTEdufCTHEdoc+XbaIrev	AGGTACTTTATATGTCATTCTAGATGGCCTGCCAGAAATTCTTTCT TAAT
CSTEcdCTHEdoc+XbaIfwd	TAGCATTCTGTTTGGCAGCTCTAGAATGACATATAAAGTACCTGGT ACTCCTTCTACT
CSTEdufCTHEdoc+XbaIfwd	AATTTCTGGCAGGCCAGTCTAGAATGACATATAAAGTACCTGGTAC TCCTTCTACT
CSTEwt-rev-nostop	TCGAGTGCGGCCGCCGGCTCGATACCATAA

Supplementary Table 3.5: Primer names and primer sequences used for parent constructs.

Primer name	Sequence (5' to 3')
CCEL0fwd	CCGACTAGTGACTCATATGGCTTCAAGTCCTGCAAAC
CTHE0fwd	CCGACTAGTGACTCATATGGGTCCTACAAAGGCACC
CSTE0fwd	CCGACTAGTGACTCATATGGCCAGCAGCGATGATCCG
CCEL1fwd	TGGCAGAACGGACTCTCCGCTAGCCCCGACAGAAAAGGATCAGCCCAATA
CCEL1rev	GGAGAGTCCGTTCTGCCAGACGGGATCAAATACTGTTCGGTAACAG
CTHE1fwd	TGGCAGAACGGACTCTCCGCTAGCCCCGGACAGCACAGAGCAGCCGG
CTHE1rev	GGAGAGTCCGTTCTGCCAGACGGAATTATCCAATCCTCCATAACTTTC
CSTE1fwd	TGGCAGAACGGACTCTCCGCTAGCCCCGACCGAGCAGGATCAACCGAA
CSTE1rev	GGAGAGTCCGTTCTGCCAGACGGAATCATGTACTTCTCAATGGTTTC
CCEL2fwd	CAACCGTACCGGTACTCCGCTAGCGGCCGTTCAAGCAACTTACTGGGC
CCEL2rev	GGAGTACCGGTACGGTTGCCGGCACGAGCATCAGCATCTGGTG
CTHE2fwd	CAACCGTACCGGTACTCCGCTAGCGGCCATACAGGCTGTTTACTGGGC
CTHE2rev	GGAGTACCGGTACGGTTGCCGGCACGGCCTTCTGCGTCAGGTG
CSTE2fwd	CAACCGTACCGGTACTCCGCTAGCGGCCATCCAGGCGACGTACTGGGC
CSTE2rev	GGAGTACCGGTACGGTTGCCGGCACGCGCATCCGCGTCCGGTG
CCEL3fwd	GCACGATATACCACGTCTCCGCTAGCAGACTACCTTAGATATTCATTCTTTGATAAG
CCEL3rev	GGAGACGTGGTATATCGTGCGTCACCCATCTTTGTTGCCTTACC
CTHE3fwd	GCACGATATACCACGTCTCCGCTAGCAGACTTCTTGAGAAACGACATGTTCG
CTHE3rev	GGAGACGTGGTATATCGTGCGTCACCCATCTTTGCAGCCTTGGA
CSTE3fwd	GCACGATATACCACGTCTCCGCTAGCAGACTACCTGCGCTATGCAATGTT
CSTE3rev	GGAGACGTGGTATATCGTGCGTCGCCCATGCGGCTTGCCTTCT
CCEL4fwd	CCGCACTAGTGCTCTTCTCATTTCGGTTACCAGAACCCATTT
CCEL4rev	TTTTCCGCGGGCTCTTCTATGATTATGACTGCTACCGATTATCC
CTHE4fwd	CCGCACTAGTGCTCTTCTCATTTCGGATATCAGAACCCATTCC
CTHE4rev	TTTTCCGCGGGCTCTTCTATGTGCGTGGCTGCATCCGATCT
CSTE4fwd	CCGCACTAGTGCTCTTCTCATGCAGGCTACCAGAATCCGAT
CSTE4rev	TTTTCCGCGGGCTCTTCTATGAACGTGGGAACAGCCAATAAT
CCEL5fwd	TGGCAGAACGGACTCTCCGCTAGCCGGTGCTATTGCCGGTGGAGCTAC
CCEL5rev	GGAGAGTCCGTTCTGCCAGAACCTTCTGCTGACTGCAACCACT
CTHE5fwd	TGGCAGAACGGACTCTCCGCTAGCCGGTGGTATTGCCGGTGGAGCAAC
CTHE5rev	GGAGAGTCCGTTCTGCCAGAACCTTCAGCCGACTGCAACCACT
CSTE5fwd	TGGCAGAACGGACTCTCCGCTAGCCGGTGCAATCGCAGGTGGTGCGAC
CSTE5rev	GGAGAGTCCGTTCTGCCAGAACCCTCAGCGCTCTGCAGCCATT
CCEL6fwd	CAACCGTACGCCATCTCCGCTAGCGAACACTTGGTTTGGTATGCAGGTAT
CCEL6rev	GGAGATGGCGTACGGTTGCCGTTACTACCTGGGTCAGCATATACA
CTHE6fwd	CAACCGTACGCCATCTCCGCTAGCGAACCAGTGGTTCGGATTCCAGGC
CTHE6rev	GGAGATGGCGTACGGTTGCCGTTACTACCCGGGTCAGCGTATA
CSTE6fwd	CAACCGTACGCCATCTCCGCTAGCGAACACGTGGTTTGGCTTTCAGGC
CSTE6rev	GGAGATGGCGTACGGTTGCCGTTGCTACCCGGATCCAGGTACA
CCEL7fwd	GCACGATATACCACGACTCCGCTAGCATTACTTGACGCTATGTGGAATAACT
CCEL7rev	GGAGTCGTGGTATATCGTGCTAATTTCTGTGCATTCTGCCTTGA
CTHE7fwd	GCACGATATACCACGACTCCGCTAGCATTAGTTAACCGTGCATGGTACAAC
CTHE7rev	GGAGTCGTGGTATATCGTGCTAATTCAGCAGCCATGTCTCTTG
CSTE7fwd	GCACGATATACCACGACTCCGCTAGCATTACTGGACCGTATGTGGAACCT
CSTE7rev	GGAGTCGTGGTATATCGTGCTAATTCTTTCGCCAGATTACGTG
CTHEdocrev	TTTTCCGCGGTTTGCGGCCGCGTTCTTGTACGGC
CSTE4rev	TTTTCCGCGGGCTCTTCTATGAA
CSTE4fwd	CCGCACTAGTGCTCTTCTCATG

Supplementary Table 3.6: List of primers used for construction of SCHEMA library.

Block	A	B	C	D	E	F	G	H
CelF	26	7	29	24	1	14	9	11
CelY	11	6	7	5	25	27	34	31
CelS	17	41	18	25	28	13	11	12

Supplementary Table 3.7: The quality of the built library: distribution of blocks for 54 randomly picked, and correctly assembled, chimeras. Every parent is present in every position. CelF block E appears to be underrepresented in the library.

$$T_{50} = a_0 + \Sigma_{i=1}^{8}\Sigma_{j=2}^{3} a_{ij}x_{ij} + a_E E$$

Supplementary Equation 3.1: The thermostability model. a_0 is a constant (in this case the T_{50} of CelY-2), a_{ij} is the contribution to activity of block i from parent j, x_{ij} is a dummy variable representing that block i either comes from CelF-1 $x_{i2} = 1$ or from CelS $x_{i3} = 1$, a_E is the contribution to stability of the E value per contact broken, E is the E value (the number of contacts disrupted in a chimera).

$$P(active) = f_{link}(a_0 + \Sigma_{i=1}^{8}\Sigma_{j=2}^{3} a_{ij}x_{ij} + a_E E)$$

Supplementary Equation 3.2: The functionality model. P(active) is the probability a chimera is active, a_0 is a constant, a_{ij} is the contribution to activity of block i from parent j, x_{ij} is a dummy variable representing that block i either comes from CelF-1 $x_{i2} = 1$ or from CelS $x_{i3} = 1$, a_E is the contribution to activity of the E value per contact broken, E is the E value (the number of contacts broken in a chimera), $f_{link}()$ is the logistic linking function that scales the output to a value between 0 and 1.

Chapter 4

Chimeragenesis of distantly-related proteins by non-contiguous recombination

A modified version of this chapter appears in: Smith M. A., Romero P. A., Wu T., Brustad E. M., and Arnold F. H. (2013) Chimeragenesis of distantly-related proteins by non-contiguous recombination, *Protein Sci.* **22**, 231-238, and is reprinted with permission from Wiley-VCH.

4.1 Abstract

We introduce a method for identifying elements of a protein structure that can be shuffled to make chimeric proteins from two or more homologous parents. Formulating recombination as a graph partitioning problem allows us to identify non-contiguous segments of the sequence that should be inherited together in the progeny proteins. We demonstrate this non-contiguous recombination approach by constructing a chimera of β-glucosidases from two different kingdoms of life. Although the proteins alpha-beta barrel fold has no obvious sub-domains for recombination, non-contiguous SCHEMA recombination generated a functional chimera that takes approximately half its structure from each parent. The x-ray crystal structure shows that the structural blocks that make up the chimera maintain the backbone conformations found in their respective parental structures. Although the chimera has lower β-glucosidase activity than the parent enzymes, the activity was easily recovered by directed evolution. This simple method, which does not rely on detailed atomic models, can be used to design chimeras that take structural, and functional, elements from distantly-related proteins.

4.2 Introduction

Swapping sequence elements among related proteins [1] can produce chimeric proteins with novel behaviors [2,3] and improved properties such as enhanced stability [4]. Although homologous mutations are much more conservative than random mutations, chimeras of distantly-related proteins have a low probability of retaining fold and function [5]. Selecting crossover locations that minimize disruption of the folded structure increases the likelihood that a chimeric protein will be functional.

To design libraries of chimeric proteins we have used structural information to select crossover locations that minimize the average number of non-native residue-residue contacts in the resulting chimeras [6]. The sequence elements are then shuffled and reassembled in the correct order to generate the chimeric progeny. We have used this SCHEMA recombination method to make large numbers of functional enzyme chimeras, with which we have explored the benefits and costs of recombination [3, 7-9]. We have also shown that stabilities and other properties of these recombined enzymes - the 'recombination landscape' - can be predicted with high accuracy using models built by sampling small numbers of chimeras [4, 10].

To date, we have only considered recombination of sequence blocks that are contiguous along the polypeptide chain. Sequence blocks that are contiguous in the primary structure, however, are not necessarily optimal elements for recombination [11]. Here, we introduce a new tool for protein recombination that identifies structural blocks that can be swapped among homologous proteins with minimal disruption. Because elements that are distant in the primary structure are often brought together in the folded protein, structural blocks may not be contiguous in the polypeptide chain. This non-contiguous recombination approach enables design of chimeras and libraries of chimeras with less disruption than can be achieved by swapping blocks of sequence. Less disruption means that we can generate libraries with higher fractions of functional enzymes and enables recombination of more distant homologs.

We demonstrate this new tool by constructing a functional β-glucosidase that derives approximately half of its sequence from each of two distantly-related parents. The crystal structure of this prokaryote-eukaryote chimera illustrates the structurally conservative nature of this recombination: the hybrid structure retains the overall function as well as the detailed structural features of the parental enzymes.

4.3 Results

4.3.1 Non-contiguous protein recombination

The goal is to identify blocks that can be shuffled among related parent proteins to create chimeras with minimal disruption. The overall process is illustrated in Figure 1 for the simple case of 2 parents, but can be extended easily to any number of parents. Starting from one or more structures and a parental sequence alignment (Figure 4.1a), non-contiguous recombination involves splitting the proteins into a set of blocks (Figure 4.1b) which are swapped to create chimeras (Figure 4.1c). Similar to previous work with recombination of contiguous sequence elements, our disruption metric is the number of non-native residue-residue contacts that are broken in the recombined sequence; we call this the SCHEMA disruption [6]. To minimize disruption, the residue-residue contacts that are not shared among the parents and therefore could be broken upon recombination are converted into a graph, with residues as nodes and non-native contacts as edges (Figure 4.1d). Assigning residues to blocks is then equivalent to partitioning the graph to minimize the number of edges that are cut (Figure 4.1e). This is an NP-complete problem [12], but there are heuristic algorithms that can find near optimal solutions very quickly [13]. We use hmetis [14, 15], a suite of graph partitioning tools. The hmetis suite assigns each node to a partition, which corresponds to assigning each residue to a block. The non-contiguous chimeras are then assembled from the shuffled blocks, where a block can comprise multiple sequence fragments that should be inherited together (Figure 4.1f).

4.3.2 Chimeric β-glucosidase design

We chose to test this non-contiguous SCHEMA recombination approach by making a chimera of two distantly-related GH1 β-glucosidases, one from a prokaryote, the thermophilic *T. maritima* BglA [16, 17] (TmBglA), and the other from a eukaryote, the mesophilic *T. reesei* Bgl2 [18, 19] (TrBgl2). These enzymes share 41% sequence identity, with a conserved active site. The TIM-barrel enzyme fold has no obviously interchangeable subdomains.

We generated various 2-block chimera designs that are predicted to have low disruption and picked the one shown in Figure 4.2 for construction and characterization. Chimera NcrBgl would have approximately half its sequence from TmBglA and half from TrBgl2; it would have 144 mutations, corresponding to ∼31% of its sequence, from the closest parent (TmBglA). Figure 4.2a shows NcrBgl on the sequence alignment of TmBglA and TrBgl2. The non-contiguous nature of the two blocks on the polypeptide chain is readily apparent - the red TrBgl2 block has 7 separate sequence fragments, and the green TmBglA block has 8. These blocks are contiguous, however, on the 3-dimensional structure, as shown in Figure 4.2b.

We predicted that this choice of crossovers should be minimally disruptive. The number of residue-residue contacts in NcrBgl that are not found in any of the parent contact maps is only 27.5, an average of 25 broken contacts based on TmBglA's structure 2WBG.pdb and 30 based on TrBgl2's structure 3AHY.pdb. By comparison, swapping half the proteins structure randomly breaks on average 155 contacts (Figure 4.3a), and the best design of 10,000 random designs breaks more than 70 contacts (see Materials and Methods). Designs with many broken contacts are unlikely to lead to properly folded, functional enzymes [7]. Figure 4.3b shows the optimized non-contiguous chimera design on a plot of the residue-

residue contacts that could be broken (SCHEMA contacts). Most SCHEMA contacts are sequestered within a block in this design, and thus few contacts are disrupted upon recombination.

4.3.3 Structural conservation

The gene encoding the eukaryotic-prokaryotic NcrBgl chimera was synthesized and expressed under the control of an arabinose-inducible promoter in Top10 *E. coli* cells. TrBgl2 and TmBglA break down cellobiose and other short oligosaccharides into glucose. Both parent enzymes are active over a range of pH, from 4 to 7, and TrBgl2 is active between 30°C and 55°C [19], while TmBgl2 is highly thermostable with significant activity between 60°C and 100°C [16]. NcrBgl is catalytically active over the temperature range 30°C to 60°C and is approximately a factor of 10^3 less active than TrBgl2 at 37°C. The activity is easily recovered, however, to TrBgl2 levels, by directed evolution (see below). We also synthesized the gene for the 'mirror' chimera (with the parental identities of each block swapped), but it was not expressed as a functional protein in *E. coli*.

For structure determination, the NcrBgl chimera was expressed in *E. coli* BL21 DE3 with an N-terminal his6 tag and purified from cell lysate on a Ni-NTA column followed by an anion exchange column. Crystals were grown using the vapor-diffusion method, and NcrBgl's structure was solved from x-ray diffraction data using MOLREP [20] and REFMAC5 [21] (see Materials and Methods).

The crystal structure of NcrBgl (4GXP.pdb), determined at 3.0 Å, shows that both blocks retain the structures of their respective parents. Chimera NcrBgl has the TIM-barrel fold and catalytic residues E170 and E374 (numbering based on the alignment shown in Figure 4.2a) of the parent enzymes. Figure 4.4a illustrates the blocks on the parent struc-

tures and the structure of the chimera. The structural independence of recombined blocks is pronounced: there are significant differences between the aligned structures of the parents (Figure 4.4b), particularly on the surface where there are multiple insertions and deletions in loop regions (Figure 4.4c). These structurally disparate regions are apparently unaffected by the chimeragenesis and maintain their backbone conformations when reassembled in the chimera.

We tested whether we could model the structure of the chimera by combining the parental structures of the chimeras blocks, using an alignment of the parental structures to position each block. Thus, for NcrBgl we combined the structures of the TrBgl2 block and the TmBglA block to predict the structure of NcrBgl. This model does a good job at capturing variations in the backbone and loops (Figure 4.4d). Our ability to predict finer structural features is limited by the current low resolution of the chimera structure.

4.3.4 Recovering activity with directed evolution

We performed five rounds of random mutagenesis and screening for higher activity on the fluorescent β-glucosidase substrate, 4-nitrophenyl β-D-glucopyranosidase (pNPG) (see Materials and Methods). Figure 4.5 shows the activity of the best mutant from each round, relative to NcrBgl. Activity increased almost 1000-fold in just five rounds. The resulting mutant has 149 mutations from the closest known natural sequence (TmBglA) and activity comparable to TrBgl2.

4.4 Discussion

Structure-guided recombination is a powerful tool for generating novel enzymes with diverse sequences. We have presented a new method that splits proteins into elements of sequence

that should be inherited together in order to minimize structural disruption. The resulting blocks can be non-contiguous along the polypeptide chain. We have developed tools to efficiently design chimeras and chimera libraries. These non-contiguous block designs disrupt far fewer SCHEMA contacts than equivalent designs that require contiguous sequence blocks. Indeed, contiguous block designs are a (suboptimal) subset of the non-contiguous block design space.

This approach does not rely on detailed atomistic models of the parent and progeny proteins. Indeed the only structural information used is a set of residue-residue contacts, which, with the parent sequences, is sufficient to design functional chimeras of distantly-related proteins that do not have obvious subdomains. Simply minimizing the number of broken parental contacts seems to be sufficient to generate functional chimeras with a good success rate, as has been shown for contiguous SCHEMA recombination [7].

To test the method, we designed and constructed a chimeric β-glucosidase that takes large blocks from a prokaryotic parent and a eukaryotic parent. While we designed a 2-block, 2-parent chimera for this example, the graph partitioning method can easily produce non-contiguous designs for libraries of chimeric proteins having multiple parents and multiple blocks.

On solving the crystal structure of the chimeric enzyme, we discovered that each block retains the structure of its corresponding parent (within the limits of the 3.0 $\mathring{A}$ resolution), suggesting that it may be possible to predict the structures of chimeric enzymes from the parent enzymes by simply combining the known parent structures. Alternatively, structures of the chimeric proteins could provide detailed and accurate information on the structures of the parent proteins. This can be very useful for eukaryotic protein structure determination, for example, where chimeragenesis enables production in a microbial recombinant host [22,

23]. The fact that the recombined blocks retain their parental structure could also be very useful for creating protein chimeras that acquire the functions (e.g. allosteric regulation, interactions with other proteins, or substrate specificity) of their parent blocks.

That the chimera is somewhat compromised in β-glucosidase activity compared to its parents is not surprising, considering the simplicity of the design approach and also that 144 mutations were introduced. However, the chimera was easily fine-tuned for native-like activity levels in just five rounds of random mutagenesis and screening. This example offers promise for exploring distant parts of sequence space, perhaps never explored by nature, for novel enzymes.

4.5 Materials and methods

4.5.1 Non-contiguous recombination

A structure-based sequence alignment of the parental enzymes T. maritima BglA [16, 17] (TmBglA) and T. reesei Bgl2 [18, 19] (TrBgl2) was created using PROMALS3D [24]. For a given structure, two residues are in contact if any atoms from each residue were within $4.5\mathring{A}$ of each other, excluding hydrogen atoms. A SCHEMA contact map contains those contacts that are not conserved among the parental enzymes. Since the TmBglA and TrBgl2 structures vary considerably, a SCHEMA contact map was built for each parent, and a final average SCHEMA contact map weighted each contact depending on the number of parents in which it was present (0.5 if in a single parent, 1 if in both parents). PDB structures 2WBG.pdb chain A and 3AHY.pdb chain A were used to create the TmBglA and TrBgl2 SCHEMA contact maps, respectively.

The SCHEMA contact map was abstracted as a graph. Each non-conserved residue rep-

resented a node, and each edge represented an average weighted SCHEMA contact between two residues. Finding crossover locations that minimize the average number of SCHEMA contacts in the chimeras was reformulated as a problem of minimizing the cut edges when partitioning a graph. The hmetis graph partitioning suite [14, 15] was used to find 2-way partitions of the SCHEMA contact map - these partitions gave designs for 2-block chimeragenesis of TmBglA and TrBgl2. A design was selected that would produce a chimera with a SCHEMA energy (number of disrupted contacts) of 27.5 and 144 mutations from the closest parent.

4.5.2 Random chimeragenesis designs

This analysis was carried out with PDB structure 2WBG chain A. The structure was partitioned into two blocks by a randomly-generated cut plane through the protein's center. Each residue was assigned to one of the two blocks based on the coordinates of its alpha carbon. Swapping the residues of the blocks among the parents TmBglA and TrBgl2 created two possible chimeras with equal SCHEMA energies. The chimera SCHEMA energies were calculated using the SCHEMA contact map from 2WBG chain A.

4.5.3 Gene synthesis

The NcrBgl gene (Supplementary Figure 4.6) was optimized for expression in *E. coli* and synthesized by DNA2.0, Menlo Park, CA, USA.

4.5.4 Protein preparation and crystallization

A 1L baffled flask of Luria broth (LB) with 100 mg/L ampicillin was inoculated with 5 mL of an overnight culture of *E. coli* BL21 DE3 cells containing the NcrBgl gene with an

N-terminal his6 tag on a pET-22(+) vector. The flask was grown for 4 hours at 37°C, 250 rpm before being induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 10 μM and incubated for 16 hours at 16°C and 250 rpm. The cells were pelleted by centrifugation at 5000 g and frozen at -20°C. The cells were resuspended in 10 mM Tris, pH 7.4 and lysed by sonication. The lysate was spun at 60,000 g for 20 minutes and the supernatant filtered with a Nalgene 0.2 μm aPES filter. The supernatant was loaded onto a 5 mL Ni-NTA His-trap HP column (GE Healthcare) and purified by washing with 1% elution buffer (20 mM Tris, pH 7.4, 100 mM NaCl, 300 mM imidazole) for 15 column volumes (CV), followed by a gradient elution (increase to 80% elution buffer in 10 CV). Fractions containing the NcrBgl protein were buffer exchanged to 20 mM Tris, pH 7.4 and loaded onto a 5 mL HiTrap Q HP column (GE healthcare). The column was washed with 1% elution buffer (20 mM Tris, 1 M NaCl, pH 7.4) for 15 column volumes (CV) and the protein purified by a gradient elution (increase to 80% elution buffer in 10 CV). Fractions containing the NcrBgl protein were pooled and concentrated using 30,000 molecular weight cut-off protein concentrators with cellulose-free membranes (Vivaspin). Buffer was exchanged to 10 mM Tris, pH 8.0 by repeated refills and the protein flash frozen and stored at -20°C. The protein was crystallized by vapor diffusion of a 4:3 mixture of 20 g/L protein in 10 mM Tris, pH 8.0 and 20% polyethylene glycol 3350, 0.4 M sodium malonate, pH 7.0 in 24-well sitting drop plates (Hampton Research). Crystal growth occurred over a period of 2-3 days and larger, higher-resolution crystals were obtained by microseeding with pieces of sonicated crystals. Crystals were frozen in 25% glycerol for structure determination.

4.5.5 Structure determination and refinement

X-ray diffraction data were collected on a Dectris Pilatus 6M detector at 100K at the Stanford Synchrotron Radiation Lightsource, beamline 12-2. The wavelength of the beam was 0.9795 $\AA$. Diffraction data were integrated using XDS [25] and scaled using SCALA [26]. A homology model of the NcrBgl was constructed in MODELLER [27] using 2WBG.pdb, chain A and 3AHY.pdb, chain C. This model was used by MOLREP [20], a molecular replacement tool that is part of the CCP4 crystallography software [28], to determine the initial phases of the X-ray data. The structure was refined with several rounds of manual model building within Coot [29] and automated refinement using REFMAC5 [21] within CCP4. Data refinement and collection statistics are given in Supplementary Table 4.1.

4.5.6 Error-prone library construction

For expression in *E. coli* TOP10 cells, the NcrBgl gene and N-terminal his6 tag was subcloned into the arabinose-inducible pBAD vector using Gibson assembly [30]. A library of mutants with 3.4 nucleotide mutations per gene was generated by error-prone PCR using 50 μM $MnCl_2$ and Applied Biosystems AmpliTaq polymerase. The pBAD backbone was amplified by regular PCR. Both PCR products were digested for 30 minutes by Dpn1 (New England Biolabs), purified on an agarose gel and ligated together using Gibson assembly. The library was transformed into electrocompetent *E. coli* TOP10 cells and plated on LB-agar media with 100 mg/L ampicillin.

4.5.7 Library expression in 96-well plates

Individual mutant colonies from the library plates were picked into 96-well plates containing 300 μL LB with 100 mg/L ampicillin and grown at 37°C, 250 rpm, and 80% humidity.

Each plate contained four null-control wells with an empty pBAD plasmid, four wells with the NcrBgl gene and four wells with the parent gene from the previous round of directed evolution. After 16 hours, 50 μL of each culture was expanded into 96-well plates containing 900 μL LB with 100 mg/L and grown at 37°C for a further 4 hours. The plates were then induced with 50 μL of 0.8% arabinose to give a final concentration of 0.04% arabinose. The plates were incubated for 16 hours at 16°C and 250 rpm and the cells pelleted by centrifugation at 4000 g and frozen at -20°C.

4.5.8 Enzyme activity screen

The cell pellets were lysed by adding 300 μL of 10 mM HEPES pH 8.0, 10 mM $MgCl_2$, 0.7 mg/L lysozyme and 0.1 units of DNAase I (Sigma) to each well and incubating at 37°C for 1 hour. 50 μL of lysate was transferred to a PCR plate containing 150 μL of 10 mM 4-nitrophenyl β-D-glucopyranosidase (pNPG) and incubated at 37°C for 1 hour. The reaction was stopped by adding 20 μL of 1M sodium hydroxide and absorbance was read at 410 nm. Twenty plates were screened in each round. The best mutants were streaked onto an LB plate with 100 mg/L ampicillin and individual colonies used to rescreen in quadruplicate.

4.6 Figures

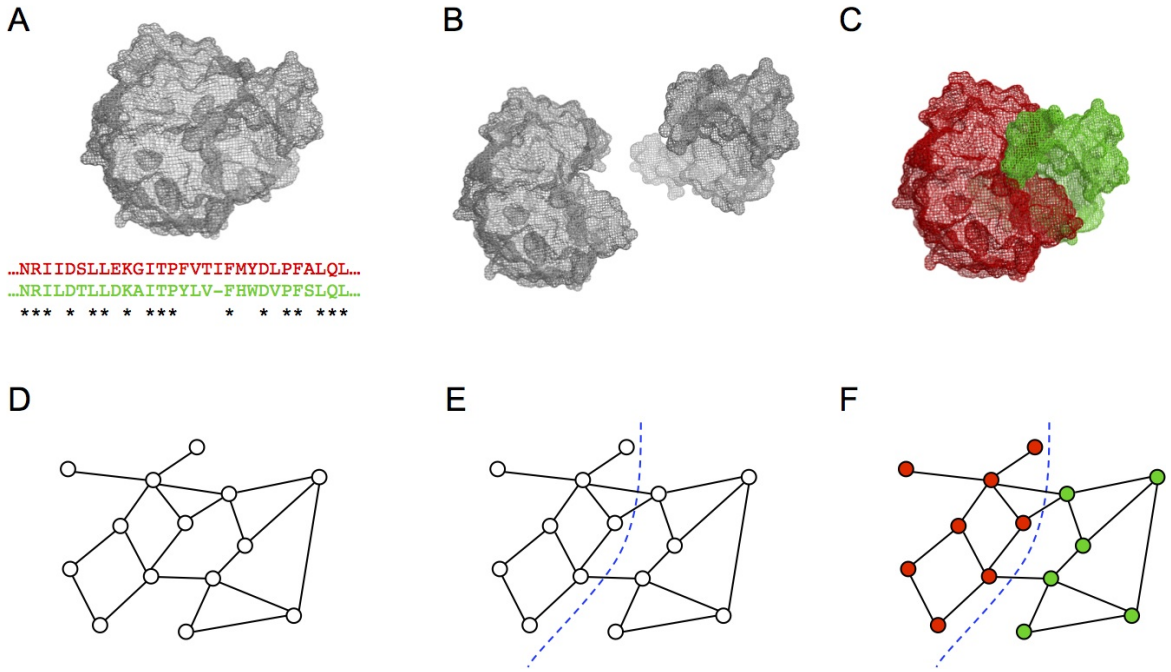


Figure 4.1: Non-contiguous recombination. A) One or more structures and a parental sequence alignment are used to identify contacts that are not conserved and can be disrupted upon recombination (SCHEMA contacts). B) Sequence elements that should be inherited together (blocks) are identified based using the SCHEMA contact map. Optimal blocks are often non-contiguous along the polypeptide chain but are contiguous on the 3D structure. C) The chimeras are reassembled using blocks from different parents. D) The SCHEMA contact map can be reformulated as a graph, where nodes represent residues and edges represent SCHEMA contacts. E) To design non-contiguous recombination chimera libraries, the graph is partitioned, with each residue assigned to a block. Partitions are chosen to minimize the edges between blocks. F) Graph schematic of a chimeric protein.

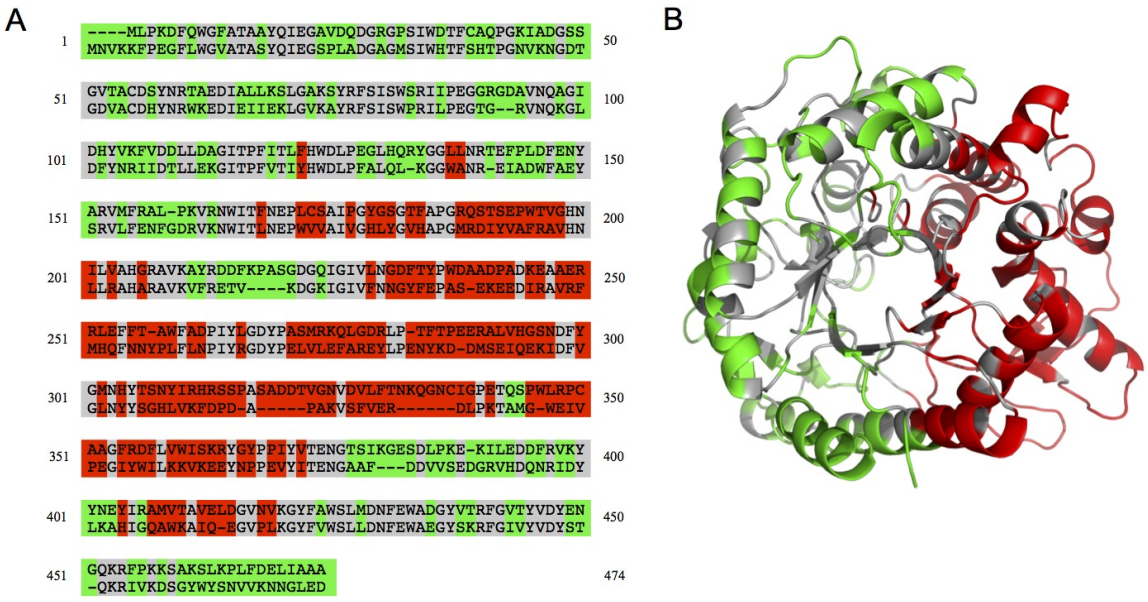


Figure 4.2: β-glucosidase non-contiguous chimera design chosen for construction. A) Numbered sequence alignment of the eukaryotic (top) and prokaryotic (bottom) β-glucosidases. Conserved residues are in gray, the block of eukaryotic mutations are in red, and the block of prokaryotic mutations are in green. B) The 2-block design illustrated on the structure of the prokaryotic enzyme, TmBglA (2WBG.pdb).

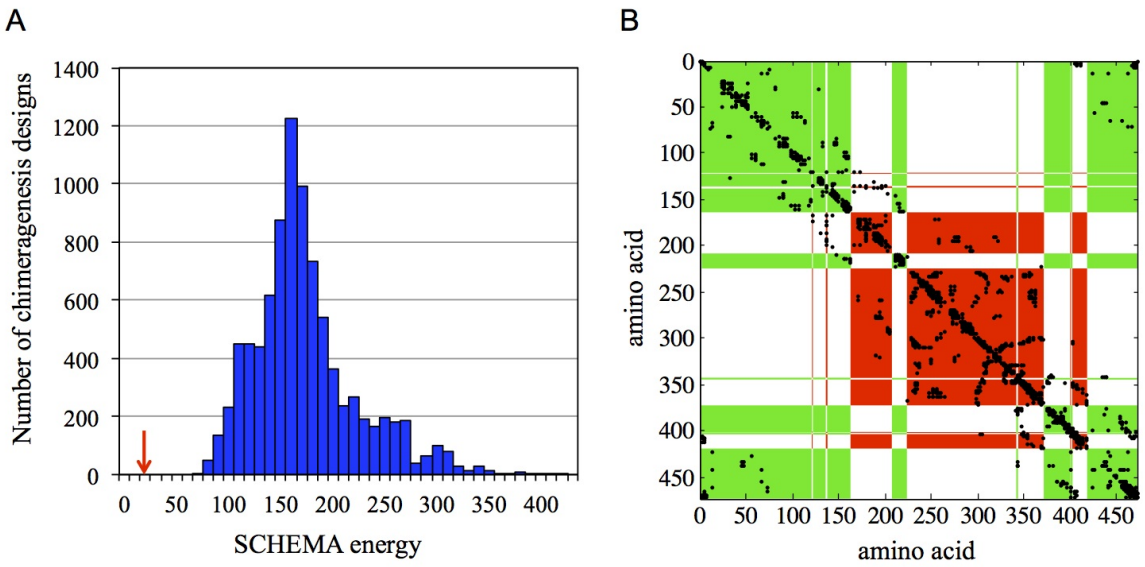


Figure 4.3: The optimal non-contiguous design breaks far fewer contacts than random 2-block partitions of the structure. A) A histogram of the SCHEMA energies of 10,000 random 2-block chimeragenesis designs. The SCHEMA energy of the optimized non-contiguous design is highlighted with a red arrow. B) The SCHEMA contact map for the optimized non-contiguous 2-block design. Most of the SCHEMA contacts are within the two blocks and thus are not disrupted upon recombination. The numbering is based on the parent alignment, and SCHEMA contacts are shown in black. Red and green areas show the two blocks. (For greater clarity, the conserved residues have been assigned to one of the two blocks based on structural proximity.)

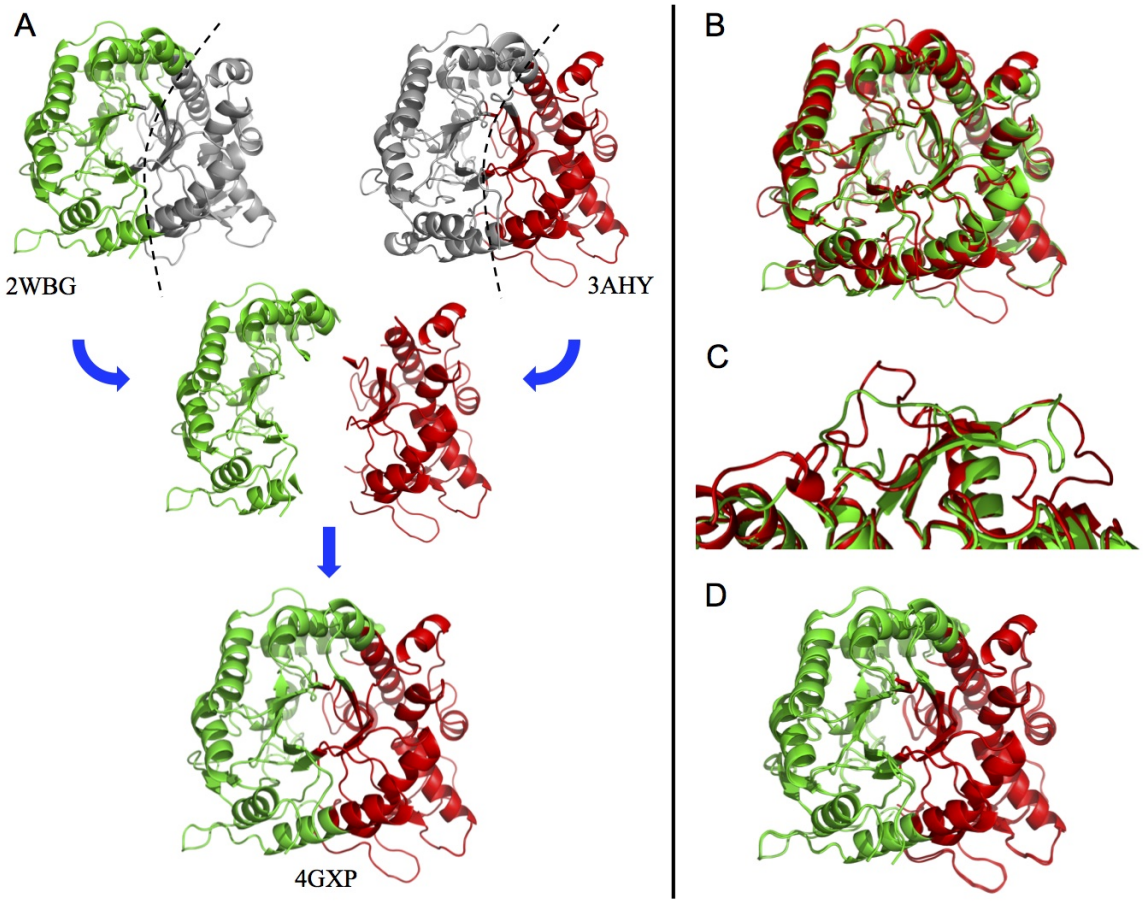


Figure 4.4: Structural elements are conserved upon recombination. A) The structure of chimera NcrBgl (4GXP.pdb), bottom, is nearly identical to the assembled structure of its component blocks from TrBgl2 (3AHY.pdb) and TmBglA (2WBG.pdb), top. The eukaryotic TrBgl2 residues and the prokaryotic TmBglA residues are highlighted in red and green, respectively. (For greater clarity, the conserved residues have been assigned to one of the two blocks based on structural proximity.) B) A structural alignment of TmBglA 2WBG.pdb and TrBgl2 3AHY.pdb (RMSD = 3.34 $\mathring{A}$) shows significant variation between these two homologs. C) An example of significant variations in loop regions. D) Model of NcrBgl constructed simply by stitching together the parental blocks closely aligns with NcrBgl's actual structure (RMSD = 1.15 $\mathring{A}$).

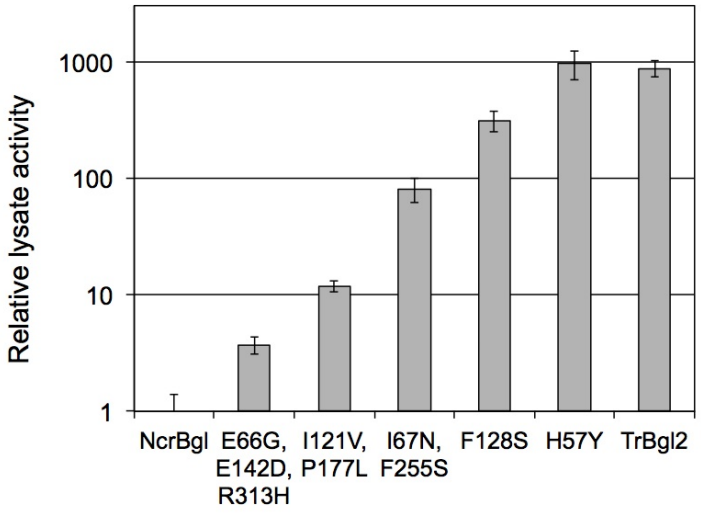


Figure 4.5: Directed evolution recovers the activity of NcrBgl to wild-type levels. Activity is measured in lysate with a 1-hour assay on pNPG at 37°C and normalized relative to NcrBgl. The new mutations found at each round are listed (numbering based on the parental alignment). Five rounds of directed evolution increased the activity of NcrBgl almost 1000-fold.

4.7 References

1. Crameri A., Raillard S. A., Bermudez E., and Stemmer W. P. C. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288-291.

2. Dueber J. E., Yeh B. J., Chak K., and Lim W. A. (2003) Reprogramming control of an allosteric signaling switch through modular recombination. *Science* **301**, 1904-1908.

3. Otey C. R., Landwehr M., Endelman J. B., Hiraga K., Bloom J. D., and Arnold F. H. (2006) Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol.* **4**, e112.

4. Li Y., 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. *Nat. Biotechnol.* **25**, 1051-1056.

5. Romero P. A. and Arnold F. H. (2012) Random field model reveals structure of the protein recombinational landscape. *PLoS Comput. Biol.* **8**, e1002713.

6. Voigt C. A., Martinez C., Wang Z.-G., Mayo S.L., and Arnold F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553-558.

7. Meyer M., Hochrein L., and Arnold F. H. (2006) Structure-guided SCHEMA recombination of distantly related β-lactamases. *Protein Eng. Des. Sel.* **19**, 563-570.

8. Heinzelman P., Snow C. D., Wu I., Nguyen C., Villalobos A., Govindarajan S., Minshull J., and Arnold F. H. (2009) A family of thermostable fungal cellulases created by structure-guided recombination. *Proc. Natl. Acad. Sci. USA* **106**, 5610-5615.

9. Romero P., Stone E., Lamb C., Chantranupong L., Krause A., Miklos A., Hughes R., Fechtel B., Ellington A. D., Arnold F. H., and Georgiou G. (2012) SCHEMA-designed variants of human arginase I and II reveal sequence elements important to stability and catalysis. *ACS Synth. Biol.* **1**, 221-228.

10. Romero P. A., Krause A., and Arnold F. H. (2012) Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. USA* **110**, e193-e201.

11. Pantazes R. J., Saraf M. C., and Maranas C. D. (2007) Optimal protein library design using recombination or point mutations based on sequence-based scoring functions. *Protein Eng. Des. Sel.* **20**, 361-373.

12. Garey M. R., Johnson D. S., and Stockmeyer L. (1976) Some simplified NP-complete graph problems. *Theor. Comput. Sci.* **1**, 237-267.

13. Kernighan B. W. and Lin S. (1970) An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.* **49**, 291-307.

14. Karypis G., Aggarwal R., Kumar V., and Shekhar S. (1997) Multilevel hypergraph partitioning: application in VLSI domain. In: *Proceedings of the 34th annual Design Automation Conference*, ACM Press, New York, USA.

15. Karypis G, Kumar V (2000) Multilevel k-way hypergraph partitioning. *VLSI Des.* **11**, 285-300.

16. Gabelsberger J., Liebl W., and Schleifer K.-H. (1993) Purification and properties of recombinant β-glucosidase of the hyperthermophilic bacterium Thermotoga maritima. *Appl. Microbiol. Biotechnol.* **40**, 44-52.

17. Zechel D., Boraston A., Gloster T., Boraston C. M., Macdonald J. M., Tilbrook D. M. G., Stick R. V., and Davies G. J. (2003) Iminosugar glycosidase inhibitors: structural and thermodynamic dissection of the binding of isofagomine and 1-deoxynojirimycin to β-glucosidases. *J. Am. Chem. Soc.* **125**, 14313-14323.

18. Takashima S., Nakamura A., Hidaka M., Masaki H., and Uozumi T. (1999) Molecular cloning and expression of the novel fungal β-glucosidase genes from Humicola grisea and Trichoderma reesei. *J. Biochem.* **125**, 728-736.

19. Jeng W. Y., Wang N. C., Lin M. H., Lin C. T., Liaw Y. C., Chang W. J., Liu C. I., Liang P. H., Wang A. H. J. (2011) Structural and functional analysis of three β-glucosidases from bacterium Clostridium cellulovorans, fungus Trichoderma reesei and termite Neotermes koshunensis. *J. Struct. Biol.* **173**, 46-56.

20. Vagin A. and Teplyakov A. (1997) MOLREP: an automated program for molecular replacement. *J. Appl. Crystallogr.* **30**, 1022-1025.

21. Murshudov G. N., Vagin A. A., Dodson E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240-255.

22. Duret G., Van Renterghem C., Weng Y., Prevost M., Moraga-Cid G., Huon C., Sonner J. M., and Corringer P.-J. (2011) Functional prokaryotic-eukaryotic chimera from the pentameric ligand-gated ion channel family. *Proc. Natl. Acad. Sci. USA* **108**, 12143-12148.

23. Shimoji M., Yin H., Higgins L., and Jones J. P. (1998) Design of a novel P450: a functional bacterial-human cytochrome P450 chimera. *Biochemistry* **37**, 8848-8852.

24. Pei J., Kim B.-H., and Grishin N. V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295-2300.

25. Kabsch W. (2010) XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125-132.

26. Evans P. (2005) Scaling and assessment of data quality. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 72-82.

27. Eswar N., Webb B., Marti-Renom M. A., Madhusudhan M., Eramian D., Shen M. Y., Pieper U., and Sali A. (2007) Comparative protein structure modeling using Modeller. *Curr. Protoc. Protein Sci.* **2**, 15-32.

28. Bailey S. (1994) The CCP4 suite programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 760-763.

29. Emsley P. and Cowtan K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126-2132.

30. Gibson D. G., Young L., Chuang R.-Y., Venter J. C., Hutchison C. A., and Smith H. O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343-345.

4.8 Supplementary information

```
ATGCATCACCACCACCATCACATGAACGTTAAGAAATTCCCAGAAGGCTTCCTGTGGGGCGTTGCGACC
GCGTCTTACCAGATTGAGGGTTCCCCGCTGGCAGATGGTGCGGGCATGAGCATTTGGCACACCTTTAGC
CATACCCCGGGTAATGTTAAGAATGGCGATACGGGCGATGTTGCTTGCGACCATTACAATCGTTGGAAA
GAAGATATTGAGATTATCGAAAAGCTGGGCGTCAAGGCGTACCGCTTCAGCATCTCCTGGCCGCGTATC
CTGCCGGAAGGCACGGGCCGTGTCAATCAGAAAGGTCTGGATTTCTATAACCGCATCATTGACACCCTG
CTGGAGAAAGGTATTACCCCGTTTGTCACCATCTTCCACTGGGATCTGCCGTTTGCGCTGCAACTGAAG
GGCGGTTTGCTGAATCGTGAGATTGCCGATTGGTTCGCAGAGTACAGCCGCGTGCTGTTCGAGAACTTC
GGCGACCGTGTCAAGAATTGGATTACCTTTAACGAACCGCTGTGTAGCGCGATTCCGGGTTACGGTTCT
GGCACGTTCGCTCCAGGTCGTCAAAGCACGAGCGAGCCGTGGACGGTGGGTCATAACATTCTGGTGGCC
CACGGTCGTGCGGTCAAGGTCTTTCGTGAAACGGTTAAGGACGGTAAAATCGGTATTGTTCTGAACGGC
GACTTCACGTACCCGTGGGACGCAGCGGACCCGGCAGACAAAGAGGCCGCAGAGCGCCGTCTGGAGTTC
TTCACTGCATGGTTTGCAGACCCGATCTATCTGGGCGACTATCCAGCCAGCATGCGTAAGCAGTTGGGT
GACCGTCTGCCGACCTTTACCCCGGAAGAACGTGCGCTGGTTCACGGTAGCAACGACTTTTACGGTATG
AACCATTATACCTCGAACTATATCCGCCACCGCTCCAGCCCTGCGTCTGCGGACGATACGGTTGGCAAT
GTTGATGTGCTGTTTACCAATAAACAAGGTAACTGCATTGGCCCGGAGACTGCGATGCCGTGGCTGCGT
CCGTGTGCGGCTGGTTTCCGCGACTTTCTGGTTTGGATTAGCAAACGTTATGGTTATCCTCCGATCTAT
GTGACCGAAAATGGTGCGGCCTTCGATGATGTGGTTAGCGAGGATGGTCGCGTTCACGATCAGAATCGT
ATCGACTACCTGAAAGCATATATCGGTGCAATGGTGACCGCCGTGGAATTGGACGGTGTGAATGTAAAA
GGTTACTTTGTCTGGAGCTTGCTGGATAACTTCGAGTGGGCGGAAGGTTACAGCAAGCGTTTTGGCATC
GTGTACGTGGATTACAGCACCCAAAAACGCATCGTGAAGGACAGCGGTTATTGGTACTCCAATGTCGTC
AAAAACAACGGCTTGGAGGACTGA
```

Supplementary Figure 4.6: DNA sequence of NcrBgl.

Data collection	
Space group	P 31 1 12
Cell dimensions	
a, *b*, *c* (Å)	115.38, 115.38, 282.54
a, *b*, *g* (°)	90.0, 90.0, 120.0
Resolution (Å)	37.8 - 3.0 (3.00 - 3.15)
R_{merge}	5.2 (3.8)
I / sI	5.2 (3.8)
Completeness (%)	99.6 (99.6)
Redundancy	4.9 (4.8)
Refinement	
Resolution (Å)	37.4 - 3.0
No. reflections	41124
R_{work} / R_{free}	0.24 / 0.29
No. atoms	
Protein	9869
Ligand/ion	-
Water	2
B-factors Protein	13.1
Ligand/ion	-
Water	-
R.m.s. deviations	
Bond lengths (Å)	0.014
Bond angles (°)	1.73

Supplementary Table 4.1: Data collection and refinement statistics for 4GXP. All data were collected from a single crystal. Values in parentheses are for highest-resolution shell.

Chapter 5

Non-contiguous SCHEMA protein recombination

5.1 Summary

SCHEMA is a method of designing protein recombination libraries that contain a large fraction of functional proteins with a high degree of mutational diversity. In the previous chapter we illustrated the method for designing libraries by swapping contiguous sequence elements. Here, we introduce the NCR ("noncontiguous recombination") algorithm to identify optimal designs for swapping elements that are contiguous in the 3-D structure but not necessarily in the primary sequence. Our design recombines 3 fungal cellobiohydrolases (CBH1s) to produce a library containing more than 500,000 novel chimeric sequences.

5.2 Introduction

As discussed in Chapter 2, SCHEMA [1] seeks to maximize the probability that a library of chimeric proteins will be functional by using structural information to identify swappable sequence elements ("blocks"). We want to minimize the average SCHEMA energy ($< E >$) of all the chimeras in a library, as this increases the fraction of functional chimeras [2]. When recombining sequence elements that are contiguous along the polypeptide chain, we use RASPP [3] to identify optimal crossovers that minimize $< E >$.

In this chapter, we design SCHEMA libraries with even lower $< E >$s by removing the constraint that blocks be contiguous along the polypeptide chain. These non-contiguous blocks of sequence are still contiguous blocks of structure in the folded protein. We use non-contiguous recombination (NCR) (Smith et al. submitted) to computationally search for optimal non-contiguous SCHEMA library designs. This approach to chimera design has become feasible now that the genes can be made by total gene synthesis.

Here, we design a SCHEMA library that recombines 3 fungal cellobiohydrolases (CBH1s)

splitting each homolog into 12 blocks. Shuffling these blocks generates a chimera library of $312 = 531{,}441$ possible sequences. We have previously designed a library very similar to this one (Smith et al. in preparation) and identified several stabilizing sequence elements. NCR-designed libraries can have significantly lower disruption than RASPP (contiguous) designs from the same parent sequences. Alternatively, NCR enables recombination of parents with lower sequence identity. We recommend analysis of NCR-designed libraries by making an informative sample set of genes and using those to build predictive models, as we have done for RASPP-designed libraries [4].

5.3 Materials

1. A Unix-based computer that can run python scripts (*see* **Note 1**). Python can be downloaded from:

   http://www.python.org/download/

2. Download and unpack the NCR toolbox. This is available from:

   http://cheme.che.caltech.edu/groups/fha/media/ncr.zip

3. Download MUSCLE (*see* **Note 2**). This is available from:

   http://www.drive5.com/muscle/downloads.htm

   Unpack the compressed file and place the executable in the directory 'ncr/tools/muscle' (*see* **Note 3**).

4. Download hmetis (*see* **Note 4**). This is available from:

   http://glaros.dtc.umn.edu/gkhome/metis/hmetis/download

   Unpack the compressed file and place the hmetis folder in the directory 'ncr/tools' (*see* **Note 3**).

5. A multiple sequence alignment of the parental sequences we wish to recombine (*see* **Note 5**). This alignment should be in FASTA format (*see* **Note 6**) and the file should be named 'alignment.fasta'. As recombination parents, we pick the CBH1 sequences from *C. thermophilum*, *H. jecorina*, and *T. emersonii*, which have about 60% pairwise sequence identity. These CBH1s have a catalytic domain, a linker and a cellulose-binding domain. The available crystal structures are for the catalytic domain, thus we only considered this domain for recombination (*see* **Note 7**). To eliminate the possibility of generating unpaired disulfide bonds, we mutated two residues in the *T. emersonii* CBH1 sequence to cysteine (*see* **Note 8**). We used PROMALS3D [5] to align the parental sequences.

6. A PDB structure file of one of the parental sequences (*see* **Note 9**). We use the *T. emersonii* structure, '1Q9H.pdb'. Alternatively, if no structure is provided, the NCR tools can search for suitable structures from the PDB database (*see* **Note 10**).

5.4 Methods

1. Place the parent sequence alignment file (alignment.fasta) in the 'ncr' folder. Place the PDB structure file (1Q9H.pdb) in the directory 'ncr/structures'.

2. Set the 'Number of blocks' to 12 and 'Find all PDB structures' to 0 in the 'init.txt' file (*see* **Note 10**).

3. Run the following command (*see* **Note 11**) in the 'ncr' directory:

   ```
   python ncr.py
   ```

   This NCR script identifies a set of candidate libraries with low $< E >$ and sends these results to the terminal window (*see* **Note 12**) (Figure 5.1). These libraries are saved

in the directory 'ncr/output' and listed in the text file 'library12_result_list.csv' (*see* **Note 13**).

4. Pick an NCR library (*see* **Note 14**). In this case, we pick the library 'library12_2.output', with $< E > = 16.8$ and $< m > = 83.9$ (Figure 5.2).

5. Certain non-conserved residues still need to be assigned to blocks (*see* **Note 15**). Open 'ncr/output/library12_2.output' and assign residues 41, 175, 197, 199, 202, and 442 to blocks G, C, A, A, A, and J, respectively (*see* **Note 16**).

6. Run the following command (*see* **Note 17**) in the 'ncr' directory:

   ```
   python picklibrary.py library12_2
   ```

   This generates a list of all the chimeras in the chosen library along with their SCHEMA energies, number of mutations, and sequences (*see* **Note 18**). This list is saved as a text file 'chimeras.output' in the directory 'ncr/picked_libraries/library12_2'.

7. We synthesize the genes encoding a subset of the chimera library (*see* **Note 19**). Before expressing the CBH1 chimeras, we add a linker and cellulose-binding domain to the recombined catalytic domains.

5.5 Notes

1. The NCR toolbox 'ncr' is written for python 2.6 on a Unix-based system. We recommend using this python release for the NCR toolbox.

2. Ensure you download the correct distribution of MUSCLE for your system. For example, on Apple OS X it might be 'muscle3.8.31_i86darwin64.tar.gz'. The NCR tools were written for MUSCLE 3.8.

3. The NCR toolbox unpacks as a folder called 'ncr'. Directories are given relative to this folder. For example there is a folder in 'ncr' called 'tools' and the directory would be 'ncr/tools'.

4. Ensure you download the correct distribution of hmetis for your system. For example, on Apple OS X it might be 'hmetis-1.5-osx-i686.tar.gz'. The NCR tools were written for hmetis 1.5.

5. We assume the parental proteins share the same structural fold. If structures are available for more than one parental protein, we confirm the parents have the same fold by aligning the parental structures. It is important that the sequence alignment is accurate, especially when the parent sequence identities are low.

6. In FASTA format, the name of each sequence begins with '>', for example, '>Temersonii'. After each name there should be a return, followed by the corresponding aligned sequence.

7. SCHEMA library designs require a protein structure. If no structural information is available for a parent sequence, but there are structures of homologs, we can use MODELLER to build a structure model [6]. An inaccurate homology model hinders SCHEMA library design; an actual structure is preferred.

8. We assumed but did not verify that broken disulfide bonds are destabilizing. In this case, *C. thermophilum* and *H. jecorina* CBH1s have 10 disulfide bonds while *T. emersonii* has 9 disulfide bonds. If the cysteines from the missing disulfide bond are in separate sequence blocks, chimeras with unpaired cysteines can result. We avoided this by modifying the parental sequence of *T. emersonii* to include the remaining cysteine pair.

9. One or more structures is needed to identify the residue-residue contacts. When possible, we pick high-resolution structures ($< 2.0\mathring{A}$). If a PDB file contains more than one chain, each chain is split into its own structure file labeled XXXX.A.pdb, XXXX.B.pdb, etc. The NCR tools can handle multiple structures. Residue-residue contacts from multiple structures of the same parent form a parental contact map if these contacts are present in at least 50% of the structures. If structures from multiple parents are used, each contact is weighted by the fraction of parental contact maps it appears in.

10. The 'init.txt' file is in the 'ncr' folder. It specifies two parameters for the NCR toolbox:

    - 'Number of blocks': The number of blocks in the designed libraries. It can either be a number (e.g. 8) or a range of numbers (e.g. 2-6) for designing a range of libraries with different block sizes.
    - 'Find all PDB structures': If 1, the NCR script will search, download and use all suitable structures from the PDB database. If 0, the user will provide one or more structures.

    Increasing the number of blocks in a library increases library size and reduces the average number of mutations in a block. The user may want smaller blocks if searching for single mutations that cause specific functional changes. However, it is harder to find desirable chimeras in larger libraries and increasing the number of blocks increases a library's $< E >$. We chose to split our 3 parent proteins into 12 blocks.

11. The python script 'ncr.py' generates one or more parental contact maps, calculates the SCHEMA contacts and searches for low $< E >$ libraries. This script may take several hours to complete, depending on protein size and computer specifications. Progress

is displayed in the terminal window. The script uses heuristic algorithms to find near optimal solutions, thus results will vary each time 'ncr.py' is run.

12. In the terminal window, NCR lists $< E >$ and $< m >$ for each library as well as the distribution of mutations among the 12 blocks. This distribution is given as a list of 12 numbers, each referring to the number of mutations in a block with blocks counting A, B, C, etc. There is a trade-off between the average SCHEMA energy of a library ($< E >$) and how evenly distributed mutations are among the blocks. If all the blocks are evenly sized, the solution space of possible libraries is small and so $< E >$ is large. As block sizes become uneven, the solution space of possible libraries increases. This enables NCR to find libraries with lower $< E >$, but libraries with very unevenly sized blocks may not be useful. NCR is designed to find low $< E >$ libraries for a range of block sizes.

13. In non-contiguous recombination, a library is defined by assigning every non-conserved residue to a block. In the library text file 'library12_2.output', a designated block (named 'A', 'B', 'C', etc.) appears beside every non-conserved residue. A dash ('-') is placed next to every conserved residue. Residues are numbered based on the parental sequence alignment. The results file 'library12_result_list.csv' lists $< E >$ and $< m >$ for each library.

14. NCR returns a set of candidate libraries with a range of $< m >$ values. A lower $< E >$ implies more functional chimeras in the library. For moderately-sized proteins (250-500 amino acids) we try to pick SCHEMA libraries with $< E >$ less than 30. For non-contiguous recombination of homologs with >55% sequence identity, often all the candidate libraries have $< E >$ below 30. In our case, we pick a library with evenly

sized blocks. This will make it easier to identify stabilizing point mutations within a stabilizing block. Protein-specific biochemical and structural knowledge may also help users pick from the candidate libraries. Note that the $< E >$ value is lower and the $< m >$ value higher in this NCR design than the previously described RASPP design.

Blocks are not always one contiguous piece of structure. Sometimes, a group of residues will only have SCHEMA contacts with one another and not with the rest of the protein. These 'disconnected blocks' can belong to any block without altering $< E >$. NCR will assign these disconnected blocks to blocks such that $< m >$ is maximized. This can result in a block comprising two separate pieces of structure. These disconnected blocks are apparent when blocks are visualized on the PDB structure. In this case, blocks 'A', 'G', and 'J' each contain a disconnected block.

15. Some non-conserved residues do not have any SCHEMA contacts. These residues often appear on the surface of the protein, in a region that is highly conserved or in a region where structural information is missing. NCR does not assign these residues to a block and instead the decision is left to the user. Unassigned residues are printed to the terminal. In this case residues 41, 175, 197, 199, 202, and 442 have not been assigned a block.

16. Looking at the structure '1Q9H.pdb', we designate each unassigned residue to the same block as one of its neighboring residues. This will slightly alter $< m >$ for the library, but leave $< E >$ unaffected. We can alter the block assignments by editing the text file 'ncr/output/library12_2.output'. In this file unassigned residues, like conserved residues, have a dash ('-') in place of a block ('A', 'B', 'C', etc.).

17. The python script 'picklibrary.py' generates all the chimeras in a given library. The name of the library 'library12_2' needs to be provided as an argument. Any non-conserved residues that have not been assigned to a block will be automatically assigned to block A. For a large library such as this one (more than 500,000 chimeras), this script may take several hours to complete.

18. Chimeras are numbered according to the parental sequence of each block with the numbers ordered from the first to the last block. Parents are numbered based on the order they appear in the parental sequence alignment. For example, chimera '132213131322' has parent 1 as the sequence of its first block ('A'), parent 3 as its second block ('B'), etc. The amino acid sequence provided alongside each chimera in 'chimeras.output' is built from the parent sequence alignment. It contains dashes ('-') where there are gaps in the alignment. These dashes should be removed when ordering the synthetic genes.

19. These chimeras are very difficult to construct with traditional cloning techniques. We pick a subset of the library to synthesize and analyze. We ensure every block from every parent is represented independently of one another in this subset. This enables us to model the effects of the different blocks on biochemical properties such as stability [7]. We pick a set of chimeras to be most informative using the Submodular Function Optimization Matlab toolbox [8, 9]. Alternatively, we could have picked a set of chimeras that substitute one block at a time into the background of a parent that expresses well, such as *T. emersonii* CBH1 [10].

5.6 Figures

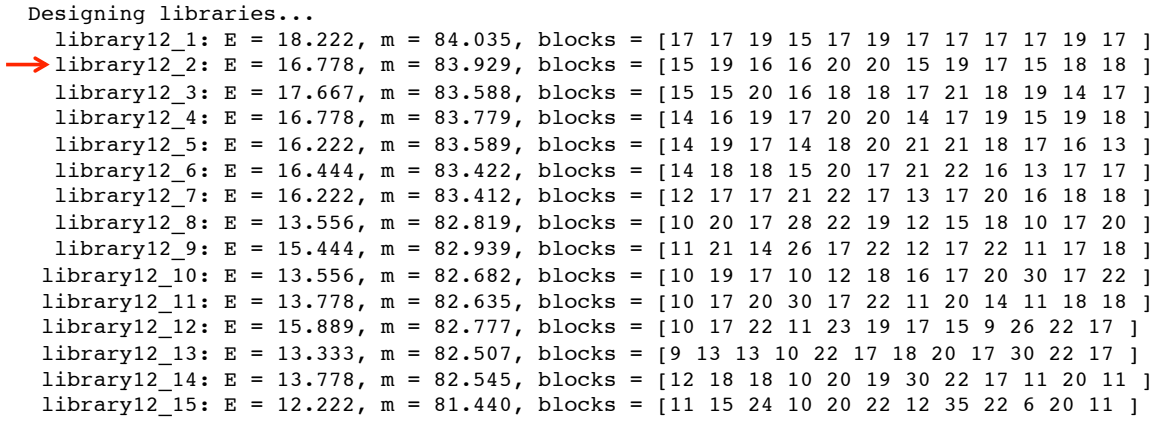


Figure 5.1: Libraries returned by NCR. The average SCHEMA energy ($< E >$) and average number of mutations ($< m >$) for each library is printed to the terminal window. In addition, the output displays the distribution of the mutations among the 12 blocks. Libraries with higher $< E >$ have more evenly sized blocks. The chosen library is highlighted with an arrow.

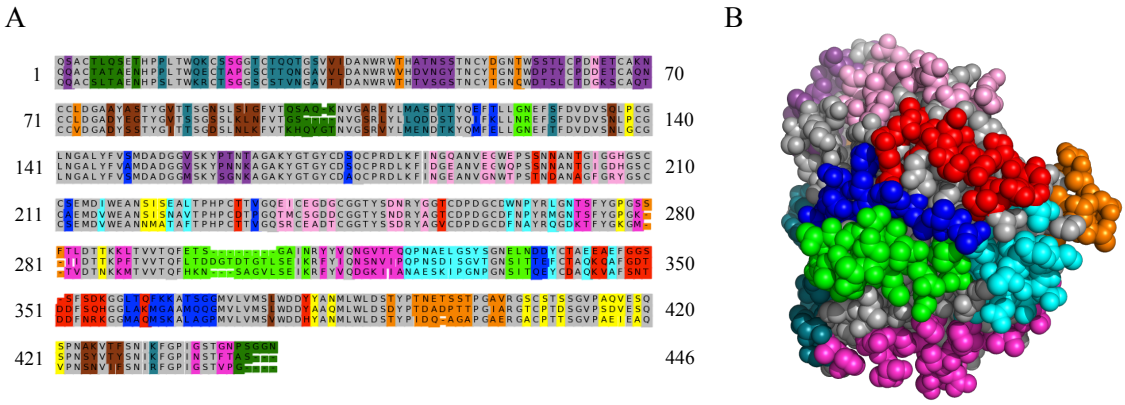


Figure 5.2: Visualizing the chosen NCR design. A) The multiple sequence alignment of the parent CBH1s with each of the 12 blocks highlighted in a different color. Conserved residues are colored gray. It is clear that the blocks are non-contiguous along the polypeptide chain. B) The blocks highlighted on the CBH1 structure '1Q9H.pdb'. Most of the blocks are contiguous structural elements in 3-D.

5.7 References

1. Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L., and Arnold, F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553-558

2. Meyer, M., Hochrein, L., and Arnold, F. H. (2006) Structure-guided SCHEMA recombination of distantly related β-lactamases. *Protein Eng. Des. Sel.* **19**, 563-570

3. Endelman, J., Silberg, J., Wang, Z., and Arnold, F. H. (2004) Site-directed protein recombination as a shortest-path problem. *Protein Eng. Des. Sel.* **17**, 589-594

4. Heinzelman, P., Romero, P. A., and Arnold, F. H. (2013) Efficient sampling of SCHEMA Chimera Families for Identification of Useful Sequence Elements. In: Keasling, A (ed) *Methods in Enzymology: Methods in Protein Design*, Elsevier Ltd, Oxford, U.K.

5. Pei, J., Kim, B.-H., and Grishin, N. V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295-2300

6. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A. (2007) Comparative protein structure modeling using Modeller. *Curr. Protoc. Protein Sci.* **2**, 15-32

7. Li, Y., Drummond, D. A., Sawayama, A. M., Snow, C. D., Bloom, J. D., and Arnold, F. H. (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25**, 1051-1056

8. Krause, A. (2010) SFO: A Toolbox for Submodular Function Optimization. *J. Mach. Learn. Res.* **11**, 1141-1144

9. Romero, P., Stone, E., Lamb, C., Chantranupong, L., Krause, A., Miklos, A., Hughes, R., Fechtel, B., Ellington, A. D., Arnold, F. H., and Georgiou, G. (2012) SCHEMA-designed variants of human arginase I and II reveal sequence elements important to stability and catalysis. *ACS Synth. Biol.* **1**, 221-228

10. Heinzelman, P., Komor, R., Kanaan, A., Romero, P. A., Yu, X., Mohler, S., Snow, C., and Arnold, F. H. (2010) Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Eng. Des. Sel.* **23**, 871-880

Chapter 6

H. jecorina cellobiohydrolase I stabilizing mutations identified using non-contiguous recombination

6.1 Abstract

Non-contiguous recombination (NCR) is a simple method to identify pieces of structure that can be swapped among homologous proteins. These 'blocks' are encoded by elements of sequence that are not necessarily contiguous along the polypeptide chain. We used NCR to design a library in which blocks of structure from *H. jecorina* cellobiohydrolase I (Cel7A) and its two thermostable homologs from *T. emersonii* and *C. thermophilum* are shuffled to create 531,438 possible chimeric enzymes. We constructed a maximally informative subset of 35 chimeras to analyze this library and found that the blocks contribute additively to the stability of a chimera. Within two highly stabilizing blocks, we uncovered six single amino acid substitutions that each improve the stability of *H. jecorina* cellobiohydrolase I by 1 - 3°C. The small number of measurements required to find these mutations demonstrates that non-contiguous recombination is an efficient strategy for identifying stabilizing mutations.

6.2 Introduction

Highly thermostable cellulases are desirable for the production of sugars from cellulosic substrates. Thermo-tolerant mixtures of fungal cellulases have been shown to degrade cellulose faster at elevated temperatures than mixtures from mesophilic fungi [1]. At elevated temperatures, cellulolytic processes can benefit from reduced contamination and viscosity of the biomass slurry as well as increased cellulase hydrolysis rates.

The success of the filamentous fungus *Hypocrea jecorina* (anamorph *Trichoderma reesei*) as an industrial cellulase producer derives from its ability to secrete cellulases at up to 100 g/L. Cellobiohydrolase I (CBHI, Cel7A) is one of the most important cellulase components. Removal of the *cbh1* gene reduces the cellulolytic activity of the fungus by 70% and the

total secreted protein by 40% [2]. Low expression levels and altered glycosylation patterns [3, 4], however, make this enzyme difficult to engineer in heterologous expression systems.

There have been various efforts to engineer improved CBHI variants, including screening random mutants [5], engineering disulfide bonds [6], and DNA shuffling [7]. In addition, we have sought to enhance CBHI stability through protein recombination [8] and predictive methods [9]. The most stable CBHI enzymes from these latter works have more than 150 mutations from *H. jecorina* CBHI, which could adversely affect the high titers of secreted protein in fungal expression systems. We therefore sought to stabilize the *H. jecorina* CBHI with minimal mutation to its sequence and in a way that does not decrease its activity.

We recently introduced a method for non-contiguous protein recombination [10] that identifies elements of structure ('blocks') that can be shuffled among homologous proteins. Unlike previous SCHEMA recombination libraries that swap elements of sequence [11], these elements of structure are not necessarily contiguous polypeptide sequences. Here we show how non-contiguous recombination can be used to efficiently identify stabilizing mutations that have been incorporated into CBHI homologs in nature.

Swapping structural blocks among *H. jecorina* CBH1 and two thermostable homologs from *T. emersonii* and *C. thermophilum*, we analyze a subset of CBHIs from a library containing more than 500,000 possible chimeric sequences. We predict the thermostabilities of all library members using data from a maximally informative subset of just 32 chimeras (and 3 parents) and identify several blocks that are predicted to stabilize *H. jecorina* CBHI. Searching within these blocks, we find six single amino acid substitutions that stabilize *H. jecorina* CBHI by more than 1°C. One previously undiscovered mutation improves its thermostability by 3°C.

6.3 Results

6.3.1 Non-contiguous protein recombination library design

We wish to shuffle elements of sequence among homologous proteins to create a library of chimeras highly enriched in functional sequences. A good metric for the functional impairment of a chimeric protein is its SCHEMA disruption [12], which is the number of non-native residue-residue contacts formed in the recombined sequence. We have used this metric previously to design recombination libraries that shuffled contiguous blocks of sequence [11, 13]. Recombination of structural elements can be significantly less disruptive than recombining sequence elements. We recently presented a method for finding the optimal structural blocks for any given set of parent proteins, based on a graph partitioning algorithm [10].

For NCR, we create a graph from the non-native residue-residue contacts, with nodes corresponding to residues and edges corresponding to non-native contacts. NCR minimizes the SCHEMA disruption by identifying minimal cuts that partition the graph [10]. We partition the graph with hmetis [14, 15], a suite of graph partitioning tools. Residues are assigned to blocks based on how nodes are assigned to partitions. Blocks can have non-contiguous sequences but will be contiguous pieces of structure in 3 dimensions. Shuffling these blocks generates a library of non-contiguous chimeras.

As parental enzymes we chose the catalytic domains of three fungal CBHI cellulases: *H. jecorina* CBHI (P1), *T. emersonii* CBHI (P2), and *C. thermophilum* CBHI (P3). *T. emersonii* CBHI has one fewer disulfide bond than *H. jecorina* CBH1 and *C. thermophilum* CBH1, which each have 10. To ensure unpaired cysteines do not appear in the chimeras, we mutated P2 to include the missing cysteine pair (G4C, A72C). This extra disulfide bond is

known to increase the stability of P2 [6]. These three cellulase catalytic domains were used in a previous study to create a contiguous block SCHEMA recombination library [8].

For ease of identifying stabilizing point mutations within a block, we divided mutations among blocks so each block contained only a small number of mutations. We also required the blocks to be of equal size to ensure a fair comparison of block stability contributions. We designed a 12-block library where each block contained approximately 18 non-conserved residues (see Materials and Methods). The design has an average SCHEMA disruption (number of disrupted contacts) of 24.8 and an average of 83.4 mutations from the closest parent. Most non-native residue-residue contacts are sequestered within blocks (Figure 6.1A), which increases the fraction of the library that is likely to be folded and functional. While almost all the blocks are contiguous pieces of structure (Figure 6.1B), they each comprise many fragments of the polypeptide chain (Figure 6.1C).

Some groups of residues only have SCHEMA contacts with one another and not with the rest of the protein. These disconnected 'sub-blocks' can belong to any block without altering the SCHEMA disruption and they appear separate from the rest of the block. Blocks 'A', 'D', 'E' and 'J' contain disconnected sub-blocks and thus contain several separate pieces of structure.

6.3.2 Stabilities of an informative subset

Our 3-parent, 12-block library contains more than half a million chimeras. The nature of non-contiguous recombination makes it very difficult to construct these chimeras with traditional cloning techniques. Because it is neither feasible nor necessary to synthesize and analyze the entire library, we selected a highly informative subset of 35 chimeras to construct and characterize. These chimeras were chosen to maximize mutual information

about the sequences (see Materials and Methods), as described previously for a library of chimeric arginases [16]. At the same time, the chosen sequences had low SCHEMA disruption in order to enrich the chimera subset in functional sequences. To the C-terminus of each chimeric catalytic domain we added the linker and carbohydrate binding module from *H. jecorina* CBHI.

Of the 35 chimeras synthesized, 32 (91%) were expressed with detectable levels of activity. We quantified the thermostabilities of the 32 chimeras and three parents using two measures. We define T_{R50} as the incubation temperature at which an enzyme loses half its (unincubated) activity. We incubated the enzymes at a range of temperatures for 10 minutes without substrate and measured the residual activities. These T_{R50}s are plotted in Supplementary Figure 6.7. Most of the chimeras have stabilities that lie between those of the parents, but several were more stable than the most stable parent, P2.

We define T_{A50} as the elevated temperature at which an enzyme loses half its activity measured at its optimum temperature. We ran a 2-hour activity assay at a range of temperatures and measured the total enzyme activities. Whereas T_{R50} is a measure of enzyme tolerance to thermal stress, T_{A50} measures an enzyme's ability to function at elevated temperature.

Values of T_{A50} are plotted in Figure 6.2A. T_{R50}s and T_{A50}s are correlated for the chimeras (Figure 6.2B), but there are some outliers where the T_{R50} greatly exceeds the T_{A50}. Even though the enzymes are incubated in 1 mM DTT, these are cases where the CBHIs are able to refold and regain activity once the temperature is reduced for the assay (Supplementary Figure 6.8).

6.3.3 Modeling thermostability

We have previously shown that contiguous blocks of sequence contribute additively to the stabilities of chimeras and that these stabilities are predictable with simple additive block models trained on a small sample of a library [17, 18]. Here we used the same linear regression model to demonstrate that contiguous blocks of structure (with non-contiguous blocks of sequence) also contribute additively to the stabilities of recombined enzymes. We constructed predictive models of T_{R50} and T_{A50} based on the sequences of the 32 functional chimeras and three parental cellulases (see Materials and Methods). As shown in Figure 6.3A, the T_{R50} model accurately predicts the stabilities of the library sample ($r^2 = 0.81$). This model provides the predicted contributions of each structural block to T_{R50} (Figure 6.3B). Similarly, we trained a model that fits the T_{A50} stability data ($r^2 = 0.74$, Figure 6.3C). The predicted block contributions to T_{A50} are shown in Figure 6.3D. In both models, block G appears to be highly stabilizing to parent P1 when taken from either parent P2 or P3. There are two mutations common to P2 and P3 in this block, T360A and F362M.

With the stability models constructed from this highly informative sample set, we can predict the T_{R50}s and T_{A50}s of all the untested chimeras in the library. We correctly identified seven chimeras from the library expected to have both high T_{R50}s and T_{A50}s (Figure 6.4A and B). While two of the predicted chimeras had T_{R50}s 2°C higher than the most stable parent (P2), none of the chimeras had T_{A50}s above the most stable parent.

6.3.4 Stabilizing point mutations

We wish to stabilize *H. jecorina* CBHI (P1) with minimal disruption to its amino acid sequence. Using linear regression, we have identified two highly stabilizing blocks, block G from P2 and block G from P3. We placed each of these blocks in place of block G in

H. jecorina CBHI and found they were indeed stabilizing, improving *H. jecorina* CBHI's T_{R50} by 1.7°C and 1.1°C, respectively (Supplementary Table 6.3). Given that a single block is made up of a combination of stabilizing and destabilizing mutations, we wanted to identify the individual amino acids that have the most significant positive contribution to stability. Similar to the approach we used on Cel6 cellulases [19], we searched these blocks for individual mutations that stabilize P1 by substituting each of the 23 point mutations into P1 and measuring the T_{A50} (Figure 6.5A and B). Most of the amino acid substitutions have only a slight effect on *H. jecorina* CBHI thermostability (less than 1°C). Of the remaining mutations, three are destabilizing and six are stabilizing. One of the stabilizing mutations, F362M, present in both P2 and P3, is stabilizing by a full 3°C. This mutation allows *H. jecorina* CBHI to retain higher levels of activity at elevated temperatures (Figure 6.6).

We created seven mutants that were combinations of the most stabilizing mutations. None were more stable than the F362M single mutant (Supplementary Table 6.4).

6.4 Discussion

We have utilized a new non-contiguous recombination method to design a library with more than 500,000 sequences enriched in functional family 7 cellobiohydrolases. NCR identifies swappable elements of structure that are not necessarily contiguous pieces of polypeptide. Because library designs with contiguous sequence elements are a small subset of the large number of possible non-contiguous designs, this approach identifies libraries that disrupt fewer SCHEMA contacts and therefore contain more functional chimeric proteins than contiguous block design algorithms such as RASPP [20]. Indeed, in the library 39 of the 42 synthesized chimeras were functional even though on average they had 79 mutations from the closest parent.

By measuring the thermostabilities of a maximally informative subset of 32 chimeras, we showed that the structural blocks identified by NCR contribute additively to protein stability. Furthermore, we used a block-additive stability model to correctly identify several stable chimeras in the library.

H. jecorina CBHI, *T. emersonii* CBHI, and *C. thermophilum* CBHI differ at 213 residues. It is difficult to identify the stabilizing mutations from just analyzing the sequences, and it would be cumbersome to construct and test all 273 single mutants. The 'divide and conquer' method we present first measures the stabilities of functional groups of mutations (blocks) and then identifies stabilizing single mutants within the most stable blocks. With a small number of experimental measurements, we identified two blocks that significantly stabilize *H. jecorina* CBHI and uncovered a single amino acid substitution that stabilizes this important industrial enzyme by 3°C.

Despite much previous work on stabilizing *H. jecorina* CBHI, the F362M mutation has not been described previously. The mutation is located close to the surface of the protein facing inwards and the sulfur atom is proximal to the sulfur of another methionine residue. The enhanced stability may come from interaction of these two residues, possibly a hydrogen bond if one of the methionines is oxidized to methionine sulfoxide.

NCR identifies elements of structure that, when swapped, preserve protein function [10]. Splitting the CBHI structure into a relatively large number of equally sized blocks and swapping these structural elements with stable homologs has proven to be an efficient strategy to search for stabilizing mutations. While we tested all 23 single mutations from the two most stabilizing blocks, the most stable single mutation was present in both blocks. A method of prioritizing point mutations within a stable block, such as using consensus mutagenesis, may further improve the speed with which valuable mutations are identified.

6.5 Materials and methods

6.5.1 Non-contiguous recombination

PROMALS3D [21] was used to create a structure-based sequence alignment of the catalytic domains from *H. jecorina* CBHI (P1), *T. emersonii* CBHI (P2), and *C. thermophilum* CBHI (P3). Residues that have (non-hydrogen) atoms closer than 4.5$\mathring{A}$ are considered to be in contact with one another. All residue-residue contacts were identified in PDB structure 1Q9H.pdb chain A. Contacts not conserved among the three parent enzymes form the SCHEMA contact map.

Designing libraries that minimize the average number of SCHEMA contacts in the resulting chimeras was reformulated as a graph partitioning problem. The SCHEMA contact map was transformed into a graph with each node representing a non-conserved residue and each weighted edge representing an average SCHEMA contact between two residues. Residues were assigned to blocks such that the sum of weighted edges between blocks was minimized. For the 12-block library designs, the hmetis graph partitioning suite [14, 15] was used to perform a series of 12-way partitions of the SCHEMA contact map. A library design was chosen with an average SCHEMA energy (number of disrupted contacts) of 24.8 and an average of 83.4 mutations from the closest parent. Residues 41, 175, 197, 199, 202, and 442 have no SCHEMA contacts and were not partitioned into blocks; we assigned these residues to blocks D, G, B, A, A, J, respectively, based on their spatial proximity to those blocks. The C-terminal linker and carbohydrate binding module from *H. jecorina* CBHI was appended to each chimera.

6.5.2 Optimal experimental design

A greedy algorithm was employed to find a subset of sequences from the library with low SCHEMA disruption and maximized mutual information, as described [16]. Due to computational constraints, the informative set of chimeras was identified from 50,000 randomly chosen chimeras with a SCHEMA disruption below 30, rather than the entire library. This optimized experimental design was carried out with the Submodular Function Optimization Matlab toolbox [22].

6.5.3 Gene synthesis

The chimeric CBHI genes were optimized for expression in *S. cerevisiae* and synthesized by DNA2.0 (Menlo Park, CA, USA).

6.5.4 Protein expression

The genes encoding parental and chimeric CBHI catalytic domains were cloned into the yeast expression vector Yep352/PGT91-1-αss with an N-terminal His6 tag and the *H. jecorina* CBHI linker and cellulose binding domain attached to the C-terminus. These vectors were transformed into yeast strain YDR483W BY4742 ($Mat\alpha$ $hus3\Delta1$ $leu2\Delta0$ $lys2\Delta0$ $ura3\Delta0$ $\Delta kre2$, ATCC No. 4014317) as described [23] and plated on synthetic dropout-uracil medium with 10 g/L agar. The plates were incubated for 2 days at 30°C. 5 mL of synthetic dropout-uracil medium was inoculated by a single yeast colony from a plate and incubated for 1 day at 30°C, with shaking at 250 rpm. Cultures were expanded at a 1:10 ratio into either 10 mL or 50 mL of yeast peptone dextrose (YPD) medium (10 g yeast extract, 20 g peptone, 20 g dextrose) and incubated for 2 days at 30°C, with shaking at 250 rpm. The cells were pelleted by centrifugation at 5000 g for 10 min and the supernatant,

containing the secreted cellulases, was decanted and separated through a 0.20 μm pore size conical filter unit from Nalgene (Rochester, NY, USA). The supernatant was concentrated up to 4-fold using Vivaspin 20 spin columns with a 30 kDa MWCO PES membrane from GE Healthcare (Little Chalfont, UK) and stored at 4°C with 0.02% sodium azide and 1 mM phenylmethanesulfonyl-fluoride.

6.5.5 Thermostability residual activity assay (T_{R50} measurement)

In a 96-well PCR plate, 100 μL of supernatant is added to 25 μL of 625 mM sodium acetate, pH 4.8 with 5 mM dithiothreitol (DTT), giving a final concentration of 125 mM sodium acetate, pH 4.8 and 1 mM DTT, as described [9]. The plate is incubated in a gradient thermocycler for 10 min at a range of temperatures, and then cooled to 4°C. To each well, 25 μL of 1.8 mM 4-methylumbelliferyl lactopyranoside (MUL) from Sigma-Aldrich (St. Louis, MI, USA) dissolved in 18% DMSO and 125 mM sodium acetate was added. The heat-treated cellulases were incubated in a thermocycler for 90 min at 45°C. The reaction was quenched by adding 150 μL of 1 M Na2CO3 and cellulase activity was quantified by measuring the fluorescence of released 4-methylumbelliferone with excitation at 364 nm and emission at 445 nm.

6.5.6 Thermostability activity assay (T_{A50} measurement)

In a 96-well PCR plate, 100 μL of supernatant is added to 25 μL of 625 mM sodium acetate, pH 4.8, and 25 μL of 1.8 mM 4-methylumbelliferyl lactopyranoside (MUL) dissolved in 18% DMSO and 125 mM sodium acetate. The plate is incubated in a gradient thermocycler for 90 min at a range of temperatures, and then cooled to 4°C. The reaction was quenched by adding 150 μL of 1 M Na_2CO_3 and cellulase activity was quantified by measuring the

fluorescence of released 4-methylumbelliferone with excitation at 364 nm and emission at 445 nm.

6.5.7 Linear regression

Stability models for T_{R50} and T_{A50} were constructed as described previously [17] and trained using Matlab's 'regress' function.

6.6 Figures

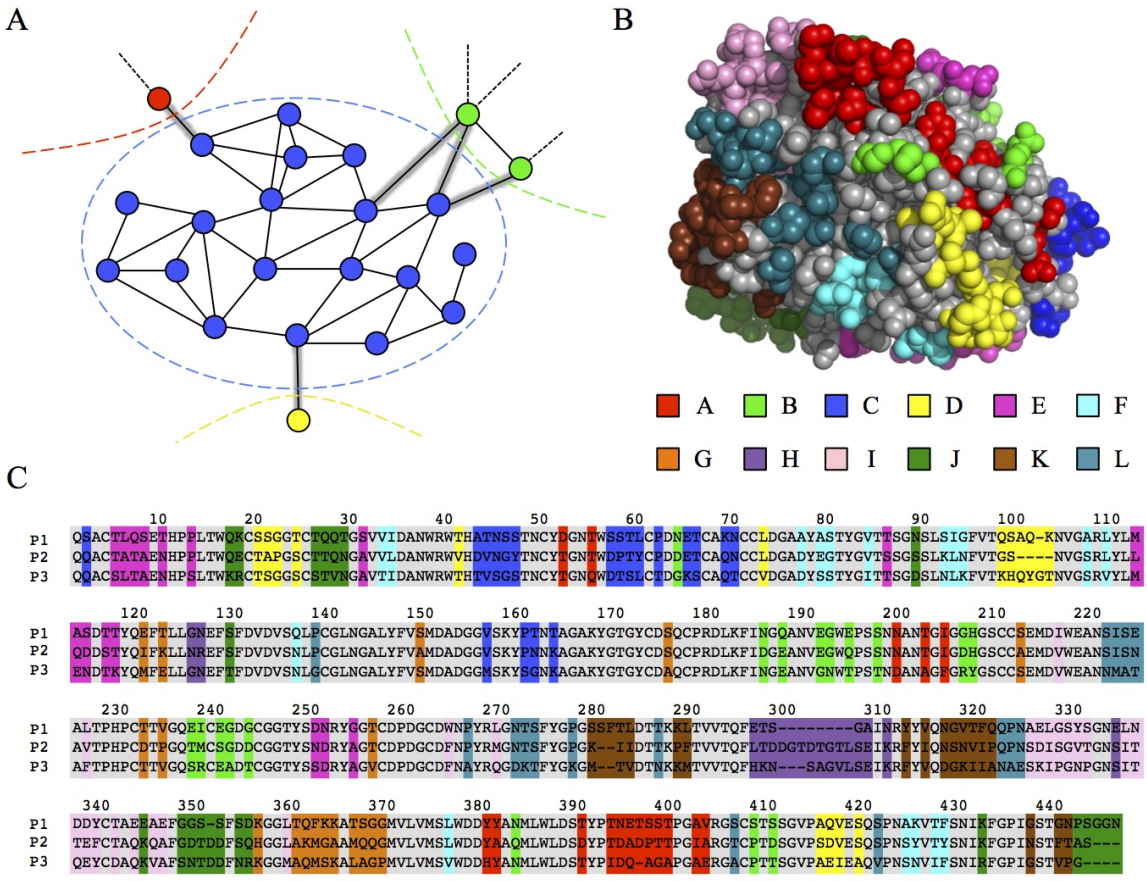


Figure 6.1: Non-contiguous recombination library design. A) A graph view of the blue block and neighboring residues. Nodes represent residues and edges represent residue-residue contacts. Colored, dashed lines define the graph partitions for each block. Contacts to residues from other blocks (highlighted) will be broken upon recombination. B) The 12-block design displayed on the structure of P2 (1Q9H.pdb). Each block (labeled A to L) is represented by a different color and conserved residues are in gray. C) The 12-block design displayed on the numbered sequence alignment of the catalytic domains of the three parental enzymes.

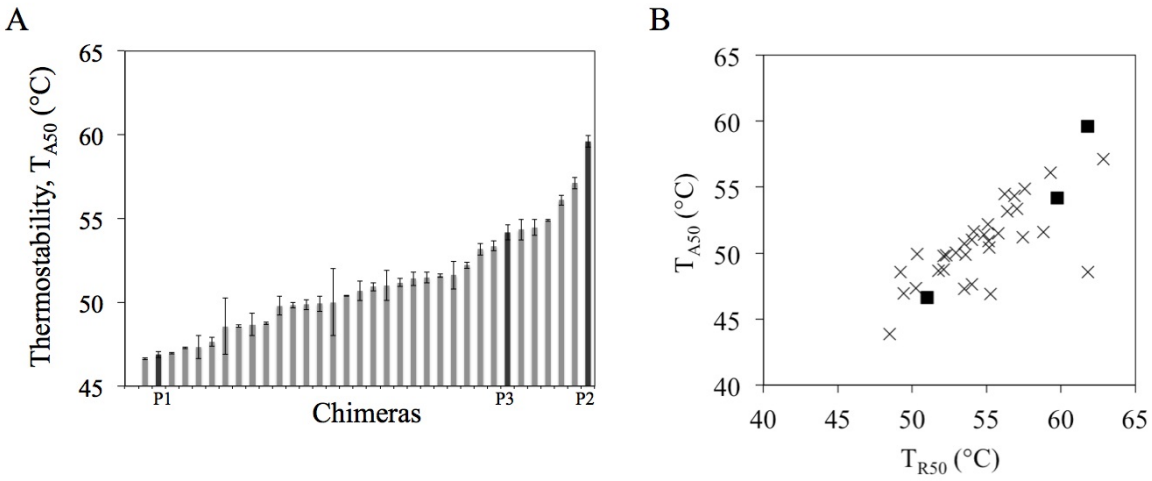


Figure 6.2: Thermostabilities of a maximally informative subset of the library. A) T_{A50}: the elevated temperature at which a chimera's activity is half its maximum. Measurements were performed in duplicate. The parental enzymes are highlighted. B) A plot of the elevated temperature at which an enzyme loses half its activity (T_{A50}) against the incubation temperature at which an enzyme loses half its (unincubated) activity (T_{R50}). The parental cellulases are highlighted with black squares. While most of the T_{R50} and T_{A50} measurements are similar, several cellulases have significantly higher T_{R50} than T_{A50}.

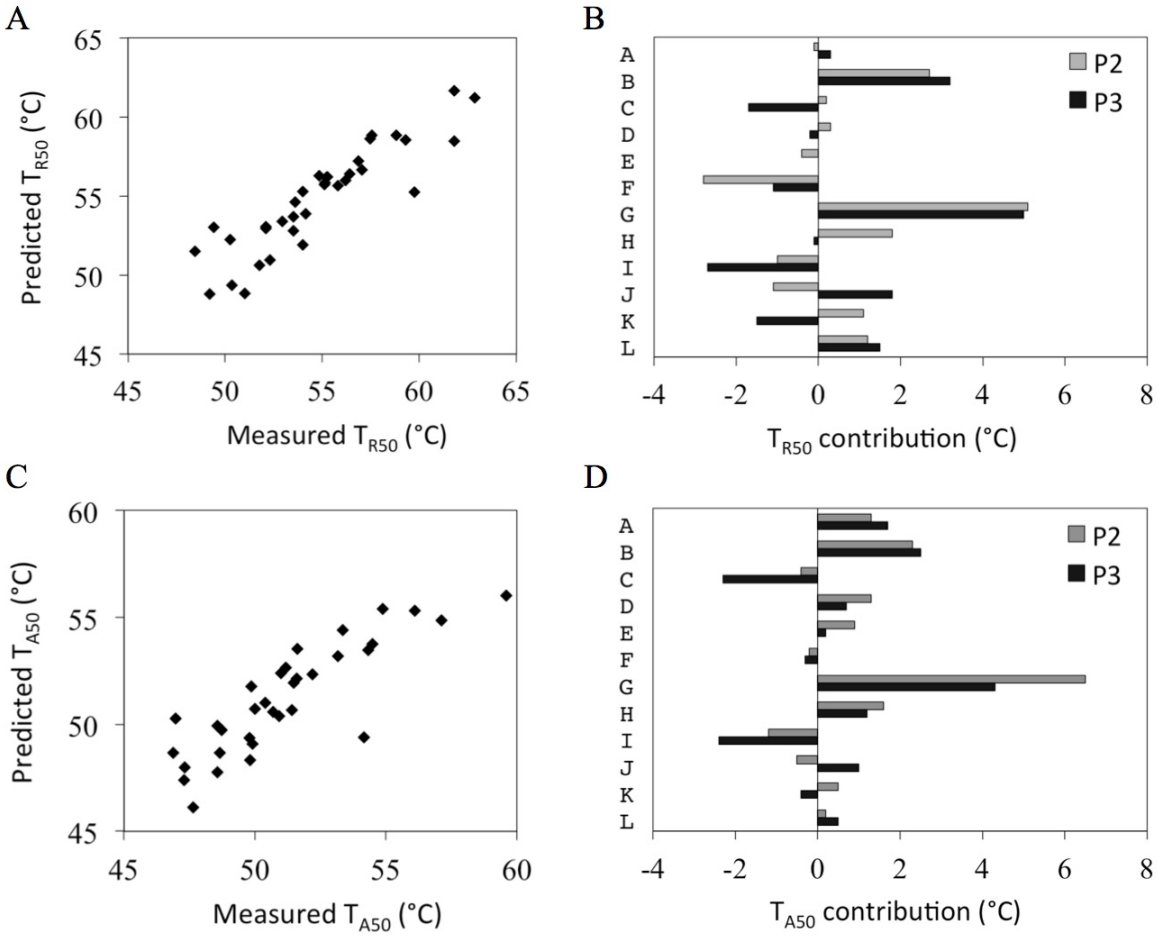


Figure 6.3: The thermostability of a chimera can be predicted with a simple linear model that sums the contributions from each block. A) A linear thermostability model trained on the T_{R50}s of the chimeras accurately predicts the measured values ($r^2 = 0.81$). Blocks G and B from *T. emersonii* and *C. thermophilum* are predicted to be significantly stabilizing relative to those blocks from *H. jecorina*. B) The predicted T_{R50} contributions of each block from parents P2 and P3 relative to parent P1. C) A linear thermostability model trained on the T_{A50}s of the chimeras accurately predicts the measured values ($r^2 = 0.74$). D) The predicted T_{A50} contributions of each block from parents P2 and P3 relative to parent P1.

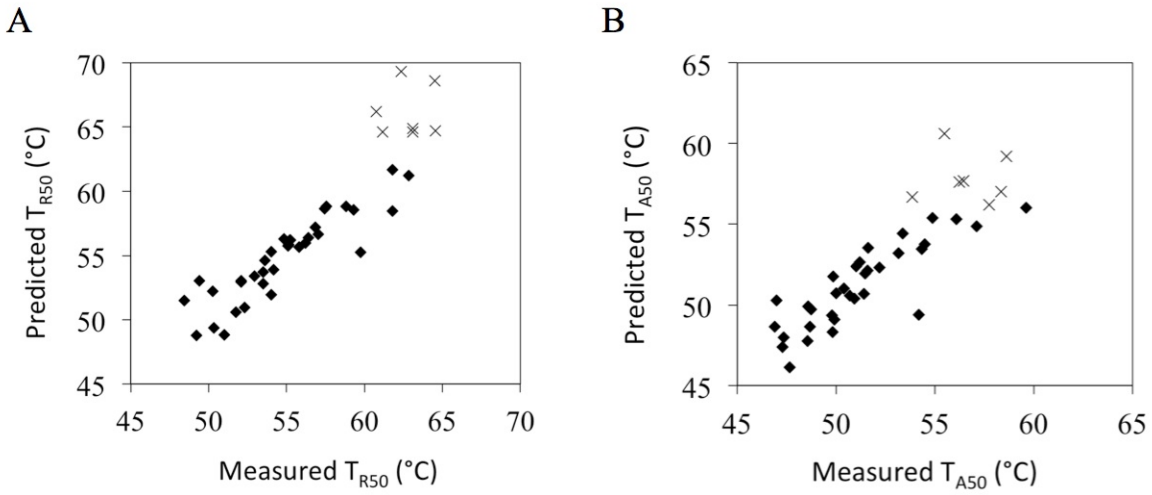


Figure 6.4: The thermostability models identify stable CBHI chimeric cellulases in the library. A) Predicted T_{R50} against measured T_{R50} for seven chimeras predicted to have high stabilities (crosses). The original data used to train the model are represented as filled diamonds. B) Predicted T_{A50} against measured T_{A50} for seven chimeras predicted to have high stabilities (crosses). The original data used to train the model are shown as filled diamonds.

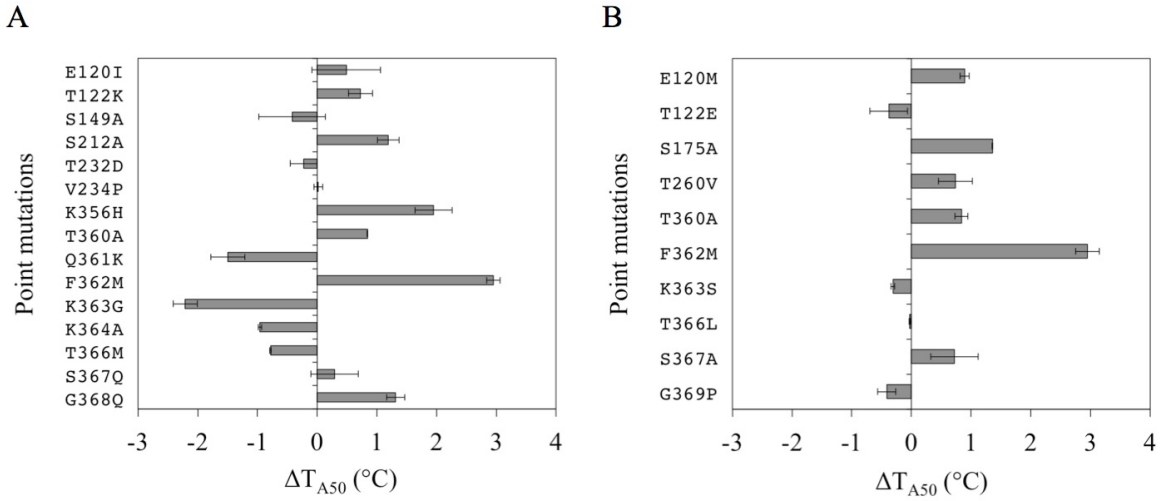


Figure 6.5: The effect on *H. jecorina* CBHI thermostability (T_{A50}) for a series of point mutations from two of the most stabilizing blocks. A) Block G, parent P2. B) Block G, parent P3. Two of the mutations (T360A and F362M) are present in both blocks. F362M is stabilizing by 3°C.

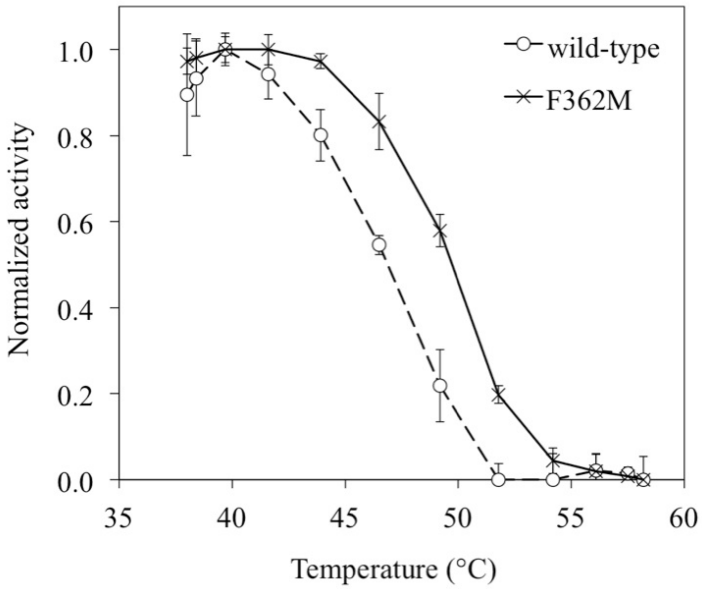


Figure 6.6: Effect of the mutation F362M (x) on the activity of *H. jecorina* CBHI (o) in a 90-minute assay on MUL over a range of temperatures, performed in quadruplicate. To account for differences in levels of secreted cellulase, CBHI activity was normalized by the activity at 40°C.

6.7 References

1. Viikari L., Alapuranen M., Puranen T., Vehmaanpera J., and Siika-Aho M. (2007) Thermostable enzymes in lignocellulose hydrolysis. *Adv. Biochem. Eng. Biotechnol.* **108**, 121-145.

2. Suominen P. L., Mantyla A. L., Karhunen T., Hakola S., and Nevalainen H. (1993) High frequency one-step gene replacement in Trichoderma reesei. II. Effects of deletions of individual cellulase genes. *Molec. Gen. Genet.* **241**, 523-530.

3. Boer H., Teeri T. T., and Koivula A. (2000) Characterization of Trichoderma reesei cellobiohydrolase Cel7A secreted from Pichia pastoris using two different promoters. *Biotechnol. Bioeng.* **69**, 486-494.

4. Jeoh T., Michener W., Himmel M. E., Decker S. R., and Adney W. S. (2008) Implications of cellobiohydrolase glycosylation for use in biomass conversion. *Biotechnol. Biofuels* **1**, 10.

5. Voutilainen S., Boer H., and Alapuranen M. (2009) Improving the thermostability and activity of Melanocarpus albomyces cellobiohydrolase Cel7B. *Appl. Microbiol. Biotechnol.* **83**, 261-272.

6. Voutilainen S. P., Murray P. G., Tuohy M. G., and Koivula A. (2010) Expression of Talaromyces emersonii cellobiohydrolase Cel7A in Saccharomyces cerevisiae and rational mutagenesis to improve its thermostability and activity. *Protein Eng. Des. Sel.* **23**, 69-79.

7. Dana C. M., Saija P., Kal S. M., Bryan M. B., Blanch H. W., and Clark D. S. (2012) Biased clique shuffling reveals stabilizing mutations in cellulase Cel7A. *Biotechnol.*

Bioeng. **109**, 2710-2719.

8. Heinzelman P., Komor R., Kanaan A., Romero P. A., Yu X., Mohler S., Snow C., and Arnold F. H. (2010) Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Eng. Des. Sel.* **23**, 871-880.

9. Komor R. S., Romero P. A., Xie C. B., and Arnold F. H. (2012) Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. *Protein Eng. Des. Sel.* **25**, 827-833.

10. Smith M. A., Romero P. A., Wu T., Brustad E. M., and Arnold F. H. (2013) Chimeragenesis of distantly-related proteins by noncontiguous recombination. *Protein Sci.* **22**, 231-238.

11. Otey C. R., Landwehr M., Endelman J. B., Hiraga K., Bloom J. D., and Arnold F. H. (2006) Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol.* **4**, e112.

12. Voigt C. A., Martinez C., Wang Z.-G., Mayo S. L., and Arnold F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553-558.

13. Meyer M., Hochrein L., and Arnold F. H. (2006) Structure-guided SCHEMA recombination of distantly related β-lactamases. *Protein Eng. Des. Sel.* **19**, 563-570.

14. Karypis G., Aggarwal R., Kumar V., and Shekhar S. (1997) Multilevel hypergraph partitioning: application in VLSI domain. In: *Proceedings of the 34th annual Design Automation Conference*, ACM Press, New York, USA.

15. Karypis G., and Kumar V. (2000) Multilevel k-way hypergraph partitioning. *VLSI Des.* **11**, 285-300.

16. Romero P., Stone E., Lamb C., Chantranupong L., Krause A., Miklos A., Hughes R., Fechtel B., Ellington A. D., Arnold F. H., and Georgiou G. (2012) SCHEMA-designed variants of human arginase I and II reveal sequence elements important to stability and catalysis. *ACS Synth. Biol.* **1**, 221-228.

17. Li Y., Drummond D. A., Sawayama A. M., Snow C. D., Bloom J. D., and Arnold F. H. (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25**, 1051-1056.

18. Romero P. A., Krause A., and Arnold F. H. (2013) Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. USA* **110**, E193-E201.

19. Heinzelman P., Snow C. D., Smith M. A., Yu X., Kannan A., Boulware K., Villalobos A., Govindarajan S., Minshull J., and Arnold F. H. (2009) SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *J. Biol. Chem.* **284**, 26229-26233.

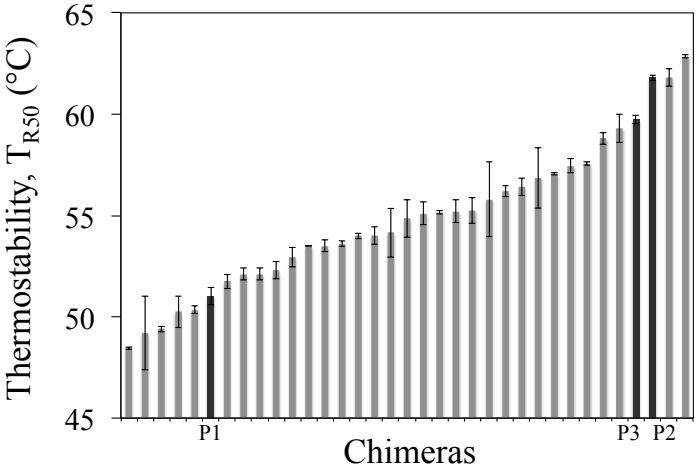
20. Endelman J., Silberg J., Wang Z., and Arnold F. H. (2004) Site-directed protein recombination as a shortest-path problem. *Protein Eng. Des. Sel.* **17**, 589-594.

21. Pei J., Kim B.-H., and Grishin N. V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295-2300.

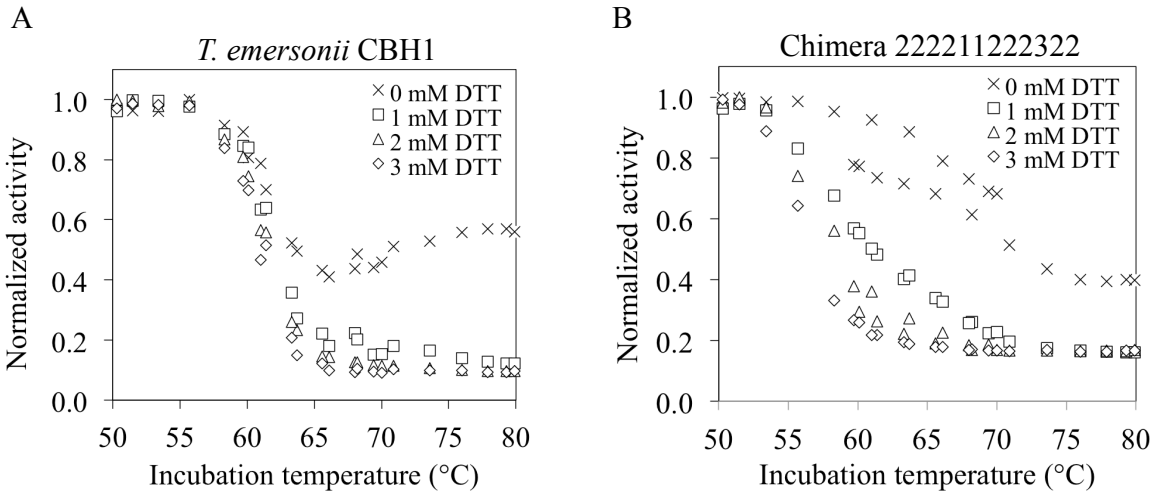
22. Krause A. (2010) SFO: A Toolbox for Submodular Function Optimization. *J. Mach. Learn. Res.* **11**, 1141-1144.

23. Heinzelman P., Snow C. D., Wu I., Nguyen C., Villalobos A., Govindarajan S., Minshull J., and Arnold F. H. (2009) A family of thermostable fungal cellulases created by structure-guided recombination. *Proc. Natl. Acad. Sci. USA* **106**, 5610-5615.

6.8 Supplementary information



Supplementary Figure 6.7: Thermostabilities of a maximally informative subset (32 chimeras) of the library. The T_{R50} is the incubation temperature at which a chimera's residual activity after 10 minutes is half its activity before incubation. Measurements were performed in duplicate. The parental enzymes are highlighted.



Supplementary Figure 6.8: T_{R50} measurements for a range of DTT concentrations. A) Parental cellulase *T. emersonii* CBHI recovers some activity after incubation if there is no DTT present. The enzyme does not recover activity when incubated with 1 mM DTT. B) Chimeric cellulase 222211222322 is able to recover some activity after incubation with 1 mM DTT.

123113322331

QQACTLQSETHPPLTWKRCSSGGTCSTVNGSVTIDANWRWTHTVSGSTNCYDGNTWDT
SLCTDDKSCAQTCCLDGADYSSTYGITTSGDSLNLKFVTQSAQKNVGSRVYLMASDTTYQ
MFELLNREFTFDVDVSNLPCGLNGALYFVSMDADGGMSKYSGNKAGAKYGTGYCDAQC
PRDLKFIDGEANVEGWQPSSNNANTGIGDHGSCCSEMDVWEANSISEAVTPHPCTTVGQT
MCSGDDCGGTYSDNRYGGVCDPDGCDFNPYRMGNTSFYGPGMTVDTTKKMTVVTQFL
TDDGTDTGTLSEIKRFYVQDGKIIAQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRKG
GLAQMSKALAGPMVLVMSVWDDYYAQMLWLDSTYPTNETSSTPGAVRGSCPTDSGVPA
QVESQSPNSNVIFSNIRFGPIGSTVPGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGG
IGYSGPTVCASGTTCQVLNPYYSQCL

211213233112

QSACTLQSETHPPLTWQKCTAPGSCTQQTGSVTIDANWRWVHATNSSTNCYTGNTWSS
TLCPDNETCAKNCCLDGADYSSTYGITTSGNSLNLKFVTGSNVGSRVYLMASDTTYQIFK
LLGNEFSFDVDVSNLPCGLNGALYFVAMDADGGVSKYPTNTAGAKYGTGYCDSQCPRDL
KFINGQANVEGWEPSSNNANTGIGGHGSCCAEMDVWEANSISNAFTPHPCDTPGQEICEG
DGCGGTYSDNRYGGTCDPDGCDFNPYRQGNTSFYGPGSSFTLDTTKKLTVVTQFHKNSA
GVLSEIKRYYVQNGVTFQQPNSKIPGNPGNSITQEYCDAQEVAFGGSSFSDHGGMAKMGA
AMQQGMVLVMSVWDDYAANMLWLDSDYPTDADPTTPGIARGTCSTSSGVPSDVESQSP
NSNVIFSNIKFGPIGSTGNPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGY
SGPTVCASGTTCQVLNPYYSQCL

232232332132

QQACSLTAENHPSLTWQKCTAPGSCTQQTGAVVLDANWRWVHDVNGYTNCYTGNTW
DPTYCPDGETCAQNCCLDGADYEGTYGVTTSGNSLKLNFVTGSNVGSRLYLMENDTKYQ
MFELLGNEFSFDVDVSNLPCGLNGALYFVSMDADGGVSKYPNNKAGAKYGTGYCDAQC
PRDLKFINGEANVGNWTPSTNNANTGIGRYGSCCSEMDVWEANSISNAVTPHPCTTVGQ
SRCEADTCGGTYSSDRYAGVCDPDGCDFNPYRMGNTSFYGPGMTVDTTKKMTVVTQF
HKNSAGVLSEIKRFYVQDGKIIAQPNSDISGVTGNSITTEFCTAQEQAFGGSSFSDKGGLAQ
MSKALAGPMVLVMSLWDDYAANMLWLDSDYPTDADPTTPGIARGTCPTTSGVPSDVES
QSPNSYVTYSNIKFGPIGSTVPPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCG
GIGYSGPTVCASGTTCQVLNPYYSQCL

322321222212

QQACTATAENHPPLTWQECTSGGSCTTQNGAVVIDANWRWTHDVNGYTNCYTGNQWD
PTYCPDDETCAQNCCVDGAAYASTYGVTSSGSSLSIGFVTKHQYGTNVGARLYLLQDDST
YQIFKLLNREFSFDVDVSQLPCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQ
CPRDLKFIDGEANVEGWQPSSNDANAGFGDHGSCCAEMDVWEANSISNAVTPHPCDTPG
QTMCSGDDCGGTYSNDRYAGTCDPDGCDFNPYRMGNTSFYGPGSSFTLDTTKKLTVVT
QFLTDDGTDTGTLSEIKRYYVQNGVTFQQPNSDISGVTGNSITTEFCTAQKQAFGDTDDF
SQHGGLAKMGAAMQQGMVLVMSLWDDHYAQMLWLDSTYPIDQAGAPGAERGTCPTDS
GVPAEIEAQSPNAKVTFSNIKFGPIGSTGNASPPGGNRGTTTTRRPATTTGSSPGPTQSHY
GQCGGIGYSGPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases. The *H. jecorina* CBHI linker and cellulose binding domain are attached to the C-terminus. The chimera nomenclature is a series of numbers, each representing a parent for the blocks A-L. For example, chimera 123113322331 has parent 1's sequence for block A, parent 2 for block B, etc. The three chimeras with no detectable activity are highlighted with an asterisk.

331212211233 *

QSACTLQSETHPPLTWQECTAPGSCTTQNGSVVLDANWRWVHATNSSTNCYTGNQWSS
TLCPDGETCAKNCCLDGADYEGTYGVTTSGSSLKLNFVTGSNVGSRLYLMASDTTYQIFK
LLGNEFSFDVDVSNLGCGLNGALYFVAMDADGGVSKYPTNTAGAKYGTGYCDSQCPRD
LKFINGEANVGNWTPSTNDANAGFGRYGSCCAEMDIWEANNMATALTPHPCDTPGQSR
CEADTCGGTYSDNRYGGTCDPDGCDWNAYRLGDKTFYGKGMTVDTNKKMTVVTQFE
TSGAINRFYVQDGKIIANAEAELGSYSGNELNDDYCTAEKAEFGDTDDFSQHGGLAKMGA
AMQQGMVLVMSLWDDHYANMLWLDSTYPIDQAGAPGAERGACPTTSGVPSDVESQVPN
SYVTYSNIKFGPIGSTVPASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYSGPT
VCASGTTCQVLNPYYSQCL

311112222333

QSACTLQSETHPPLTWKRCSSGGTCSTVNGSVVLDANWRWTHATNSSTNCYTGNQWSS
TLCPDNETCAKNCCLDGADYEGTYGVTTSGDSLKLNFVTQSAQKNVGSRLYLMASDTTY
QIFKLLNREFTFDVDVSNLGCGLNGALYFVAMDADGGVSKYPTNTAGAKYGTGYCDSQC
PRDLKFINGQANVEGWEPSSNDANAGFGGHGSCCAEMDVWEANNMATAVTPHPCDTPG
QEICEGDGCGGTYSDNRYGGTCDPDGCDFNAYRMGDKTFYGKGMTVDTNKKMTVVTQ
FLTDDGTDTGTLSEIKRFYVQDGKIIANAESDISGVTGNSITTEFCTAQKQAFSNTDDFNR
HGGLAKMGAAMQQGMVLVMSLWDDHYANMLWLDSTYPIDQAGAPGAERGACSTSSGV
PAQVESQVPNSYVTYSNIRFGPIGSTVPGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQ
CGGIGYSGPTVCASGTTCQVLNPYYSQCL

331211323222

QSACTLQSETHPPLTWQECTAPGSCTTQNGSVVIDANWRWVHATNSSTNCYTGNQWSS
TLCPDGETCAKNCCLDGAAYASTYGVTTSGSSLSIGFVTGSNVGARLYLMASDTTYQMFE
LLNREFSFDVDVSQLPCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDAQCPRD
LKFINGEANVGNWTPSTNDANAGFGRYGSCCSEMDVWEANSISNAFTPHPCTTVGQSRC
EADTCGGTYSDNRYGGVCDPDGCDFNPYRQGNTSFYGPGKIIDTTKPFTVVTQFLTDDG
TDTGTLSEIKRFYIQNSNVIPQPNSKIPGNPGNSITQEYCDAQKVAFGDTDDFSQKGGMAQ
MSKALAGPMVLVMSLWDDHYANMLWLDSTYPIDQAGAPGAERGTCPTTSGVPSDVESQ
SPNAKVTFSNIKFGPINSTFTASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYS
GPTVCASGTTCQVLNPYYSQCL

211331132321

QSACSLTAENHPSLTWKRCTSGGSCSTVNGAVVIDANWRWTHATNSSTNCYTGNTWSST
LCPDNETCAKNCCVDGAAYASTYGVTTSGDSLSIGFVTKHQYGTNVGARLYLMENDTKY
QEFTLLGNEFTFDVDVSQLPCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDSQC
PRDLKFINGQANVEGWEPSSNNANTGIGGHGSCCSEMDVWEANSISEAVTPHPCTTVGQE
ICEGDGCGGTYSSDRYAGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQFHKNS
AGVLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRKGGLTQFKK
ATSGGMVLVMSLWDDYAANMLWLDSDYPTDADPTTPGIARGSCSTSSGVPAEIEAQSPN
AKVTFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYSGPT
VCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases (continued).

212322323311

QQACTATAENHPPLTWKRCTSGGSCSTVNGAVVLDANWRWTHDVNGYTNCYTGNTW
DPTYCPDNETCAQNCCVDGADYEGTYGVTSSGDSLKLNFVTKHQYGTNVGSRLYLLQDD
STYQMFELLNREFTFDVDVSNLPCGLNGALYFVSMDADGGVSKYPNNKAGAKYGTGYC
DAQCPRDLKFINGQANVEGWEPSSNNANTGIGGHGSCCSEMDVWEANSISEAFTPHPCTT
VGQEICEGDGCGGTYSNDRYAGVCDPDGCDFNPYRQGNTSFYGPGSSFTLDTTKKLTVV
TQFLTDDGTDTGTLSEIKRYYVQNGVTFQQPNSKIPGNPGNSITQEYCDAQKVAFSNTDD
FNRKGGMAQMSKALAGPMVLVMSLWDDYAANMLWLDSDYPTDADPTTPGIARGSCSTS
SGVPAEIEAQSPNSYVTYSNIRFGPIGSTGNGPPGGNRGTTTTRRPATTTGSSPGPTQSHY
GQCGGIGYSGPTVCASGTTCQVLNPYYSQCL

323212111312

QQACTLQSETHPPLTWKRCTAPGSCSTVNGSVVLDANWRWVHTVSGSTNCYTGNQWD
TSLCTDDKSCAQTCCLDGADYEGTYGVTTSGDSLKLNFVTGSNVGSRLYLMASDTTYQE
FTLLGNEFTFDVDVSNLPCGLNGALYFVSMDADGGMSKYSGNKAGAKYGTGYCDSQCP
RDLKFIDGEANVEGWQPSSNDANAGFGDHGSCCSEMDIWEANSISNALTPHPCTTVGQT
MCSGDDCGGTYSDNRYGGTCDPDGCDWNPYRLGNTSFYGPGSSFTLDTTKKLTVVTQF
ETSGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEKAEFSNTDDFNRKGGLTQF
KKATSGGMVLVMSLWDDHYAQMLWLDSTYPIDQAGAPGAERGTCPTDSGVPSDVESQS
PNSYVTYSNIRFGPIGSTGNGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYSG
PTVCASGTTCQVLNPYYSQCL

312223311121

QQACTATAENHPPLTWQKCTAPGSCTQQTGAVTIDANWRWVHDVNGYTNCYTGNQW
DPTYCPDNETCAQNCCLDGADYSSTYGITSSGNSLNLKFVTGSNVGSRVYLLQDDSTYQM
FELLGNEFSFDVDVSNLPCGLNGALYFVSMDADGGVSKYPNNKAGAKYGTGYCDAQCPR
DLKFINGQANVEGWEPSSNDANAGFGGHGSCCSEMDIWEANSISEALTPHPCTTVGQEIC
EGDGCGGTYSNDRYAGVCDPDGCDWNPYRLGNTSFYGPGKIIDTTKPFTVVTQFETSGA
INRFYIQNSNVIPQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGLAQMSKALAGPM
VLVMSVWDDHYANMLWLDSTYPIDQAGAPGAERGSCSTSSGVPSDVESQSPNSNVIFSNIK
FGPINSTFTPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYSGPTVCASG
TTCQVLNPYYSQCL

221221333333

QSACTATAENHPPLTWKRCTAPGSCSTVNGAVVIDANWRWVHATNSSTNCYTGNTWSS
TLCPDDETCAKNCCLDGAAYASTYGVTSSGDSLSIGFVTGSNVGARLYLLQDDSTYQMFE
LLGNEFTFDVDVSQLGCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDAQCPRD
LKFIDGEANVEGWQPSSNNANTGIGDHGSCCSEMDVWEANNMATAFTPHPCTTVGQTM
CSGDDCGGTYSNDRYAGVCDPDGCDFNAYRQGDKTFYGKGMTVDTNKKMTVVTQFHK
NSAGVLSEIKRFYVQDGKIIANAESKIPGNPGNSITQEYCDAQKVAFSNTDDFNRKGGMAQ
MSKALAGPMVLVMSLWDDYAAQMLWLDSDYPTDADPTTPGIARGACPTDSGVPSDVES
QVPNAKVTFSNIRFGPIGSTVPGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGY
SGPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases (continued).

322132333121

QQACSLTAENHPSLTWQKCSSGGTCTQQTGAVVLDANWRWTHDVNGYTNCYTGNQWD
PTYCPDDETCAQNCCLDGADYEGTYGVTTSGNSLKLNFVTQSAQKNVGSRLYLMENDTK
YQMFELLGNEFSFDVDVSNLPCGLNGALYFVSMDADGGVSKYPNNKAGAKYGTGYCDA
QCPRDLKFIDGEANVEGWQPSSNDANAGFGDHGSCCSEMDVWEANSISEAFTPHPCTTV
GQTMCSGDDCGGTYSSDRYAGVCDPDGCDFNPYRQGNTSFYGPGKIIDTTKPFTVVTQF
HKNSAGVLSEIKRFYIQNSNVIPQPNSKIPGNPGNSITQEYCDAQEVAFGGSSFSDKGGMA
QMSKALAGPMVLVMSLWDDHYAQMLWLDSTYPIDQAGAPGAERGSCPTDSGVPAQVES
QSPNSYVTYSNIKFGPINSTFTPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCG
GIGYSGPTVCASGTTCQVLNPYYSQCL

231123332212

QSACTATAENHPPLTWQECSSGGTCTTQNGAVTIDANWRWTHATNSSTNCYTGNTWSS
TLCPDGETCAKNCCLDGADYSSTYGITSSGSSLNLKFVTQSAQKNVGSRVYLLQDDSTYQ
MFELLGNEFSFDVDVSNLPCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDAQC
PRDLKFINGEANVGNWTPSTNNANTGIGRYGSCCSEMDVWEANSISNAVTPHPCTTVGQ
SRCEADTCGGTYSNDRYAGVCDPDGCDFNPYRMGNTSFYGPGSSFTLDTTKKLTVVTQF
HKNSAGVLSEIKRYYVQNGVTFQQPNSDISGVTGNSITTEFCTAQKQAFGDTDDFSQKGG
LAQMSKALAGPMVLVMSVWDDYAANMLWLDSDYPTDADPTTPGIARGTCPTTSGVPAQ
VESQSPNSNVIFSNIKFGPIGSTGNASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGI
GYSGPTVCASGTTCQVLNPYYSQCL

121233122211

QSACSLTAENHPSLTWQECTAPGSCTTQNGAVTIDANWRWVHATNSSTNCYDGNTWSS
TLCPDDETCAKNCCLDGADYSSTYGITTSGSSLNLKFVTGSNVGSRVYLMENDTKYQEFT
LLNREFSFDVDVSNLPCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDSQCPRDL
KFIDGEANVEGWQPSSNNANTGIGDHGSCCSEMDVWEANSISEAVTPHPCTTVGQTMCS
GDDCGGTYSSDRYAGTCDPDGCDFNPYRMGNTSFYGPGSSFTLDTTKKLTVVTQFLTDD
GTDTGTLSEIKRYYVQNGVTFQQPNSDISGVTGNSITTEFCTAQKQAFGDTDDFSQKGGL
TQFKKATSGGMVLVMSVWDDYYAQMLWLDSTYPTNETSSTPGAVRGSCPTDSGVPSDV
ESQSPNSNVIFSNIKFGPIGSTGNASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIG
YSGPTVCASGTTCQVLNPYYSQCL

221332311112

QSACSLTAENHPSLTWQKCTSGGSCTQQTGAVVLDANWRWTHATNSSTNCYTGNTWSS
TLCPDDETCAKNCCVDGADYEGTYGVTTSGNSLKLNFVTKHQYGTNVGSRLYLMENDT
KYQMFELLGNEFSFDVDVSNLPCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCD
AQCPRDLKFIDGEANVEGWQPSSNNANTGIGDHGSCCSEMDIWEANSISNALTPHPCTTV
GQTMCSGDDCGGTYSSDRYAGVCDPDGCDWNPYRLGNTSFYGPGSSFTLDTTKKLTVV
TQFETSGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGLAQ
MSKALAGPMVLVMSLWDDYAAQMLWLDSDYPTDADPTTPGIARGTCPTDSGVPAEIEA
QSPNSYVTYSNIKFGPIGSTGNPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCG
GIGYSGPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases (continued).

212312332233

QQACTLQSETHPPLTWQECTSGGSCTTQNGSVVLDANWRWTHDVNGYTNCYTGNTWD
PTYCPDNETCAQNCCVDGADYEGTYGVTTSGSSLKLNFVTKHQYGTNVGSRLYLMASDT
TYQMFELLGNEFSFDVDVSNLGCGLNGALYFVSMDADGGVSKYPNNKAGAKYGTGYCD
AQCPRDLKFINGQANVEGWEPSSNNANTGIGGHGSCCSEMDVWEANNMATAVTPHPCT
TVGQEICEGDGCGGTYSDNRYGGVCDPDGCDFNAYRMGDKTFYGKGMTVDTNKKMTV
VTQFHKNSAGVLSEIKRFYVQDGKIIANAESDISGVTGNSITTEFCTAQKQAFGDTDDFSQ
KGGLAQMSKALAGPMVLVMSLWDDYAANMLWLDSDYPTDADPTTPGIARGACSTSSGV
PAEIEAQVPNSYVTYSNIKFGPIGSTVPASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQ
CGGIGYSGPTVCASGTTCQVLNPYYSQCL

313322332122

QQACTATAENHPPLTWQKCTSGGSCTQQTGAVVLDANWRWTHTVSGSTNCYTGNQWD
TSLCTDNKSCAQTCCVDGADYEGTYGVTSSGNSLKLNFVTKHQYGTNVGSRLYLLQDDS
TYQMFELLGNEFSFDVDVSNLPCGLNGALYFVSMDADGGMSKYSGNKAGAKYGTGYCD
AQCPRDLKFINGQANVEGWEPSSNDANAGFGGHGSCCSEMDVWEANSISNAVTPHPCTT
VGQEICEGDGCGGTYSNDRYAGVCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQ
FHKNSAGVLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQEQAFGGSSFSDKGGLA
QMSKALAGPMVLVMSLWDDHYANMLWLDSTYPIDQAGAPGAERGTCSTSSGVPAEIEAQ
SPNSYVTYSNIKFGPINSTFTPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGI
GYSGPTVCASGTTCQVLNPYYSQCL

213131232211

QQACSLTAENHPSLTWQECSSGGTCTTQNGAVVIDANWRWTHTVSGSTNCYTGNTWDT
SLCTDNKSCAQTCCLDGAAYASTYGVTTSGSSLSIGFVTQSAQKNVGARLYLMENDTKYQ
IFKLLGNEFSFDVDVSQLPCGLNGALYFVAMDADGGMSKYSGNKAGAKYGTGYCDSQCP
RDLKFINGQANVEGWEPSSNNANTGIGGHGSCCAEMDVWEANSISEAVTPHPCDTPGQEI
CEGDGCGGTYSSDRYAGTCDPDGCDFNPYRMGNTSFYGPGSSFTLDTTKKLTVVTQFHK
NSAGVLSEIKRYYVQNGVTFQQPNSDISGVTGNSITTEFCTAQKQAFGDTDDFSQHGGLA
KMGAAMQQGMVLVMSLWDDYAANMLWLDSDYPTDADPTTPGIARGSCSTSSGVPAQV
ESQSPNAKVTFSNIKFGPIGSTGNASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGI
GYSGPTVCASGTTCQVLNPYYSQCL

131232232321

QSACSLTAENHPSLTWKRCTAPGSCSTVNGAVVLDANWRWVHATNSSTNCYDGNTWSS
TLCPDGETCAKNCCLDGADYEGTYGVTTSGDSLKLNFVTGSNVGSRLYLMENDTKYQIF
KLLGNEFTFDVDVSNLPCGLNGALYFVAMDADGGVSKYPTNTAGAKYGTGYCDSQCPR
DLKFINGEANVGNWTPSTNNANTGIGRYGSCCAEMDVWEANSISEAVTPHPCDTPGQSR
CEADTCGGTYSSDRYAGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQFHKNS
AGVLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRHGGLAKMG
AAMQQGMVLVMSLWDDYYANMLWLDSTYPTNETSSTPGAVRGSCPTTSGVPSDVESQS
PNSYVTYSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYSGP
TVCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases (continued).

112131322322

QQACSLTAENHPSLTWKRCSSGGTCSTVNGAVVIDANWRWTHDVNGYTNCYDGNTWD
PTYCPDNETCAQNCCLDGAAYASTYGVTTSGDSLSIGFVTQSAQKNVGARLYLMENDTK
YQMFELLNREFTFDVDVSQLPCGLNGALYFVSMDADGGVSKYPNNKAGAKYGTGYCDA
QCPRDLKFINGQANVEGWEPSSNNANTGIGGHGSCCSEMDVWEANSISNAVTPHPCTTV
GQEICEGDGCGGTYSSDRYAGVCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQF
LTDDGTDTGTLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRKG
GLAQMSKALAGPMVLVMSLWDDYYANMLWLDSTYPTNETSSTPGAVRGTCSTSSGVPA
QVESQSPNAKVTFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCG
GIGYSGPTVCASGTTCQVLNPYYSQCL

111122332232

QSACTATAENHPPLTWQECSSGGTCTTQNGAVVLDANWRWTHATNSSTNCYDGNTWSS
TLCPDNETCAKNCCLDGADYEGTYGVTSSGSSLKLNFVTQSAQKNVGSRLYLLQDDSTYQ
MFELLGNEFSFDVDVSNLPCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDAQC
PRDLKFINGQANVEGWEPSSNNANTGIGGHGSCCSEMDVWEANSISNAVTPHPCTTVGQ
EICEGDGCGGTYSNDRYAGVCDPDGCDFNPYRMGNTSFYGPGMTVDTTKKMTVVTQF
HKNSAGVLSEIKRFYVQDGKIIAQPNSDISGVTGNSITTEFCTAQKQAFGDTDDFSQKGGL
AQMSKALAGPMVLVMSLWDDYYANMLWLDSTYPTNETSSTPGAVRGTCSTSSGVPAQV
ESQSPNSYVTYSNIKFGPIGSTVPASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGI
GYSGPTVCASGTTCQVLNPYYSQCL

311211331111

QSACTLQSETHPPLTWQKCTAPGSCTQQTGSVVIDANWRWVHATNSSTNCYTGNQWSS
TLCPDNETCAKNCCLDGAAYASTYGVTTSGNSLSIGFVTGSNVGARLYLMASDTTYQMF
ELLGNEFSFDVDVSQLPCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDAQCPRD
LKFINGQANVEGWEPSSNDANAGFGGHGSCCSEMDIWEANSISEALTPHPCTTVGQEICE
GDGCGGTYSDNRYGGVCDPDGCDWNPYRLGNTSFYGPGSSFTLDTTKKLTVVTQFHKN
SAGVLSEIKRYYVQNGVTFQQPNAELGSYSGNSLNDDYCTAEEAEFGGSSFSDKGGLAQM
SKALAGPMVLVMSLWDDHYANMLWLDSTYPIDQAGAPGAERGSCSTSSGVPSDVESQSP
NAKVTFSNIKFGPIGSTGNPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIG
YSGPTVCASGTTCQVLNPYYSQCL

122122133333 *

QQACTATAENHPPLTWKRCSSGGTCSTVNGAVVLDANWRWTHDVNGYTNCYDGNTW
DPTYCPDDETCAQNCCLDGADYEGTYGVTSSGDSLKLNFVTQSAQKNVGSRLYLLQDDS
TYQEFTLLGNEFTFDVDVSNLGCGLNGALYFVSMDADGGVSKYPNNKAGAKYGTGYCD
SQCPRDLKFIDGEANVEGWQPSSNNANTGIGDHGSCCSEMDVWEANNMATAFTPHPCTT
VGQTMCSGDDCGGTYSNDRYAGTCDPDGCDFNAYRQGDKTFYGKGMTVDTNKKMTV
VTQFHKNSAGVLSEIKRFYVQDGKIIANAESKIPGNPGNSITQEYCDAQKVAFSNTDDFNR
KGGMTQFKKATSGGMVLVMSLWDDYYAQMLWLDSTYPTNETSSTPGAVRGACPTDSG
VPAQVESQVPNSYVTYSNIRFGPIGSTVPGPPGGNRGTTTTRRPATTTGSSPGPTQSHYG
QCGGIGYSGPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases (continued).

233122223223

QQACTATAENHPPLTWQECSSGGTCTTQNGAVVLDANWRWTHTVSGSTNCYTGNTWD
TSLCTDGKSCAQTCCLDGADYEGTYGVTSSGSSLKLNFVTQSAQKNVGSRLYLLQDDSTY
QIFKLLNREFSFDVDVSNLGCGLNGALYFVAMDADGGMSKYSGNKAGAKYGTGYCDSQC
PRDLKFINGEANVGNWTPSTNNANTGIGRYGSCCAEMDVWEANNMATAFTPHPCDTPG
QSRCEADTCGGTYSNDRYAGTCDPDGCDFNAYRQGDKTFYGKGKIIDTNKPFTVVTQFL
TDDGTDTGTLSEIKRFYIQNSNVIPNAESKIPGNPGNSITQEYCDAQKVAFGDTDDFSQHG
GMAKMGAAMQQGMVLVMSLWDDYAANMLWLDSDYPTDADPTTPGIARGACPTTSGV
PAQVESQVPNSYVTYSNIKFGPINSTFTASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQ
CGGIGYSGPTVCASGTTCQVLNPYYSQCL

333222211111

QQACTATAENHPPLTWQKCTAPGSCTQQTGAVVLDANWRWVHTVSGSTNCYTGNQW
DTSLCTDGKSCAQTCCLDGADYEGTYGVTSSGNSLKLNFVTGSNVGSRLYLLQDDSTYQI
FKLLGNEFSFDVDVSNLPCGLNGALYFVAMDADGGMSKYSGNKAGAKYGTGYCDSQCP
RDLKFINGEANVGNWTPSTNDANAGFGRYGSCCAEMDIWEANSISEALTPHPCDTPGQSR
CEADTCGGTYSNDRYAGTCDPDGCDWNPYRLGNTSFYGPGSSFTLDTTKKLTVVTQFE
TSGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDHGGLAKMGA
AMQQGMVLVMSLWDDHYANMLWLDSTYPIDQAGAPGAERGSCPTTSGVPSDVESQSPN
SYVTYSNIKFGPIGSTGNPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYS
GPTVCASGTTCQVLNPYYSQCL

231333111132

QSACSLTAENHPSLTWQKCTSGGSCTQQTGAVTIDANWRWTHATNSSTNCYTGNTWSST
LCPDGETCAKNCCVDGADYSSTYGITTSGNSLNLKFVTKHQYGTNVGSRVYLMENDTKY
QEFTLLGNEFSFDVDVSNLPCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDSQC
PRDLKFINGEANVGNWTPSTNNANTGIGRYGSCCSEMDIWEANSISNALTPHPCTTVGQS
RCEADTCGGTYSSDRYAGTCDPDGCDWNPYRLGNTSFYGPGMTVDTTKKMTVVTQFE
TSGAINRFYVQDGKIIAQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGLTQFKKAT
SGGMVLVMSVWDDYAANMLWLDSDYPTDADPTTPGIARGTCPTTSGVPAEIEAQSPNSN
VIFSNIKFGPIGSTVPPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYSGP
TVCASGTTCQVLNPYYSQCL

233232323133

QQACSLTAENHPSLTWQKCTAPGSCTQQTGAVVLDANWRWVHTVSGSTNCYTGNTWD
TSLCTDGKSCAQTCCLDGADYEGTYGVTTSGNSLKLNFVTGSNVGSRLYLMENDTKYQM
FELLNREFSFDVDVSNLGCGLNGALYFVSMDADGGMSKYSGNKAGAKYGTGYCDAQCP
RDLKFINGEANVGNWTPSTNNANTGIGRYGSCCSEMDVWEANNMATAFTPHPCTTVGQ
SRCEADTCGGTYSSDRYAGVCDPDGCDFNAYRQGDKTFYGKGMTVDTNKKMTVVTQF
LTDDGTDTGTLSEIKRFYVQDGKIIANAESKIPGNPGNSITQEYCDAQEVAFGGSSFSDKG
GMAQMSKALAGPMVLVMSLWDDYAANMLWLDSDYPTDADPTTPGIARGACPTTSGVP
SDVESQVPNSYVTYSNIKFGPIGSTVPPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHY
GQCGGIGYSGPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases (continued).

333112122312 *

QQACTLQSETHPPLTWKRCSSGGTCSTVNGSVVLDANWRWTHTVSGSTNCYTGNQWDT
SLCTDGKSCAQTCCLDGADYEGTYGVTTSGDSLKLNFVTQSAQKNVGSRLYLMASDTTY
QEFTLLNREFTFDVDVSNLPCGLNGALYFVSMDADGGMSKYSGNKAGAKYGTGYCDSQ
CPRDLKFINGEANVGNWTPSTNDANAGFGRYGSCCSEMDVWEANSISNAVTPHPCTTVG
QSRCEADTCGGTYSDNRYGGTCDPDGCDFNPYRMGNTSFYGPGSSFTLDTTKKLTVVT
QFLTDDGTDTGTLSEIKRYYVQNGVTFQQPNSDISGVTGNSITTEFCTAQKQAFSNTDDF
NRKGGLTQFKKATSGGMVLVMSLWDDHYANMLWLDSTYPIDQAGAPGAERGTCPTTSG
VPAQVESQSPNSYVTYSNIRFGPIGSTGNGPPGGNRGTTTTRRPATTTGSSPGPTQSHYG
QCGGIGYSGPTVCASGTTCQVLNPYYSQCL

133331223212

QQACSLTAENHPSLTWQECTSGGSCTTQNGAVVIDANWRWTHTVSGSTNCYDGNTWDT
SLCTDGKSCAQTCCVDGAAYASTYGVTTSGSSLSIGFVTKHQYGTNVGARLYLMENDTK
YQIFKLLNREFSFDVDVSQLPCGLNGALYFVAMDADGGMSKYSGNKAGAKYGTGYCDSQ
CPRDLKFINGEANVGNWTPSTNNANTGIGRYGSCCAEMDVWEANSISNAFTPHPCDTPG
QSRCEADTCGGTYSSDRYAGTCDPDGCDFNPYRQGNTSFYGPGSSFTLDTTKKLTVVTQ
FLTDDGTDTGTLSEIKRYYVQNGVTFQQPNSKIPGNPGNSITQEYCDAQKVAFGDTDDFS
QHGGMAKMGAAMQQGMVLVMSLWDDYYANMLWLDSTYPTNETSSTPGAVRGTCPTT
SGVPAEIEAQSPNAKVTFSNIKFGPIGSTGNASPPGGNRGTTTTRRPATTTGSSPGPTQSH
YGQCGGIGYSGPTVCASGTTCQVLNPYYSQCL

131112222121

QSACTLQSETHPPLTWQKCSSGGTCTQQTGSVVLDANWRWTHATNSSTNCYDGNTWSS
TLCPDGETCAKNCCLDGADYEGTYGVTTSGNSLKLNFVTQSAQKNVGSRLYLMASDTTY
QIFKLLNREFSFDVDVSNLPCGLNGALYFVAMDADGGVSKYPTNTAGAKYGTGYCDSQC
PRDLKFINGEANVGNWTPSTNNANTGIGRYGSCCAEMDVWEANSISEAVTPHPCDTPGQ
SRCEADTCGGTYSDNRYGGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQFLT
DDGTDTGTLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQEQAFGGSSFSDHGGLA
KMGAAMQQGMVLVMSLWDDYYANMLWLDSTYPTNETSSTPGAVRGSCPTTSGVPAQV
ESQSPNSYVTYSNIKFGPINSTFTPSGGNPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQC
GGIGYSGPTVCASGTTCQVLNPYYSQCL

322211223233

QQACTLQSETHPPLTWQECTAPGSCTTQNGSVVIDANWRWVHDVNGYTNCYTGNQWD
PTYCPDDETCAQNCCLDGAAYASTYGVTTSGSSLSIGFVTGSNVGARLYLMASDTTYQIF
KLLNREFSFDVDVSQLGCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQCPR
DLKFIDGEANVEGWQPSSNDANAGFGDHGSCCAEMDVWEANNMATAFTPHPCDTPGQT
MCSGDDCGGTYSDNRYGGTCDPDGCDFNAYRQGDKTFYGKGMTVDTNKKMTVVTQFL
TDDGTDTGTLSEIKRFYVQDGKIIANAESKIPGNPGNSITQEYCDAQKVAFGDTDDFSQHG
GMAKMGAAMQQGMVLVMSLWDDHYAQMLWLDSTYPIDQAGAPGAERGACPTDSGVPS
DVESQVPNAKVTFSNIKFGPIGSTVPASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQC
GGIGYSGPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases (continued).

223311222322

QQACTLQSETHPPLTWKRCTSGGSCSTVNGSVVIDANWRWTHTVSGSTNCYTGNTWDT
SLCTDDKSCAQTCCVDGAAYASTYGVTTSGDSLSIGFVTKHQYGTNVGARLYLMASDTT
YQIFKLLNREFTFDVDVSQLPCGLNGALYFVAMDADGGMSKYSGNKAGAKYGTGYCDS
QCPRDLKFIDGEANVEGWQPSSNNANTGIGDHGSCCAEMDVWEANSISNAVTPHPCDTP
GQTMCSGDDCGGTYSDNRYGGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQ
FLTDDGTDTGTLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRH
GGLAKMGAAMQQGMVLVMSLWDDYAAQMLWLDSDYPTDADPTTPGIARGTCPTDSGV
PAEIEAQSPNAKVTFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQC
GGIGYSGPTVCASGTTCQVLNPYYSQCL

121213333323

QSACTLQSETHPPLTWKRCTAPGSCSTVNGSVTIDANWRWVHATNSSTNCYDGNTWSST
LCPDDETCAKNCCLDGADYSSTYGITTSGDSLNLKFVTGSNVGSRVYLMASDTTYQMFEL
LGNEFTFDVDVSNLGCGLNGALYFVSMDADGGVSKYPTNTAGAKYGTGYCDAQCPRDL
KFIDGEANVEGWQPSSNNANTGIGDHGSCCSEMDVWEANNMATAFTPHPCTTVGQTMC
SGDDCGGTYSDNRYGGVCDPDGCDFNAYRQGDKTFYGKGKIIDTNKPFTVVTQFHKNS
AGVLSEIKRFYIQNSNVIPNAESKIPGNPGNSITQEYCDAQKVAFSNTDDFNRKGGMAQMS
KALAGPMVLVMSVWDDYYAQMLWLDSTYPTNETSSTPGAVRGACPTDSGVPSDVESQV
PNSNVIFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYSGP
TVCASGTTCQVLNPYYSQCL

123333232233

QQACSLTAENHPSLTWQECTSGGSCTTQNGAVTIDANWRWTHTVSGSTNCYDGNTWDT
SLCTDDKSCAQTCCVDGADYSSTYGITTSGSSLNLKFVTKHQYGTNVGSRVYLMENDTKY
QIFKLLGNEFSFDVDVSNLGCGLNGALYFVAMDADGGMSKYSGNKAGAKYGTGYCDSQC
PRDLKFIDGEANVEGWQPSSNNANTGIGDHGSCCAEMDVWEANNMATAVTPHPCDTPG
QTMCSGDDCGGTYSSDRYAGTCDPDGCDFNAYRMGDKTFYGKGMTVDTNKKMTVVT
QFHKNSAGVLSEIKRFYVQDGKIIANAESDISGVTGNSITTEFCTAQKQAFGDTDDFSQHG
GLAKMGAAMQQGMVLVMSVWDDYYAQMLWLDSTYPTNETSSTPGAVRGACPTDSGVP
AEIEAQVPNSNVIFSNIKFGPIGSTVPASPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQC
GGIGYSGPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.1: Amino acid sequences of the maximally informative subset of 35 chimeric cellulases (continued).

222233221322

QQACSLTAENHPSLTWKRCTAPGSCSTVNGAVTIDANWRWVHDVNGYTNCYTGNTWD
PTYCPDDETCAQNCCLDGADYSSTYGITTSGDSLNLKFVTGSNVGSRVYLMENDTKYQIF
KLLNREFTFDVDVSNLPCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQCPR
DLKFIDGEANVEGWQPSSNNANTGIGDHGSCCAEMDIWEANSISNALTPHPCDTPGQTMC
SGDDCGGTYSSDRYAGTCDPDGCDWNPYRLGNTSFYGPGKIIDTTKPFTVVTQFLTDDG
TDTGTLSEIKRFYIQNSNVIPQPNAELGSYSGNSLNDDYCTAEKAEFSNTDDFNRHGGLAK
MGAAMQQGMVLVMSVWDDYAAQMLWLDSDYPTDADPTTPGIARGTCPTDSGVPSDVE
SQSPNSNVIFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYS
GPTVCASGTTCQVLNPYYSQCL

232233222322

QQACSLTAENHPSLTWKRCTAPGSCSTVNGAVTIDANWRWVHDVNGYTNCYTGNTWD
PTYCPDGETCAQNCCLDGADYSSTYGITTSGDSLNLKFVTGSNVGSRVYLMENDTKYQIF
KLLNREFTFDVDVSNLPCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQCPR
DLKFINGEANVGNWTPSTNNANTGIGRYGSCCAEMDVWEANSISNAVTPHPCDTPGQSR
CEADTCGGTYSSDRYAGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQFLTDD
GTDTGTLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRHGGLAK
MGAAMQQGMVLVMSVWDDYAANMLWLDSDYPTDADPTTPGIARGTCPTTSGVPSDVE
SQSPNSNVIFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYS
GPTVCASGTTCQVLNPYYSQCL

222211222322

QQACTLQSETHPPLTWKRCTAPGSCSTVNGSVVIDANWRWVHDVNGYTNCYTGNTWD
PTYCPDDETCAQNCCLDGAAYASTYGVTTSGDSLSIGFVTGSNVGARLYLMASDTTYQIF
KLLNREFTFDVDVSQLPCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQCPR
DLKFIDGEANVEGWQPSSNNANTGIGDHGSCCAEMDVWEANSISNAVTPHPCDTPGQTM
CSGDDCGGTYSDNRYGGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQFLTDD
GTDTGTLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRHGGLAK
MGAAMQQGMVLVMSLWDDYAAQMLWLDSDYPTDADPTTPGIARGTCPTDSGVPSDVE
SQSPNAKVTFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGY
SGPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.2: Amino acid sequences of 7 chimeric cellulases predicted to be stable. The *H. jecorina* CBHI linker and cellulose binding domain are attached to the C-terminus. The chimera nomenclature is a series of numbers, each representing a parent for the blocks A-L. For example, chimera 123113322331 has parent 1's sequence for block A, parent 2 for block B, etc.

222223222322

QQACTATAENHPPLTWKRCTAPGSCSTVNGAVTIDANWRWVHDVNGYTNCYTGNTWD
PTYCPDDETCAQNCCLDGADYSSTYGITSSGDSLNLKFVTGSNVGSRVYLLQDDSTYQIFK
LLNREFTFDVDVSNLPCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQCPRD
LKFIDGEANVEGWQPSSNNANTGIGDHGSCCAEMDVWEANSISNAVTPHPCDTPGQTMC
SGDDCGGTYSNDRYAGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQFLTDDG
TDTGTLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRHGGLAK
MGAAMQQGMVLVMSVWDDYAAQMLWLDSDYPTDADPTTPGIARGTCPTDSGVPSDVE
SQSPNSNVIFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYS
GPTVCASGTTCQVLNPYYSQCL

222211221322

QQACTLQSETHPPLTWKRCTAPGSCSTVNGSVVIDANWRWVHDVNGYTNCYTGNTWD
PTYCPDDETCAQNCCLDGAAYASTYGVTTSGDSLSIGFVTGSNVGARLYLMASDTTYQIF
KLLNREFTFDVDVSQLPCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQCPR
DLKFIDGEANVEGWQPSSNNANTGIGDHGSCCAEMDIWEANSISNALTPHPCDTPGQTMC
SGDDCGGTYSDNRYGGTCDPDGCDWNPYRLGNTSFYGPGKIIDTTKPFTVVTQFLTDDG
TDTGTLSEIKRFYIQNSNVIPQPNAELGSYSGNSLNDDYCTAEKAEFSNTDDFNRHGGLAK
MGAAMQQGMVLVMSLWDDYAAQMLWLDSDYPTDADPTTPGIARGTCPTDSGVPSDVE
SQSPNAKVTFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGY
SGPTVCASGTTCQVLNPYYSQCL

322211222322

QQACTLQSETHPPLTWKRCTAPGSCSTVNGSVVIDANWRWVHDVNGYTNCYTGNTWD
PTYCPDDETCAQNCCLDGAAYASTYGVTTSGDSLSIGFVTGSNVGARLYLMASDTTYQIF
KLLNREFTFDVDVSQLPCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQCPR
DLKFIDGEANVEGWQPSSNDANAGFGDHGSCCAEMDVWEANSISNAVTPHPCDTPGQT
MCSGDDCGGTYSDNRYGGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQFLTD
DGTDTGTLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRHGGLA
KMGAAMQQGMVLVMSLWDDHYAQMLWLDSTYPIDQAGAPGIARGTCPTDSGVPSDVES
QSPNAKVTFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYS
GPTVCASGTTCQVLNPYYSQCL

122211222322

QQACTLQSETHPPLTWKRCTAPGSCSTVNGSVVIDANWRWVHDVNGYTNCYDGNTWD
PTYCPDDETCAQNCCLDGAAYASTYGVTTSGDSLSIGFVTGSNVGARLYLMASDTTYQIF
KLLNREFTFDVDVSQLPCGLNGALYFVAMDADGGVSKYPNNKAGAKYGTGYCDSQCPR
DLKFIDGEANVEGWQPSSNNANTGIGDHGSCCAEMDVWEANSISNAVTPHPCDTPGQTM
CSGDDCGGTYSDNRYGGTCDPDGCDFNPYRMGNTSFYGPGKIIDTTKPFTVVTQFLTDD
GTDTGTLSEIKRFYIQNSNVIPQPNSDISGVTGNSITTEFCTAQKQAFSNTDDFNRHGGLAK
MGAAMQQGMVLVMSLWDDYYAQMLWLDSTYPTNETPTTPGIARGTCPTDSGVPSDVES
QSPNAKVTFSNIRFGPINSTFTGPPGGNRGTTTTRRPATTTGSSPGPTQSHYGQCGGIGYS
GPTVCASGTTCQVLNPYYSQCL

Supplementary Table 6.2: Amino acid sequences of 7 chimeric cellulases predicted to be stable (continued).

Chimera	T_{A50} (°C)
111111211111	48.3 (± 0.5)
111111311111	47.8 (± 0.1)

Supplementary Table 6.3: T_{A50} measurements of *H. jecorina* CBHI with block G from *T. emersonii* CBHI and *C. thermophilum* CBHI.

Mutations	T_{A50} (°C)
wild-type	46.9 (± 0.1)
F362M	49.6 (± 0.2)
E120I, T122K	47.4 (± 0.5)
E120M, T122K	46.6 (± 0.1)
E120M, F362M	48.8 (± 0.1)
S175A, K356H	48.4 (± 0.1)
S175A, F362M	48.3 (± 0.2)
K356H, F362M	49.3 (± 0.1)
S175A, K356H, F362M	49.1 (± 0.4)

Supplementary Table 6.4: Stability of *H. jecorina* CBHI with the single mutation F362M and other stabilizing point mutations.