ENGINEERING DIOXYGENASES BY LABORATORY EVOLUTION: A COMPARISON OF EVOLUTIONARY SEARCH STRATEGIES

Thesis by
John M. Joern

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

California Institute of Technology
Pasadena, California
2003
(Defended 1/23/03)

Acknowledgements

I am happy to acknowledge several individuals who contributed to this work and to my intellectual and personal development during my time at Caltech. As my thesis advisor, Dr. Frances Arnold has been instrumental in the success of this project. She has continuously supplied direction and fresh ideas while at the same time giving me a great deal of freedom to try new approaches and experimental designs. I can only continue to aspire to her scientific excellence and vision.

As I started my research at Caltech, I had no experience with molecular biology. These skills were to a large extent conveyed by Akira Arisawa who agreed to be my mentor in the laboratory for a summer. This was a fun and exciting time and I am especially grateful to Dr. Arisawa.

I am happy to thank two collaborators: Peter Meinhold, who developed an effective method for high-throughput sequence characterization, and Lillian Pierce, who successfully applied this method. Their data became integral to my project as a whole.

Also I would like to acknowledge Takeshi Sakamoto, who worked with me during the beginning of my project and taught me a great deal about dioxygenases. Pat Cirino, Edgardo Farinas, Hyun Joo, Zhanglin Lin, Kim Mayer, Birgit Morawski,

Daisuke Umeno and Chris Voigt were at several times useful consultants and always willing to take time to think about my problems, propose explanations and carefully explain new methods. Joff Silberg, Chris Otey, Chris Voigt and Lillian Pierce have been useful reviewers of my writing. I have had helpful discussions with more people than I can name from the Arnold group, and I have enjoyed their friendship and camaraderie over the years.

My parents are truly the best parents anyone could have and deserve special thanks. Through immense effort, they have brought out my natural talents and encouraged my interest in the physical sciences.

Finally I would like to thank my wife, Audrey, who is the love of my life and has been my best friend and supporter. She has been there for me through the inevitable frustrations and setbacks of scientific research, and always provides a sense of perspective when I most need it. Without her, I might have written this thesis, but would not have been happy in the process.

Abstract

Due to the unique and difficult chemistry they perform, the aromatic ring-hydroxylating dioxygenases are of interest as industrial catalysts. Unfortunately, an application-specific array of problems limits their utility. To address these problems through laboratory evolution, I developed methods for high-throughput screening of tens of thousands of dioxygenase variants. These methods rely on a phenol detection reagent (Gibbs reagent) and can be applied to liquid cultures or to growing bacterial colonies expressing variant enzymes.

Recombination of genes encoding homologous enzymes ("family shuffling") has emerged as a promising tool for evolutionary protein engineering. Using the dioxygenases as a model system, I have investigated the value of recombination as a search strategy for laboratory evolution. Chimeric dioxygenase libraries constructed by DNA shuffling are first evaluated for biases that limit sequence diversity using a probe hybridization approach in lieu of sequencing. This analysis shows that crossovers preferentially occur in regions with high sequence identity and that certain parent sequences can be preferred at particular gene positions.

High-throughput functional screening allowed characterization of substrate specificity for hundreds of dioxygenase chimeras. These data are coupled with sequence data to reveal sequence-function relationships and demonstrate the

utility of recombination as a tool for functional genomics. One region of sequence is shown to be a primary determinant of substrate specificity for the enzymes studied. Furthermore, several sets of variant enzymes with similar functionality are shown to have sequence similarities.

Recombination and random mutagenesis are compared as search strategies for generating functionally-diverse dioxygenases. I screened similarly sized libraries of chimeric and mutant dioxygenases for variants with altered substrate specificity or activity toward *n*-hexylbenzene, which is not accepted by the parent enzymes. Both recombination and random mutagenesis gave rise to enzymes with altered substrate specificity, although such enzymes were more frequent in the chimeric libraries and more distinct specificities were found in the chimeric libraries. Only chimeras were active toward *n*-hexylbenzene. These results support the view that recombination is an effective search strategy for evolving substrate specificity, and may be more effective than random mutagenesis.

Table of Contents

Chapter 1

Introduction

Outlook for biocatalysis and the role of directed evolution

Biocatalysis will continue to be driven by the increasing demand for more complex molecules and the growing emphasis on environmentally friendly chemical processes. Several industries, including specialty chemical, fragrance, high-tech materials and especially the pharmaceutical industry, have turned to highly complex chemicals as product candidates. For example, many of today's new drugs are large molecules with multiple chiral centers that are difficult and expensive to produce in large quantities. Nature synthesizes such compounds seemingly without effort, but not necessarily the ones we desire. As we develop methods to discover natural pathways and then engineer them to better suit the needs of industry, biocatalysis will become a fixture in the pharmaceutical and specialty chemical industries. In addition, the increased emphasis on environmental protection will continue to drive efforts to reduce waste treatment costs associated with solvent usage, toxic heavy-metal catalysts and disposal of byproducts. In contrast to the reactions of conventional chemical synthesis, enzymatic reactions typically result in few byproducts and proceed with high regio- and enantioselectivity. Enzymes require no organic solvents and are themselves completely biodegradable. As enzyme technology improves and environmental laws tighten, biocatalysis may be poised to move into large-scale chemical markets.

Despite the promise of biocatalysis, a host of problems confronts the engineer considering an enzyme or whole-cell bioconversion as part of a chemical

synthesis. These include low activity toward the substrate of interest, low stability under process conditions, low expression level and cofactor requirements. These common enzyme problems are in many cases are a direct result of constraints imposed by natural evolution. Thus it is fitting that directed evolution (which mimics natural evolution) is an effective strategy for the industrialization of natural enzymes in the laboratory (see Figure 1).

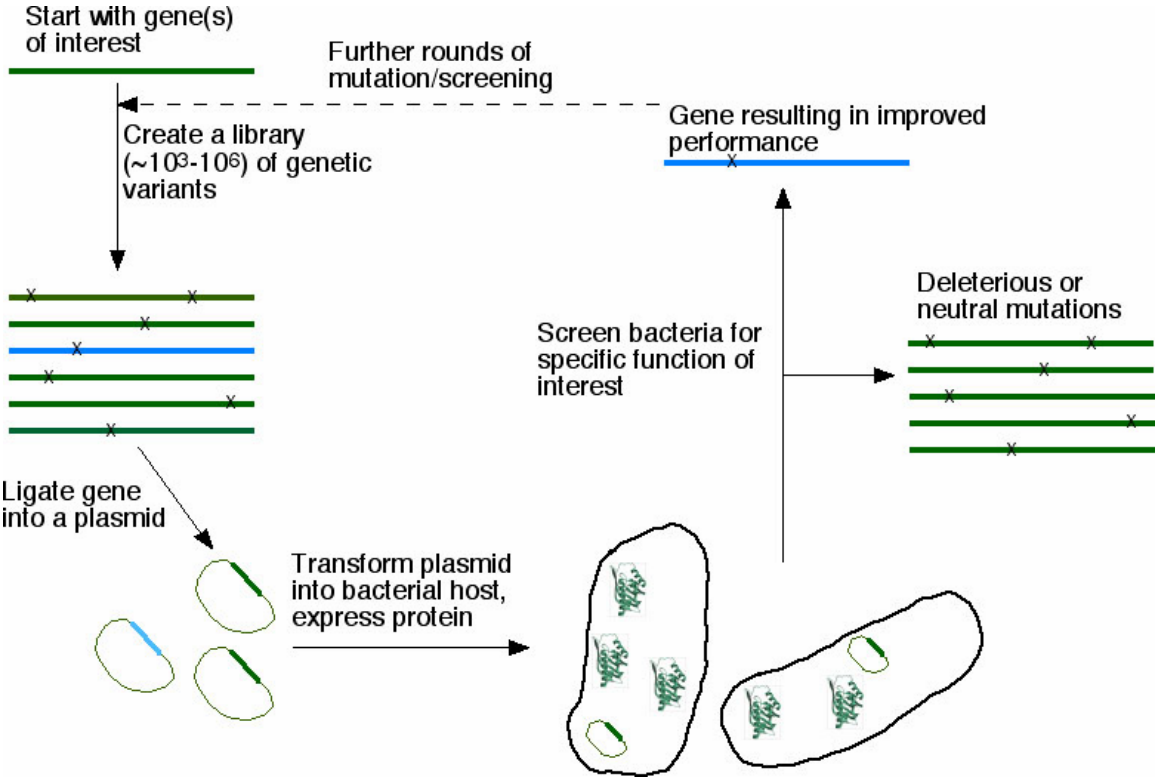


Figure 1. Schematic of directed evolution algorithm. As shown here, random mutation is applied to the gene coding for the protein to be engineered. This library of genes is ligated into an expression plasmid which is then transformed into bacteria. Individual bacterial clones are screened for the desired property. Genes from clones with improved performance are used to start the next generation. In this way, beneficial mutations can be accumulated.

While directed evolution is routinely effective in changing the substrate specificity of an enzyme and in increasing its expression level or total turnover number, it is

often too slow to apply to processes that require quick development. To successfully evolve an enzyme, the bioprocess engineer must define an assay chemistry, turn that assay into a high-throughput screen, validate that screen to ensure reproducibility, construct a library of variants, and finally screen that library and repeat the process if a suitable enzyme is not identified. For many industrially interesting enzyme reactions there is no suitable high-throughput screen for the desired product, (e.g., hydroxylation of alkanes or alkyl moieties) and most high-throughput screening methods are blind to regio- and stereospecificity--the very properties enzymes are renowned for.

These concerns should shift (and to some extent have shifted) the focus of directed evolution research to the following question: Can we advance the theory of molecular evolution to the point that we can make small (< 1000 members), highly focused libraries that can be screened on the time scale of a couple weeks by *general* chemical assay techniques (e.g., GC, HPLC or TLC) to yield significant improvements in enzyme performance or even novel functionalities? Over the last several years, a greater theoretical understanding of molecular evolution has been attained through experience with directed evolution searches and to some extent through computational studies [1-3]. Others are using computational, structure-based strategies to build more focused libraries, with the eventual goal of *a priori* prediction of function-altering mutations [4,5]. This approach is of course limited to the subset of interesting proteins that have known structures.

In this work, the *in vitro* recombination of homologous proteins is examined as a search strategy for enzyme engineering. A number of reports have shown this technique to be highly effective [6-11], but it is notoriously difficult to construct a diverse library ([9] and several researchers' unpublished observations), and the benefits of recombination over other strategies are not well-established. Thus a major goal of this work is to determine whether and how recombination can be used to create the small yet functionally diverse libraries that could potentially expand the utility of directed evolution.

Recombination as a search strategy for directed evolution

Directed evolution is an effective way to circumvent some of the problems that plague the application of enzymes to industrial processes. What is unclear is the optimal laboratory evolution method for a desired phenotypic modification. As we understand how to apply evolutionary techniques more effectively we will approach the ideal scenario discussed earlier: obtaining the desired outcome from a small library that can be screened by general analytical methods.

Two techniques are currently commonly used to generate diversity for a directed evolution search. One is random mutagenesis of the entire gene that codes for the enzyme of interest [12-16], and the other is recombination of homologous genes ("family shuffling") [7-11]. Enzymes are in general highly sensitive to

random mutations: in a library with 2-3 amino acid substitutions per gene, a significant fraction of the library (~50%) will be nonfunctional or drastically reduced in functionality [17]. So when applying random mutagenesis, we are faced with two conflicting constraints: 1) limiting the number of mutations to avoid the accumulation of deleterious mutations (and stop codons) and 2) maintaining a level of diversity that will allow for the desired phenotypic change. Recombination of homologous genes at first appears to be a better strategy because the mutations incorporated into the library have already been successful (or neutral) in the context of at least one of the parent enzymes. This allows us to change a large fraction (5-20%) of residues and still land in functional regions of sequence space [10,18]. In contrast, randomly mutating even 5% of the gene sequence of an enzyme would yield a library that is for all practical purposes completely nonfunctional. What is unclear is whether the sequence diversity accessible with recombination is enriched in *functional diversity*, or whether the amino acid changes made are too conservative.

The utility of recombination is well-demonstrated in many fields. For example, new patentable inventions often arise from a new combination of components [19,20]. Recombination is revered in the field of genetic algorithms as an efficient search technique for semi-rugged landscapes and has been used even in the evolution of artificial life *in silico* [21,22]. Nature herself employs recombination at practically every branch in the tree of life, often at the expense of immediate reproductive potential. From these examples we discover the

heuristic that recombination is most successful when the elements to be recombined are noninteracting. For instance, consider the task of building a better sailboat given a fleet of boats with known performance characteristics. One strategy might be to combine the bow of an especially speedy boat with the stern of an easily maneuverable vessel. As you cut through these unfortunate boats, you would discover that these parts are highly interacting. Because there are so many contact points, it is unlikely that the decks would be at the same level and even more unlikely that the shape of the hull would be the same at the cut point. An alternate strategy would be to put the sail, mast, and riggings of one boat on the hull of the other. We can readily see that this would be a better strategy; the abstract reason is the relative lack of interactions between hull and sail.

Enzyme structures are chock-full of interactions between noncontiguous residues, and thus it might seem that exchanging sequence elements from homologous proteins would be much like exchanging the bow from one sailboat and the stern from another. But as it turns out, much of the nonidentity we observe between homologous proteins occurs in residues that are noninteracting or weakly interacting (simultaneous nondisruptive mutation of interacting residues is quite rare), and thus important interactions are generally conserved. Current theoretical work is beginning to show that active variants from chimeric libraries have crossover points that minimize the disruption of interactions [23]. Since we can predict these points in advance, we should be able to make

chimeric libaries with a higher percentage of active members by applying this theory.

Though we are beginning to understand how to create hybrid enzymes that will remain active, we have little understanding of how to access functional diversity. We would like to know, for example, the sequence identity the parents should share, whether the parents should be functionally similar, and how many crossovers to apply in order to create libraries containing high functional diversity. In order to obtain chimeras with multiple random crossovers using current methods, the sequence identity of the parents should be >70%. In most enzyme families, these highly identical parents are functionally quite similar, and it is unclear whether functional diversity can be generated in these cases (functionally distinct parents have been used in all reported, successful studies). The crossover frequency is likely to be another crucial variable for recombination, yet no consensus has emerged in the literature. This is because generally only *evolved* chimeras are sequenced, and not chimeras from the library as a whole. Sequencing and comparison of both populations could allow assessment of whether the crossover frequency was too low or too high.

The advantages to using recombination over random mutagenesis with respect to improving catalyst performance or generating functional diversity are currently not well-understood. To date, one study has attempted a comparison of the two strategies, with the goal of creating cephalosporinases that confer improved

resistance to the antibiotic moxalactam [24]. Four homologous cephalosporinases were recombined to create a chimeric library, and also mutated individually to create four mutant libraries. Enzymes from the chimeric libraries conferred >30-fold higher resistance than the best mutants, and the authors conclude that recombination "accelerates directed evolution." However, the two best chimeric enzymes were reported to have either 14 or 33 amino acid substitutions[†], and no effort was made to determine to what extent the mutations were required for improved resistance. Furthermore, the average number of mutations incorporated in the mutant libraries was not determined; in all likelihood, few mutations were made, and this hampered the mutagenic search. In the case of this study, recombination provided the most highly active variants, but because of the lack of library characterization, this study does not allow us to attribute these functional changes to mutation or recombination.

Dioxygenases and their applications

Because of their potential utility in bioremediation and biocatalysis, I selected the aromatic ring-hydroxylating dioxygenase family as an evolution system. Degradation by dioxygenases is a major pathway by which aromatic compounds are mineralized in the environment. To cope with the enormous diversity of aromatic compounds created by diagenesis of organic material, this enzyme family has evolved remarkably broad substrate specificity. Dioxygenases are

[†] Our group's (and my) experience with random mutagenesis has been that 2-3 nucleotide mutations per gene result in inactivation of ~30-60% of the resulting clones. The high rate of mutagenesis that these clones imply would inactivate nearly all of the 50,000 chimeras that were

known to oxidize hundreds of substrates including linked and fused aromatics, aliphatic olefins, and highly substituted compounds such as tetrachlorobenzene [25]. In the prokaryotes and fungi, dioxygenases are found both chromasomally and on catabolic plasmids. In fact, in one *Sphingomonas* strain, genes encoding 49 possible dioxygenase systems were found on a single plasmid [26]. PCR-based studies of microorganisms from soil samples have revealed that dioxygenases are ubiquitous in the natural world and that we are nowhere near a full understanding of the diversity of these enzymes [27].

Since the discovery of dioxygenases over thirty years ago, much has been learned about this pathway (See Figure 1). In the first step, both atoms of dioxygen are added into the substrate to form a *cis*-dihydrodiol. Though the reaction of aromatics with oxygen is highly exergonic, the activation energy is significant. This energy barrier is overcome by transfer of electrons from NADH through a reductase and ferredoxin to the active site of the dioxygenase. NADH is regenerated in the next step which aromatizes the *cis*-dihydrodiol to a catechol. Interestingly, similar catechols are formed by the monoxygenase pathway at the additional cost of two NADHs. An extradiol dioxygenase cleaves the catechol, and the resulting compound is metabolized to carbon dioxide by the tricarboxylic acid cycle.

---

screened. Thus these sequencing results are highly suspect. The original evolved genes are no longer available for resequencing.

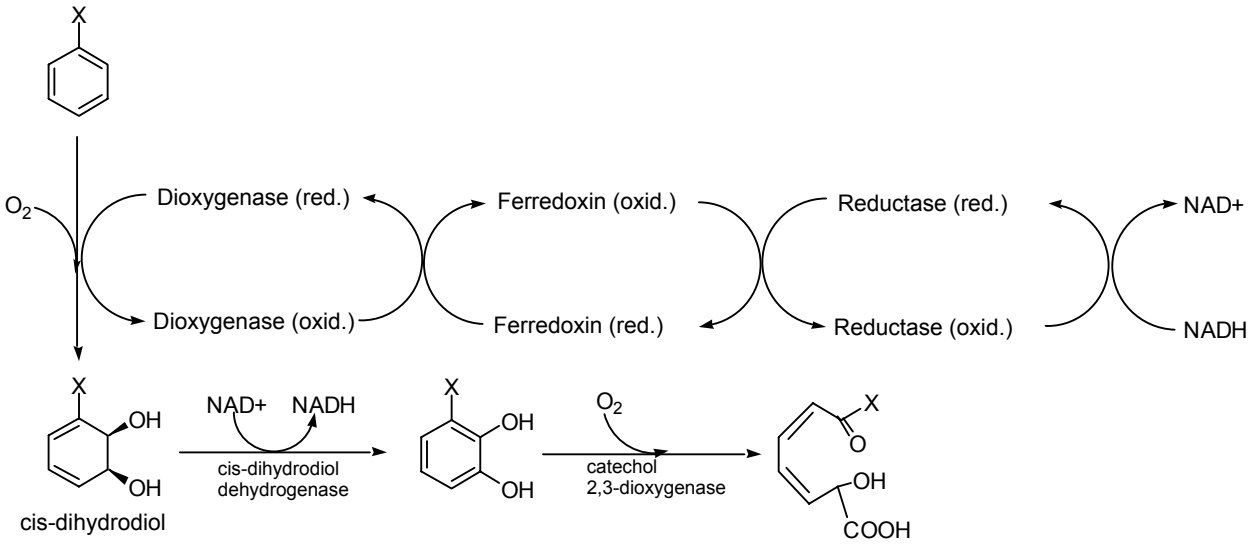


Figure 2. Degradation of a generic aromatic compound by the Class II dioxygenase pathway.

The stereospecific addition of molecular oxygen to an aromatic compound to form a *cis*-dihydrodiol (Figure 2) is the rate-limiting step of the pathway shown and results in the most interesting molecule from the standpoint of the synthetic chemist. The regio- and stereospecific oxidation of an unactivated aromatic compound is very difficult to accomplish using conventional chemical techniques, which typically produce an array of byproducts that must be separated and destroyed. Their potential for derivatization through arene functionalities makes *cis*-dihydrodiols valuable synthetic building blocks for the synthesis of biologically important pinitols, conduritols, and acyclic sugars [28] as well as the drugs Indinavir [29] and pancratistatin [30].

As discussed in the general sense in the previous section, the usefulness of dioxygenases is significantly diminished by an application-specific array of problems, including low activity toward nonnatural substrates, poor stability

(especially *in vitro*), low substrate solubility, and cofactor requirements. One of the successes of the current work is to demonstrate a general, high-throughput (10,000 variants/day) screening method suitable for directed evolution of the dioxygenase enzyme that has allowed us to address the problem of low activity toward nonnatural substrates and could be used to improve stability and expression level.

Previous application of laboratory evolution strategies to dioxygenases

Because dioxygenases are potentially useful catalysts, there has been considerable interest in developing and implementing methods for their evolution in the laboratory. The target application for many of these studies has been bioremediation of polychlorinated biphenyls (PCBs). Commercially available PCB mixtures (e.g., Aroclor) contained a variety of PCB congeners, thus enzymes with a broad substrate specificity are thought to be required. Highly substituted PCBs are highly recalcitrant and especially resistant to metabolism by natural dioxygenase pathways.

Three studies have recombined biphenyl dioxygenases from *Burkholderia cepacia* LB400 and *Pseudomonas pseudoalcaligenes* KF700, which share 95% amino acid identity, in order to evolve enzymes with broadened specificity [31,32]. Despite their high level of sequence identity, these enzymes have distinct substrate specificities [31]. Bruhlman and Chen report chimeras of these

parents that hydroxylate pentasubstituted PCBs that are not accepted by the parent enzymes [32]. Furukawa and others have evolved enzymes with altered substrate specificity and enhanced activity toward trisubstituted PCBs; some of these chimeras exhibit improved activity toward toluene and benzene, which are poorly accepted by the parent enzymes [31]. Similar results are reported in another study, where chimeras of these biphenyl dioxygenases have improved activity toward monocyclic, alkyl-substituted substrates [33].

In another study, LB400 is recombined with biphenyl dioxygenases from *Comomonas testosteroni* B-356 and *Rhodococcus globerulus* P6, which share 64-75% amino acid identity [34]. Here impressive degradation rates were reported for several PCBs not accepted by these parent dioxygenases. The broadened specificity of these evolved chimeras results primarily from just four polymorphisms located close to the active site.

Random mutagenesis of dioxygenase has also been shown to affect profound functional changes. In one case, point mutations introduced to biphenyl dioxygenase strain KF707 gave rise to variants with improved activity toward several substrates [35]. Improvements both in activity toward 4-picoline [12] and in product selectivity [36] have been accomplished through random mutagenesis of toluene dioxygenase from *Pseudomonas putida* F1.

The evolution of dioxygenases in the laboratory has been hampered by the lack of quantitative, widely applicable high-throughput screening methods. Activity toward biphenyl and PCB can be assayed by coexpressing the catechol 2,3-dioxygenase to produce yellow meta-cleavage products [31,32] or by another method with (currently) undefined chemistry [34], but these assays have low sensitivity and only apply to a limited range of substrates. More sensitive and widely applicable methods for high-throughput screening are needed and have been developed in the course of my work in collaboration with Dr. Takeshi Sakamoto.

Project aims and overview of thesis

The manuscript is composed of chapters that address the following aims:

1) Dioxygenase structure and mechanism: implications for directed evolution

In Chapter 2, I discuss what is known about the structure and mechanism of dioxygenase and how this knowledge was used to develop a directed evolution strategy.

2) Development of bacterial plasmids and high-throughput screening techniques to enable the directed evolution of dioxygenases.

Plasmids used for expression of dioxygenase in laboratory strains of bacteria are described in Chapter 3; this expression system was originally designed by Dr.

Akira Arisawa and modified by the author so that both the α and β subunits of dioxygenase could be evolved simultaneously.

In Chapter 4, schemes for the assay of dioxygenase activity are explained and the versatility and performance of each is assessed. Chapter 5 describes the application of the Gibbs' phenol detection reagent to high-throughput screening of enzyme variants, both in microtiter plates (as first conceived by Dr. Takeshi Sakamoto) and in a solid-phase approach that enables higher throughput. The screening techniques described here are versatile, cheap and fast and should find application in the evolution of several dioxygenase properties beyond what I specifically investigated, including stability, expression level and activity toward a variety of aromatic substrates.

3) Construction of highly diverse chimeric dioxygenase libraries

A variety of experimental methods have been devised to create libraries of chimeric genes from homologs sharing > 60% sequence identity [8,37-40]. Chapter 6 describes an *in vitro* recombination method based on the "DNA shuffling" protocol [8]. The chapter focuses on troubleshooting and optimization of the DNA shuffling protocol.

4) Characterization of sequence biases in chimeric dioxygenase libraries

In order to assess biases resulting from *in vitro* recombination by DNA shuffling, we devised an approach that enables partial sequencing of libraries of chimeric

genes. Peter Meinhold developed this method and subsequently used it to characterize several dioxygenase libraries. In Chapter 7, results for two dioxygenase libraries are presented which show important biases inherent to this recombination method. A predictive model is described that determines which sites are preferred for recombination.

5) Sequence-function analysis of shuffled dioxygenases

In Chapter 8, I demonstrate a systematic method for determining functional loci in protein sequences based on recombination and sequence-function analysis. Libraries of chimeric dioxygenases are screened for activity toward ten substrates. These activity data are coupled with sequence data from the probe hybridization experiments to extract sequence-function relationships. Also in Chapter 8, I identify chimeric enzymes that metabolize *n*-hexylbenzene, a substrate not measurably accepted by the parent enzymes, and show that many of the sequences of these clones are similar in some respects.

6) Comparison of functional diversity generated by recombination and mutagenesis

In Chapter 9, I compare two commonly used search strategies, random mutagenesis and recombination, with respect to their ability to generate functional diversity. I used two evolutionary tasks to evaluate these strategies: evolution of altered substrate specificity, and acquisition of activity toward *n*-

hexylbenzene. This is the first comparison of these strategies using well-characterized libraries.

References

1. Voigt, C.A., Mayo, S.L., Arnold, F.H. & Wang, Z.G. (2001). Computationally focusing the directed evolution of proteins. *J. Cellular Biochem.* **37**:58-63.
2. Voigt, C.A., Kauffman, S. & Wang, Z.G. (2001). Rational evolutionary design: The theory of *in vitro* protein evolution. *Advances in Protein Chemistry* **55**,79-160.
3. Voigt, C.A., Mayo, S.L., Arnold, F.H. & Wang, Z.G. (2001). Computational method to reduce the search space for directed protein evolution. *PNAS* **98**:3778-3783.
4. Kraemer-Pecore, C.M., Wollacott, A.M. & Desjarlais, J.R. (2001). Computational protein design. *Curr. Opin. Chem. Biol.* **5**:690-695.
5. Hayes, R.J., Bentzien, J., Ary, M.L., Hwang, M.Y., Jacinto, J.M., Vielmetter, J., Kundu, A., & Dahiyat, B.I. (2002). Combining computational and experimental screening for rapid optimization of protein properties. *PNAS* **99**,15926-15931.
6. Ness, J.E., Welch, M., Giver, L., Bueno, M., Cherry, J.R., Borchert, T.V., Stemmer, W.P.C. & Minshull, J. (1999). DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* **17**, 893-896.
7. Stemmer, W.P.C. (1994). Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389-391.
8. Stemmer, W.P.C. (1994). DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA* **91**, 10747-10751.

9. Kikuchi, M., Ohnishi, K. & Harayama, S. (1999). Novel family shuffling methods for the *in vitro* evolution of enzymes. *Gene* **236**:159-167.
10. Chang, C-C.J., Chen, T.T., Cox, B.W., Dawes, G.N., Stemmer, W.P.C., Punnonen, J. & Patten, P.A. (1999). Evolution of a cytokine using DNA family shuffling. *Nature Biotechnol.* **17**, 793-797.
11. Raillard, S., Krebber, A., Chen, Y.C., Ness, J.E., Bermudez, E., Trinidad, R., Fullem, R., Davis, C., Welch, M., Seffernick, J., Wackett, L.P., Stemmer, W.P.C. & Minshull, J. (2001). Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. *Chem. & Biol.* **8**, 891-898.
12. Sakamoto, T., Joern, J.M., Arisawa, A. & Arnold, F.H. (2001). Laboratory evolution of toluene dioxygenase to accept 4-picoline as a substrate. *Appl. and Environ. Microbiol.* **67**, 3882-3887.
13. Rai, G.P., Sakai, S., Florez, A.M., Mogollon, L. & Hager, L.P. (2001). Directed evolution of chloroperoxidase for improved epoxidation and chlorination catalysis. *Adv. Synth.Catal.* **343**, 638-645.
14. Bosma, T., Damborsky, J., Stucki, G. & Janssen, D.B. (2002). Biodegradation of 1,2,3-trichloropropane through directed evolution and heterologous expression of a haloalkane dehalogenase gene. *Appl. Environ. Microb.* **68**, 3582-3587.
15. Giver, L., Gershenson, A., Freskgard, P.O. & Arnold, F.H. (1998). Directed evolution of a thermostable esterase. *PNAS* **95**,12809-12813.

16. Wintrode, P.L., Miyazaki, K. & Arnold, F.H. (2000). Cold adaptation of a mesophilic subtilisin-like protease by laboratory evolution. *J. Biol. Chem.* **275**, 31635-31640.
17. Daugherty, P.S., Chen, G., Iverson, B.L. & Georgiou, G. (2000). Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *PNAS* **97**:2029-2034.
18. Christians,F.C., Scapozza,L., Crameri,A., Folkers,G. & Stemmer,W.P.C. (1999). Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nat. Biotechnol*. **17**, 259-264.
19. Gilfillan, S. (1935). Inventing the Ship. Follett Publishing Co., Chicago, IL.
20. Nelson, R.S. & S. Winter. (1982). An Evolutionary Theory of Economic Change. Belknap Press, Cambridge, MA.
21. Adami, C. (1998). Introduction to Artificial Life. Springer, New York.
22. Kauffman, S. (1993). The Origins of Order : Self-organization and Selection in Evolution. Oxford University Press, New York.
23. Voigt, C.A., Martinez, C., Wang, Z.G., Mayo, S.L. & Arnold, F.H. (2002). Protein building blocks preserved by recombination. *Nat. Structural Biol.* **9**, 553-558.
24. Crameri, A., Raillard, S.A., Bermudez, E. & Stemmer, W.P.C. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288-291.
25. http://umbbd.ahc.umn.edu/tol/tdo.html

26. Wackett, L.P. & Hershberger, C.D. (2001). Biocatalysis and biodegradation: Microbial transformation of organic compounds. ASM Press, Washington, D.C.

27. Yeates, C., Holmes, A.J. & Gillings, M.R. (2000). Novel forms of ring-hydroxylating dioxygenases are widespread in pristine and contaminated soils. *Environ. Microbiol.* **2**:644-653

28. Sheldrake, G.N. (1992). Biologically derived arene *cis*-dihydrodiols as synthetic building blocks. Chirality in Industry. John Wiley & Sons.

29. Buckland B.C., Drew, S.W., Connors, N.C., Chartrain, M.M., Lee, C., Salmon, P.M., Gbewonyo, K., Zhou, W., Gailliot, P., Singhvi, R., Olewinski, R.C. Jr., Sun, W.J., Reddy, J., Zhang, J., Jackey, B.A., Taylor, C., Goklen, K.E., Junker, B., Greasham, R.L. (1998). Microbial conversion of indene to indandiol: a key intermediate in the synthesis of CRIXIVAN. *Metabolic Engineering* **1**:63-74.

30. Hudlicky, T., Tian, X., Konigsberger, K., Maurya, R., Rouden, J. & Fan, B. (1996). Toluene dioxygenase-mediated *cis*-dihydroxylation of aromatics in enantioselective synthesis. Asymmetric total syntheses of pancratistatin and 7-deoxypancratistatin, promising antitumor agents. *J. Am. Chem. Soc.* **118**:10752-10765.

31. Kumamaru, T., Suenaga, H., Mitsuoka, M., Watanabe, T. & Furukawa, K. (1998). Enhanced degradation of polychlorinated biphenyls by directed evolution of biphenyl dioxygenase. *Nat. Biotech.* **16**, 663-666.

32. Bruhlmann, F. & Chen, W. (1999). Tuning biphenyl dioxygenase for extended substrate specificity. *Biotechnology and Bioengineering* **63**, 544-551.

33. Suenaga, H., Mitsuoka, M., Ura, Y., Watanabe, T. & Furukawa, K. (2001). Directed evolution of biphenyl dioxygenase: Emergence of enhanced degradation capacity for benzene, toluene, and alkylbenzenes. *J. Bacteriol.* **183**, 5441-5444.

34. Barriault, D., Plante, M.M. & Sylvestre, M. (2002). Family shuffling of a targeted bphA region to engineer biphenyl dioxygenase. *J. Bacteriol.* **184**, 3794-3800.

35. Suenaga, H., Goto, M. & Furukawa, K. (2001). Emergence of multifunctional oxygenase activities by random priming recombination. *J. Biol. Chem.* **276**, 22500-22506.

36. Zhang, N., Stewart, B.G., Moore, J.C., Greasham, R.L., Robinson, D.K., Buckland, B.C. & Lee, C. (2000). Directed evolution of toluene dioxygenase from *Pseudomonas putida* for improved selectivity toward cis-indandiol during indene bioconversion. *Metabolic Engineering* **2**, 339-348.

37. Zhao, H., Giver, L., Shao, Z., Affholter, A. & Arnold, F.H. (1998). Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol.* **16**, 258-261.

38. Volkov, A.A., Shao, Z. & Arnold, F.H. (2000). Random chimeragenesis by heteroduplex recombination. *Methods in Enzymology* **328**,456-463.

39. Shao,Z.X., Zhao,H.M., Giver,L. & Arnold,F.H. (1998). Random priming *in vitro* recombination: an effective tool for directed evolution. *Nucleic Acids Res*. **26**, 681-683.

40. Coco, W.M., Levinson W.E., Crist M.J., Hektor, H.J., Darzins, A., Peinkos, P.T., Squires, C.H. & Monticello, D.J. (2001). DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat. Biotechnol.* **19**, 354-359.

Chapter 2

Dioxygenase structure and mechanism: implications for directed evolution

Before undertaking a directed evolution experiment, it is important to consider the mechanism and structure of the enzyme system to be evolved. This is especially true with the dioxygenase enzyme system, which consists of several protein components and thus lends itself to a variety of evolutionary engineering approaches. For example, genetic variation could be applied to any combination or all of the components. Understanding the role of these components assists in tailoring the evolutionary strategy to the property to be evolved. In addition, structural and mechanistic information may suggest particular regions of a single protein that are important contributors to an evolvable property. In any case, understanding this information can help us interpret the results of a directed evolution search and may suggest alternate search strategies or add to our existing knowledge of structure-function relationships.

Functional roles for the components of the dioxygenase system

Microbial degradation of aromatic compounds through the dioxygenase pathway requires several components, as discussed in Chapter 1. Dioxygenase, ferredoxin, reductase and *cis*-dihydrodiol dehydrogenase are of the most interest here due to the synthetic utility of the *cis*-dihydrodiols and 3-substituted catechols. Dioxygenase itself carries out the initial oxidation to yield *cis*-dihydrodiols and thus is an important determinant of substrate specificity. The ferredoxin and reductase components that reduce dioxygenase are separate

polypeptides and are not strongly associated with dioxygenase [1]. Wolfe and others report that artificially reduced naphthalene 1,2-dioxygenase (NDO) does not require these electron transfer components to oxidize naphthalene and that the rate of oxidation is not accelerated by their presence [2]. Thus these components do not act as effectors and probably do not have a role in determining substrate specificity [3,4]. However, the role of these components in transferring electrons from NADH and in redistributing electrons between dioxygenase subunits [2] contributes to the overall activity of dioxygenase. For the toluene dioxygenase (TDO) [5] system, the electron transfer proteins do not limit the total activity [6].

After initial oxidation by dioxygenase, *cis*-dihydrodiol dehydrogenase converts the *cis*-dihydrodiol to a catechol while regenerating NADH. My experiments with the TDO system show that this step proceeds significantly (4-10 times) faster than initial oxidation and, in addition, that toluene *cis*-dihydrodiol dehydrogenase is highly promiscuous (data not shown).

Since dioxygenase is likely to be rate-limiting at least for the TDO system and seems to determine substrate specificity, it is probably the only component that requires modification for the evolution of enzymes with new substrate specificities or with improved activity toward a particular substrate. For other properties such as stability, activity at high or low pH, or activity in organic solvent, it may be necessary to evolve all the protein components.

Mechanism of dioxygenases

We are just beginning to understand the catalytic cycle of dioxygenases. Though conventional chemical agents such as $KMnO_4$ and OsO_4 do carry out *cis*-addition of molecular oxygen into aromatic compounds, the analogous catalytic reaction requires significant complexity. In dioxygenase, two metal cofactors are the actors for catalysis: mononuclear Fe at the active site and a 2Fe-2S Rieske cluster that transfers electrons from ferredoxin to the active site. Spectroscopic study of purified components of NDO has suggested the mechanism shown in Figure 1 [2]. At the start of the catalytic cycle, all iron atoms are in the Fe^{3+} state. Two electrons from NADH are transferred to dioxygenase through reductase and then ferredoxin to reduce iron in the Rieske cluster and active site to Fe^{2+}. The aromatic substrate binds either before or after reduction but before dioxygen is coupled to the mononuclear iron to form an apparent Fe-peroxo species. After binding the substrate and molecular oxygen, the *cis*-dihydrodiol product is formed through a still-unknown mechanism that likely involves a Fe^{3+}-OOH intermediate. This activated iron species is reactive enough to catalyze an array of diverse oxidation reactions. Dioxygenases are known to catalyze monooxygenation [7-11], oxidative dealkylation [7,8], sulfoxidation [12], desaturation [12], and oxidative dehalogenation [8]. In general, these cytochrome 450-type reactions occur with highly substituted aromatic substrates, bicyclic substrates, and highly substituted aliphatic substrates. In some cases, diols are produced along with monols [8,10]. Thus reactions besides *cis*-dihydroxylation of aromatics are possible and these capabilities could presumably be improved by evolution.

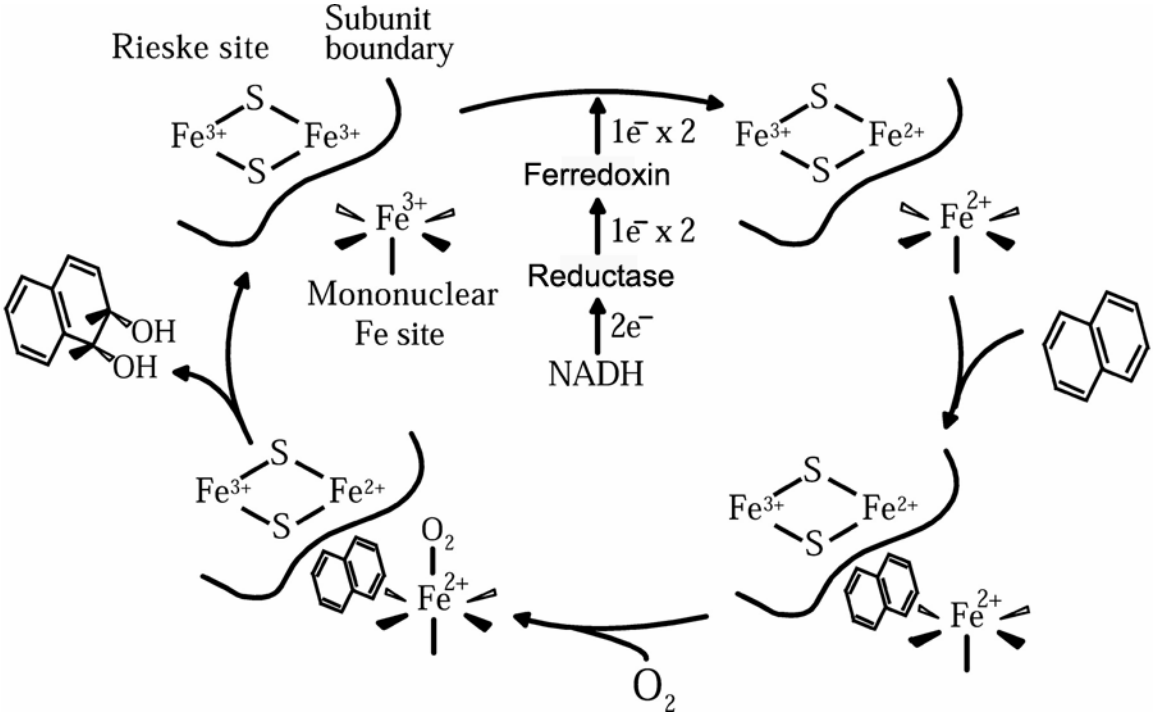


Figure 1. Catalytic cycle for naphthalene 1,2-dioxygenase. Reproduced with modification from [2] with permission.

Dioxygenase structure

The structure of naphthalene 1,2-dioxygenase (NDO) has been solved both in the presence and absence of the substrate indole [1,13]. This enzyme is a distant homolog (~30% amino acid identity) of the three dioxygenases investigated in this work, toluene dioxygenase (TDO) [5], tetrachlorobenzene dioxygenase (TCDO) [7] and biphenyl dioxygenase (BPDO) [14]. NDO is a hexamer of three large subunits (α subunits) of 449aa and three smaller subunits (β subunits) of 193aa. Many dioxygenases are thought to have this subunit arrangement, though some are composed only of α subunits [15]. The enzyme has a mushroom-like tertiary structure, with the three β subunits forming the stem, and the three α subunits forming the head (see Figure 2).

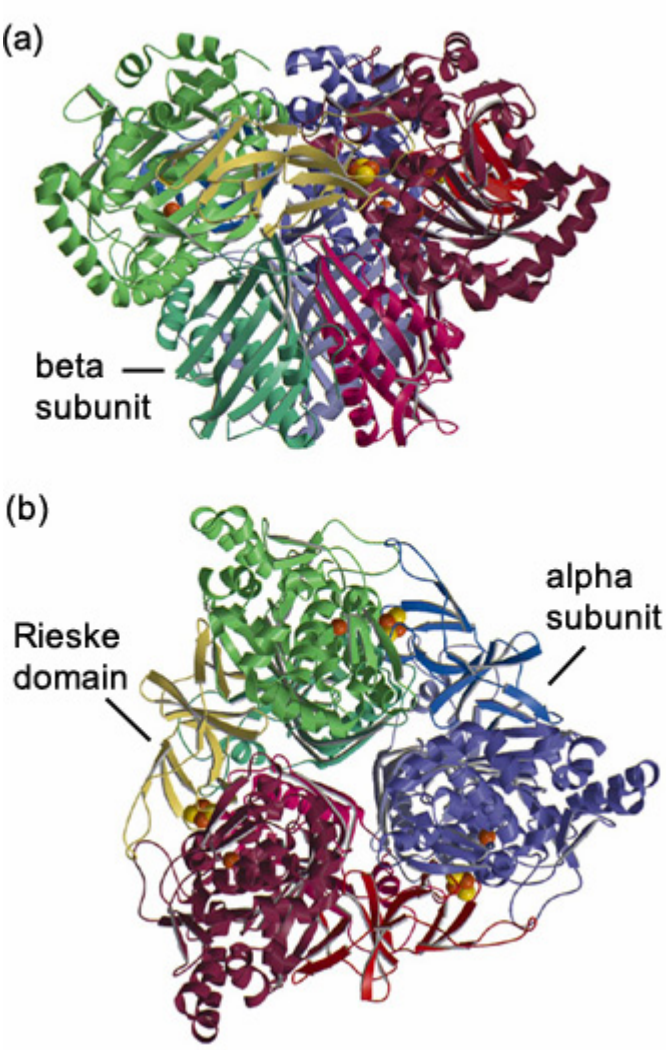


Figure 2. Orthogonal views of the tertiary structure of naphthalene dioxygenase, reproduced from [1] with permission. Subunits are differently colored, and the Rieske domain of the green α subunit is shown in yellow. Fe atoms are shown in red, and S atoms (from Rieske clusters) are shown in yellow. Panel (a) shows both α and β subunits and the mushroom configuration, and in panel (b) only α subunits are in full view.

The structure of NDO has elucidated some features of the catalytic mechanism. Because the Rieske center and active site are distant on a single α subunit but close on neighboring α subunits, it is thought that electrons are passed from the Rieske center to the active site across an α/α subunit boundary through a

conserved Asp residue [1]. The active site Fe is buried in a gorge of 15Å that is lined with hydrophobic residues. Many of these residues contact the substrate and probably determine the substrate specificity to some extent.

The β subunit is not in close proximity to either of the cofactors, and no functional role for this subunit could be ascertained from the structure. However, studies where an α subunit from one dioxygenase is coexpressed with a β subunit from another convincingly demonstrate that the β subunit can affect the substrate specificity [3,4,16,17]. For this reason, genetic variation is applied to both subunits throughout this work in the course of laboratory evolution.

References

1. Kauppi,B., Lee,K., Carredano, E., Parales,R.E., Gibson,D.T., Eklund,H. & Ramaswamy, S. (1998). Structure of an aromatic-ring-hydroxylating dioxygenase-naphthalene 1,2-dioxygenase. *Structure* **6**, 571-586.
2. Wolfe, M.D., Parales, J.V., Gibson, D.T. & Lipscomb, J.D. (2001). Single turnover chemistry and regulation of O_2 activation by the oxygenase component of naphthalene 1,2-dioxygenase. *J. Biol. Chem.* **276**,1945-1953.
3. Hirose, J., Suyama, A., Hayashida, S. & Furukawa, K. (1994). Construction of hybrid biphenyl (bph) and toluene (tod) genes for functional-analysis of aromatic ring dioxygenases. *Gene* **138**, 27-33.
4. Furukawa, K., Hirose, J., Hayashida, S. & Nakamura, K. (1994). Efficient degradation of trichloroethylene by a hybrid aromatic ring dioxygenase. *J. Bacteriol.* **176**, 2121-2123.
5. Zylstra, G.J. & Gibson, D.T. (1989). Toluene degradation by *Pseudomonas putida* F1 – Nucleotide sequence of the todC1C2BADE genes and their expression in *Escherichi coli*. *J. Biol. Chem*. **264**, 14940-14946.
6. Sakamoto, T., Joern, J.M., Arisawa, A. & Arnold, F.H. (2001). Laboratory evolution of toluene dioxygenase to accept 4-picoline as a substrate. *Appl. Environ. Microbiol.* **67**, 3882-3887.
7. Lehning, A., Fock, U., Wittich, R.M., Timmis, K.N. & Pieper, D.H. (1997). Metabolism of chlorotoluenes by *Burkholderia* sp. strain PS12 and toluene dioxygenase of *Pseudomonas putida* F1: Evidence for monooxygenation by

toluene and chlorobenzene dioxygenases. *Appl. Environ. Microbiol.* **63**, 1974-1979.

8. Lange, C.C. & Wackett, L.P. (1997). Oxidation of aliphatic olefins by toluene dioxygenase: Enzyme rates and product identification. *J. Bacteriol.* **179**, 3858-3865.
9. Spain, J.C., Zylstra, G.J., Blake, C.K. & Gibson, D.T. (1989). Monohydroxylation of phenol and 2,5-dichlorophenol by toluene dioxygenase in *Psuedomonas putida* F1. *Appl. Environ. Microbiol.* **55**, 2648-2652.
10. Resnick, S.M. & Gibson, D.T. (1996). Oxidation of 6,7-dihydro-5H-benzocycloheptene by bacterial strains expressing naphthalene dioxygenase, biphenyl dioxygenase, and toluene dioxygenase yields homochiral monol or *cis*-diol enantiomers as major products. *Appl. Environ. Microbiol.* **62**,1364-1368.
11. Robertson, J.B., Spain, J.C., Haddock, J.D. & Gibson, D.T. (1992). Oxidation of nitrotoluenes by toluene dioxygenase: evidence for a monooxygenase reaction. *Appl. Environ. Microbiol.* **58**, 2643-2648.
12. Resnick, S.M., Lee, K. & Gibson, D.T. (1996). Diverse reactions catalyzed by naphthalene dioxygenase from *Psuedomonas* sp. strain NCIB 9816. *J. of Indust. Microbiol. & Biotechnol.* **17**, 438-457.
13. Carredano, E., Karlsson, A., Kauppi, B., Choudhury, D., Parales, R.E., Parales, J.V., Lee, K., Gibson, D.T., Eklund, H. & Ramaswamy, S. (2000). Substrate binding site of naphthalene 1,2-dioxygenase: Functional implications of indole binding. *J. Mol. Biol.* **296**, 701-712.

14. Mondello, F.J. (1989). Cloning and expression in *Escherichia coli* of *Pseudomonas* strain LB400 genes encoding polychlorinated biphenyl degradation. *J. Bacteriol*. **171,** 1725-1732.

15. Mason, J.R. & Butler, C.S. (1997). Structure-function analysis of the bacterial ring-hydroxylating dioxygenases. *Adv. Microbial Phys.* **38**, 47-84.

16. Hurtubise, Y., Barriault, D. & Sylvestre, M. (1998). Involvement of the terminal oxygenase β subunit in the reactivity pattern toward chlorobiphenyls. *J. Bacteriol.* **180**, 5828-5835.

17. Parales, J.V., Parales, R.E., Resnick, S.M. & Gibson, D.T. (1998). Enzyme specificity of 2-nitrotoluene 2,3-dioxygenase from *Pseudomonas* sp. strain JS42 is determined by the C-terminal region of the α subunit of the oxygenase component. *J. Bacteriol.* **180**, 1194-1199.

Chapter 3

Construction of plasmids for expression and evolution of three dioxygenases

Introduction

A plasmid-based expression system allowing for high functional expression levels and facile cloning of gene libraries greatly facilitates the laboratory evolution of enzymes. This section describes the construction of such plasmids for this study for the expression of three dioxygenases. The approach is based on plasmids constructed by Akira Arisawa using the p*trc*99A expression system (Amersham Pharmacia). This plasmid contains the strong (and leaky) *trc* promoter and confers resistance to ampicillin. As discussed in the previous chapter, there is good reason to believe that both subunits of the dioxygenase contribute to substrate specificity. Thus the plasmid construction for laboratory evolution should allow for removal of these genes by restriction digestion so that DNA libraries created by PCR methods can be ligated into the plasmid.

Gene arrangement of three natural dioxygenase cistrons

Three parent dioxygenase systems were selected for this study. Toluene dioxygenase (TDO) from *Pseudomonas putida* F1 was obtained from D. T. Gibson whose group cloned part of the dioxygenase cistron into plasmid pDTG602 [1]. Plasmid pSTE7 containing the tetrachlorobenzene dioxygenase (TCDO) and other genes from the tetrachlorobenzene degradation pathway of *Burkholderia* sp. Strain PS12 was provided by D. H. Pieper [2]. *Pseudomonas* strain LB400 of biphenyl dioxygenase (BPDO) was obtained from F. J. Mondello

[3]. These cistrons have similar gene organization, with genes encoding the α subunit, β subunit, ferredoxin, reductase and *cis*-dihydrodiol dehydrogenase occurring in that order. For these three cistrons, a noncoding region of 97-110 bp occurs between the genes encoding the α subunit and β subunit. Genes encoding the β subunit and ferredoxin are separated by 8 bp in the TDO and TCDO systems, while BPDO contains an ORF of 420 bp between these genes. In all three cases, the start codon of reductase overlaps with the stop codon for ferredoxin.

Design of a plasmid for expression and cloning of dioxygenase variants

Figure 1 shows the plasmid design used for dioxygenase evolution. In order to express wildtype dioxygenases, plasmids pJMJ2, pJMJ6 and pJMJ7 were constructed in order to express TDO, TCDO and BPDO, respectively. KpnI and BamHI restriction sites from the multicloning site (group of unique restriction sites) of p*trc*99A flank the genes encoding the α and β subunits. Genes encoding ferredoxin, reductase and *cis*-dihydrodiol dehydrogenase are inserted between the BamHI and XbaI cloning sites. Though p*trc*99A has a ribosome binding site (rbs) upstream of the multicloning site, I have used a rbs that was shown by Dr. Akira Arisawa to give higher total activity, presumably resulting from higher expression of the α subunit (see Figure 1). In order to insert this rbs and a KpnI site during plasmid construction, I incorporated these sequences on one of the primers used for cloning the two genes encoding dioxygenase.

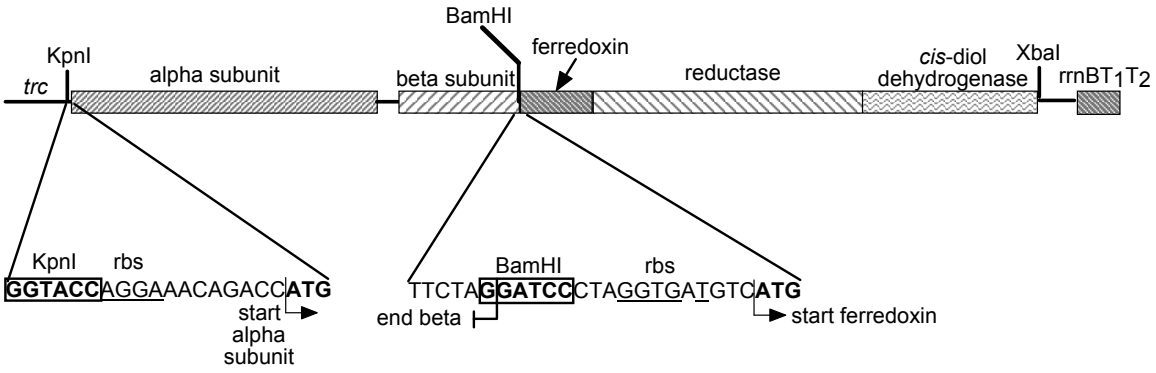


Figure 1. Schematic showing arrangment of dioxygenase genes in context of p*trc*99A expression vector.

In the natural TDO and TCDO cistrons, the β subunit and ferredoxin genes are separated by an 8bp sequence which contains an rbs. This leader sequence for ferredoxin was left intact, and a BamHI cloning site was inserted after the stop codon for the β subunit (see Figure 1). Using primers containing BamHI and XbaI sites, I amplified ferredoxin, reductase and *cis*-dihydrodiol dehydrogenase genes and ligated this PCR product to the corresponding sites in the multicloning region of p*trc*99A.

All of the plasmids and plasmid libraries used in this work contain the ferredoxin, reductase and *cis*-dihydrodiol dehydrogenase genes from the TDO parent only. These genes from TDO have previously been shown to function in concert with the dioxygenase from the TCDO system [4]. The wildtype constructs all are functioning dioxygenase systems, although decreased activity might be expected for the TCDO and BPDO expressing plasmids (pJMJ6 and pJMJ7) due to the inclusion of ferredoxin and reductase from TDO in lieu of the natural

components. I chose this design because my main goal is to evolve substrate specificity, a property which these components are thought not to influence.

Sequencing of wildtype plasmids

Genes encoding the α and β subunits from pJMJ2, pJMJ6 and pJMJ7 have been sequenced. The mutations shown in Table 1 below presumably resulted from PCR used to clone these genes. The mutations on the TDO construct are thought to be functionally neutral since pJMJ2 performed nearly as well as a similar strain constructed by Dr. Akira Arisawa when assayed contemporaneously for activity toward chlorobenzene.

plasmid	mutation	aa change	subunit
pJMJ2 (TDO)	g841a	Val281Ile	α
pJMJ2 (TDO)	g1105a	Gly369Ser	α
pJMJ2 (TDO)	t1540c	Val26Ala	β
pJMJ6 (TCDO)	g249a	*Arg83Arg*	α
pJMJ7 (BPDO)	t48c	*Val16Val*	α
pJMJ7 (BPDO)	a1599g	*Glu34Glu*	β
pJMJ7 (BPDO)	a1781g	Lys95Arg	β

Table 1. Mutations present in constructed dioxygenase-expressing plasmids. Nucleotide mutations are numbered according to the 2023-2064 bp of each αβ gene pair, while amino acid changes are numbered based on the particular subunit in which they occurred. Synonomous mutations are italicized.

As shown in Figure 2, pJMJ2 had an insertion located upstream of the start codon for the α subunit. A second KpnI cloning site is present, and between the

two sites is a 28 bp insertion that probably originated from primer concatenation during PCR. This insertion was removed by KpnI digestion followed by religation to create plasmid pJMJ11, which exhibits 3- to 4-fold higher activity than the pJMJ2 construct.

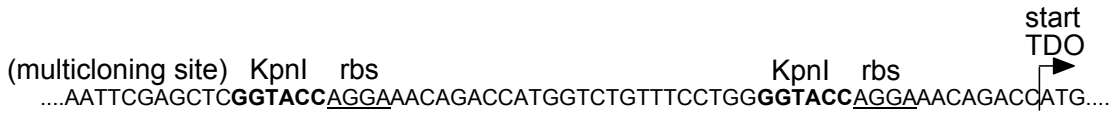


Figure 2. Sequence of pJMJ2 upstream of α subunit start codon.

Analysis of expression level by SDS-PAGE

E. coli BL21(DE3) (Stratagene) containing plasmids pJMJ11, pJMJ6, pJMJ7 and p*trc*99A were grown, induced and then lysed with Bugbuster (Novagen) according to manufacturer's instructions. Cell extracts were analyzed by SDS-PAGE as described [5] (Figure 3). Most of the dioxygenase proteins are not visible due to low expression level or the obscuring effect of native proteins. Toluene *cis*-dihydrodiol dehydrogenase (28.7 kDa) is visible, and there is increased density where the β subunit should appear, around 22 kDa. Though the calculated molecular weight for the β subunit of the three dioxygenases is nearly identical (21.9 - 22.1 kDa), the BPDO β subunit appears larger by approximately 1kDa. These results show that the dioxygenase system does not constitute a large fraction of the total soluble protein. The expression level observed is similar to that found by others upon expression of TDO in *E. coli* [1].

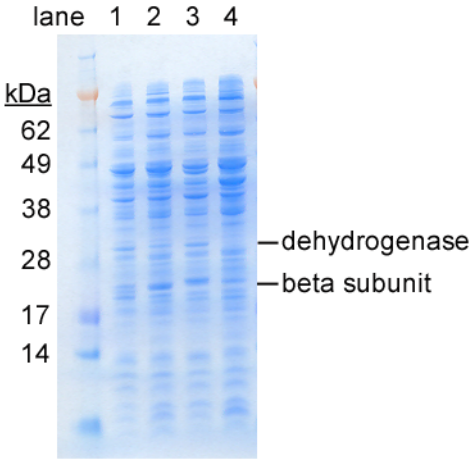


Figure 3. SDS-PAGE analysis of cell extracts from *E. Coli* BL21(DE3) expressing pJMJ11 (lane 1), pJMJ6 (lane 2), pJMJ7 (lane 3) and p*trc*99A (lane 4).

References

1. Zylstra,G.J. & Gibson,D.T. (1989). Toluene degradation by *Pseudomonas putida* F1 – Nucleotide sequence of the todC1C2BADE genes and their expression in *Escherichia-coli*. *J. Biol. Chem*. **264**, 14940-14946.
2. Beil, S., Happe,B., Timmis,K.N.& Pieper,D.H. (1997). Genetic and biochemical characterization of the broad spectrum chlorobenzene dioxygenase from *Burkholderia sp*. strain PS12-Dechlorination of 1,2,4,5-tetrachlorobenzene. *Eur. J. Biochem.* **247**,190-199.
3. Mondello,F.J. (1989). Cloning and expression in *Escherichia coli* of *Pseudomonas* strain LB400 genes encoding polychlorinated biphenyl degradation. *J. Bacteriol*. **171,** 1725-1732.
4. Beil, S., Mason, J.R., Timmis, K.N. & Pieper, D.H. (1998). Identification of chlorobenzene dioxygenase sequence elements involved in dechlorination of 1,2,4,5-tetrachlorobenzene. *J. Bacteriol.* **180**, 5520-5528.
5. Sakamoto, T., Joern, J.M., Arisawa, A. & Arnold, F.H. (2001). Laboratory evolution of toluene dioxygenase to accept 4-picoline as a substrate. *Appl. Environ. Microbiol.* **67**, 3882-3887.

Chapter 4

Colorimetric assays for dioxygenase activity

Preface

This chapter is adapted from work coauthored with Christopher R. Otey entitled "High-throughput screen for aromatic hydroxylation" to be published in the book **Methods in Molecular Biology** by Humana Press (F. H. Arnold and G. Georgiou, Eds.). The goal of this series is to publish protocols that, unlike the primary literature, provide enough detail to be followed with little difficulty the first time. Thus the protocols described here are written with frequent annotation and discussion of steps that may require optimization upon application to a new problem. I have removed material primarily contributed by the coauthor of this work, though I am indebted to him for organization of the data and rewording of some of the text presented here.

Introduction

As discussed in Chapter 1, the conventional chemical hydroxylation of unactivated aromatic compounds generally requires extremes of temperature and pressure, results in an array of byproducts, and often requires expensive and/or toxic heavy metal catalysts. One alternative that avoids these problems is enzymatic hydroxylation by enzymes such as the cytochrome P450 monooxygenases, other non-heme monooxygenases, and the dioxygenases. These catalysts are often not well-suited for industrial applications, but can be improved systematically by a laboratory evolution regime consisting of rounds of

genetic variation and screening for improvements. Thus broadly applicable methods for screening enzyme candidates quickly and reproducibly are highly valuable.

In this chapter, two colorimetric assays for hydroxylated aromatic compounds are discussed that can be implemented in high-throughput and thus are useful for biocatalyst discovery and engineering by directed evolution. These assays employ compounds that react with a variety of phenols to yield highly colored products; two such compounds are Gibbs' reagent and Fast Violet B (FVB) (Figure 1(a,b)). Both of these assay chemistries can be used to assess the performance of dioxygenases expressed in *E. coli.* Dioxygenases insert both atoms of molecular oxygen into aromatics to yield *cis*-dihydrodiols with high (>97%) enantiomeric excess (Figure 1(c)). These products are converted to phenols to enable detection using the reagents described here (Figure 1(d,e)). Other enzymes such as oxidative dealkylases or dehalogenases could be assayed similarly with only slight modification.

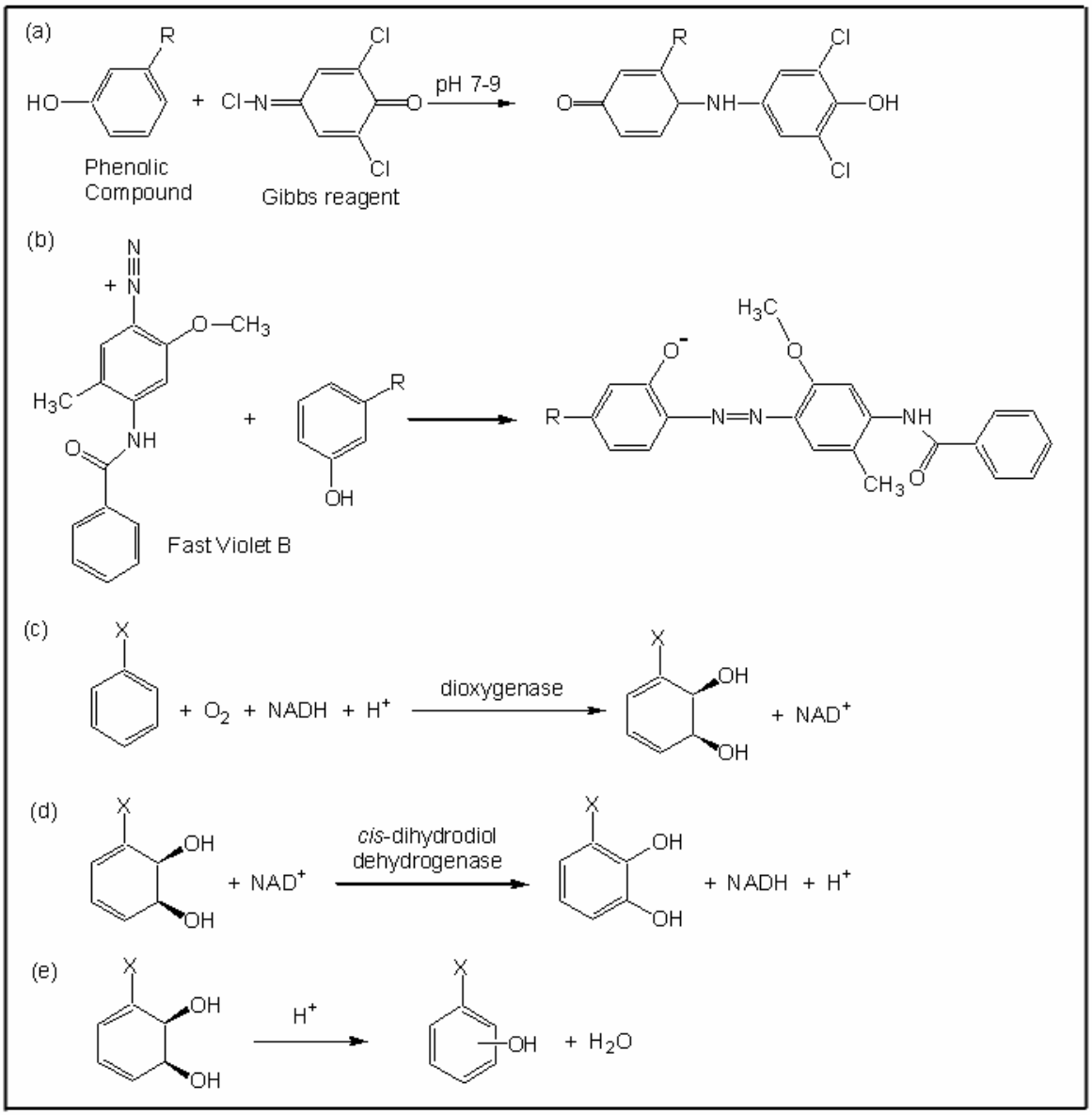


Figure 1. Chemistry of assay methods. (a) Coupling of Gibbs' reagent to a phenolic compound [1]. (b) Coupling of Fast Violet B to a phenolic compound [2]. (c) Reaction performed by dioxygenases to yield a *cis*-dihydrodiol. (d) Dehydrogenation of a *cis*-dihydrodiol to form a catechol. (e) Acidification of a *cis*-dihydrodiol to form phenols.

Some issues need to be considered when applying these assays to biotransformations using whole cells or cell extracts. When a whole cell system is used, careful consideration should be given to the method of supplying

substrate to the enzyme. To access the enzyme, the substrate must be soluble and must readily permeate the cell membrane. Solubility can be increased in most cases by adding a nontoxic organic solvent [3]. The antibiotic Polymyxin B increases the permeability of many aromatic and non-aromatic substrates, including long chain fatty acids [4,5]. Though TB or LB-media are commonly used for whole cell growth, these rich broths contribute a significant amount of background in the assays discussed here (especially the Fast Violet B assay). This is easily remedied by using a synthetic medium such as M9 minimal medium [6]. Supplying the substrate in the vapor phase is sometimes successful when the substrate is volatile and is particularly convenient when screening colonies using a solid-phase format [7].

Materials

1. 0.4% (w/v) 2,6-Dichloroquinone-4-chloroimide in ethanol (Gibbs' reagent). Store at 4°C and prepare fresh every 4 months.
2. 0.25% (w/v) Fast Violet B in *dd*H_2O (FVB). Prepare fresh every 2-3 days.
3. 1 M Tris-HCl, pH 8.5
4. 100 mM HCl.
5. 96-well microtiter plates, e.g., R-96-OAPF-ICO (Rainin, Emeryville, CA).
6. Spectrophotometer/plate reader (Model Spectra max Plus 384, Molecular Devices, Sunnyvale, CA). Software Softmax Pro 3.1.1.

7. Benchtop centrifuge that can accommodate 96-well microtiter plates: Allegra 25R Centrifuge (Beckman Coulter, Fullerton, CA).
8. Pipette robot: Multimek 96 Automated 96-Channel Pipetter (Beckman Instruments, Palo Alto, CA).
9. Multichannel pipetter.
10. Incubator at 37°C.

Methods

Gibbs' reagent was able to detect most of the ortho- and meta-substituted phenolic compounds I have tested (Table 1). It is also useful for assaying para-substituted compounds where the substituent is a halide or alkoxy group [8]. Fast Violet B (FVB) is typically less useful due to its reactivity with cells and various media, and it did not provide sensitive detection for most of the tested phenols. The absorbance recorded for these assays is linearly dependent on concentration for all of the phenols we have examined [7]. The wavelength of maximal absorbance varies based on the structure of the phenol and thus should be determined for each expected phenolic product. It may also be useful to adjust reagent concentrations to reduce background when optimizing a new assay system.

Since these assays are able to determine the products of multiple types of enzyme reactions, reaction conditions (e.g., cell growth times and temperatures,

substrate concentration, cell harvesting/lysis, etc.) will vary considerably and will not be discussed here. In the assay descriptions below, “sample” refers to the solution containing the phenolic product to be determined and may be a cell extract or supernatant depending on which type of bioconversion is chosen. The absorbance recorded after addition of the phenol detection reagent reflects the total activity of the cellular biocatalyst. Times for color development are suggested below, but this is another factor that varies from substrate to substrate and should be determined on an individual basis.

Compound	Gibbs' assay		Fast Violet B assay	
	λ_{max}	Max abs.	λ_{max}	Max abs.
3-hydroxybenzaldehyde	670	0.06	n/a	< 0.05
2-hydroxybenzaldehyde	660	0.09	380	0.09
2-hydroxybenzamide	660	2.98	n/a	< 0.05
2,3-dihydroxybenzaldehyde	570	0.63	n/a	< 0.05
catechol	460	0.44	n/a	< 0.05
3-methylcatechol	460	0.49	n/a	< 0.05
3-fluorocatechol	450	0.38	n/a	< 0.05
phenol	630	0.10	n/a	< 0.05
o-cresol	610	0.31	n/a	< 0.05
m-cresol	620	0.17	n/a	< 0.05
2-aminophenol	600	0.76	440	0.08
3-aminophenol	570	1.61	480	0.82
2-chlorophenol	660	2.77	n/a	< 0.05
3-chlorophenol	670	1.25	n/a	< 0.05
1-naphthol	580	0.31 [1]	n.d.	n.d. [2]
2-naphthol	n.d.	n.d. [2]	520	0.47
2,3-dihydroxynaphthalene	510	0.79	n.d.	n.d. [2]
4-nitrophenol	n/a	< 0.05	n/a	< 0.05
2-hydroxypyridine	n/a	< 0.05	n/a	< 0.05
3-hydroxypyridine	600	0.39	n/a	< 0.05
o-coumaric acid	650	0.42	n/a	< 0.05
m-coumaric acid	670	0.24	n/a	< 0.05
p-coumaric acid	560	0.38	n/a	< 0.05
3-hydroxybenzoic acid	640	0.25	n/a	< 0.05
3,4-dihydroxybenzoic acid	460	0.29	n/a	< 0.05
3-hydroxy-4-methylbenzoic acid	610	1.21	n/a	< 0.05
2,3-dihydroxybenzoic acid	440	0.27	n/a	< 0.05

1. Product slightly insoluble. 2. Product insoluble.
n/a - not applicable, no significant absorbance n.d. - not determined due to insolubility

Table 1. The spectroscopic signals resulting from coupling of various phenols to Gibbs' reagent and Fast Violet B. Compounds were diluted in M9 minimal medium to a concentration of 0.25 mM and assayed as described. For the Gibbs' reagent and Fast Violet B assays, 0.1ml of phenol solution was assayed in a 96-well microtiter plate, and thirty minutes or 10 minutes, respectively, were allowed for the reaction to occur before recording the visible spectra using a 96-well spectrophotometer.

Phenol quantitation with Gibbs' reagent

1. To 100 μL of sample (*see* Note 5) add 20 μL 0.4% (w/v) of Gibbs' reagent.

2. Mix and allow 3-30 minutes for color development (*see* Note 1).
3. Record spectrum or wavelength.

Phenol quantitation with Fast Violet B

1. To 100 μL of sample (*see* Note 5) add 10 μL 0.25% of (w/v) Fast Violet B.
2. Mix and allow 10 minutes for color development (*see* Note 2, Note 1).
3. Record spectrum or wavelength.

Applying phenol detection with Gibbs reagent to dioxygenases

Initial oxidation of aromatic compounds by dioxygenase results in arene *cis*-dihydrodiols, as shown in Figure 1(c). These compounds are difficult to detect in the background of a cell extract or supernatant but are easily converted to detectable phenolic compounds using one of two methods. One is to convert the *cis*-dihydrodiol to a catechol by coexpressing the *cis*-dihydrodiol dehydrogenase that resides on the dioxygenase cistron, as shown in Figure 1(d). The dehydrogenase from the toluene dioxygenase cistron of *Pseudomonas putida F1* is highly expressed in laboratory strains of *E. coli.* Another method for converting *cis*-dihydrodiols to phenols is acidification, as shown in Figure 1(e) [7] (see procedure below). The ratio of ortho- to meta-phenols is difficult to predict, but both types generally react with the detection reagents discussed here.

1. In a 96-well microtiter plate, combine 100 μL of cell extract or supernatant from a biotransformation performed in M9 minimal medium [6] with 100 μL of 0.1M HCl (*see* Notes 3-5).
2. Incubate at 37°C for 30 minutes.
3. Add 25 μL of 1M Tris-HCl, pH 8.5 (*see* Note 3).
4. Add 20 μL of 0.4% (w/v) Gibbs' reagent.
5. Record spectrum or wavelength after 3-30 minutes (*see* Note 1).

Notes

1. Optimal development time depends on the phenol assayed and, in some cases, accumulation of background absorbance over time. When assaying for improved enzyme function, only the wavelength of the product is taken and not the entire spectrum.
2. Increasing the pH to basic levels before addition of Fast Violet B can be useful in increasing the maximum absorbance value. It is not necessary, however.
3. For the *cis*-dihydrodiol products of dioxygenation of toluene and chlorobenzene, pH <2.5 should be reached after adding 0.1M HCl. Incubation at low pH may or may not be required for acidification of other *cis*-dihydrodiols. Reaction with Gibbs' reagent proceeds best at pH 7-9, thus the pH should be near or above neutral after addition of Tris buffer. If media other than M9 [6] are used, a Gibbs' reagent-compatibility check should be

made, and the amount of acid and Tris buffer added should be adjusted to match these pH ranges.

4. A pipetting robot can be useful when doing multiple 96-well microtiter plate assays, but is not necessary.
5. Removal of cell debris is not necessary for the Gibbs' assay, however it increases the sensitivity and reproducibility of the screens. It is necessary for FVB.

References

1. Quintana, M. G., Didion, C., & Dalton, H. (1997). Colorimetric method for a rapid detection of oxygenated aromatic biotransformation products. *Biotechniques* **11**, 585-587.
2. Zollinger, H. (1991). *Color Chemistry: Sytheses, Properties and Applications of Organic Dyes and Pigments*, (VCH Publishers, Inc., New York, NY).
3. Harrop, A. J., Woodley, J. M. & Lilly, M. D. (1992). Production of Naphthalene-*cis*-glycol by *Pseudomonas putida* in the Presence of Organic Solvents. *Enzyme Microb. Technol.* **14**, 725-730.
4. Schwaneberg, U. Otey, C., Cirino, P. C., Farinas, E. & Arnold, F. H. (2001). Cost-effective whole-cell assay for laboratory evolution of hydroxylases in *Escherichia coli*. *J. Biomol. Screen*. **6**, 111-117.
5. Vaara, M. (1992). Agents that increase the permeability of the outer membrane. *Microbiolgical Reviews* **56**, 395-411.
6. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, NY).
7. Joern, J. M., Sakamoto, T., Arisawa, A. & Arnold, F. H. (2001). A versatile high throughput screen for dioxygenase activity using solid-phase digital imaging. *J. Biomol. Screen.* **6**, 219-23.
8. Josephy, P. D. & Van Damme, A. (1984). Reaction of Gibbs' reagent with para-substitued phenols. *Anal. Chem*. **58**, 813-814.

Chapter 5

A versatile high-throughput screen for dioxygenase activity using solid-phase digital imaging

(Joern, J. M., Sakamoto, T., Arisawa, A. & Arnold, F. H. (2001). *J. Biomol. Screen.* 6, 219-223.)

Preface

In this chapter, we demonstrate application of the Gibbs assay described in the previous chapter to two high-throughput screening formats. Dr. Takeshi Sakamoto and I both made contributions to the microtiter plate assay method described. I developed both the application of the Gibbs chemistry to the "solid-phase" (i.e., to colonies of growing bacteria), and the quantitative image analysis techniques. This work was published previously in the *Journal of Biomolecular Screening*, Vol. 6, Issue 4, pp. 219-223, with myself, T. Sakamoto, A. Arisawa and F. H. Arnold as authors.

Abstract

We have developed a solid-phase, high-throughput (10,000 clones/day) screen for dioxygenase activity. The *cis*-dihydrodiol product of dioxygenase bioconversion is either converted to a phenol by acidification or to a catechol by reaction with *cis*-dihydrodiol dehydrogenase. Gibbs reagent reacts quickly with these oxygenated aromatics to yield colored products that are quantifiable using either a microplate reader or digital imaging and image analysis. The method is reproducible and quantitative with as little as 30μM biotransformation products, and essentially no background results from media components. This method is an effective general screen for aromatic oxidation and should be a useful tool for the discovery and directed evolution of oxygenases.

Introduction

Bacterial dioxygenases are multicomponent enzyme systems that catalyze the stereospecific introduction of molecular oxygen into a wide variety of aromatic compounds to form arene *cis*-diols (Figure 1A). In the first step of the dioxygenase mechanism, electron transfer proteins shuttle electrons from NADH to the Reiske [2Fe-2S] cluster of the terminal dioxygenase [1]. These electrons activate the mononuclear iron at the active site of the enzyme, allowing molecular oxygen and substrate to bind and react [2]. More than 300 diverse substrates ranging in size from halogenated ethylenes [3] to polyaromatic hydrocarbons such as phenanthrene and dibenzo-1,4-dioxin [2,4] can be dihydroxylated by dioxygenases. In the natural aromatic biodegradation pathway, dihydroxylation is followed by rearomatization to a catechol by *cis*-dihydrodiol dehydrogenase. Catechol is further degraded to provide a carbon and energy source for the host organism.

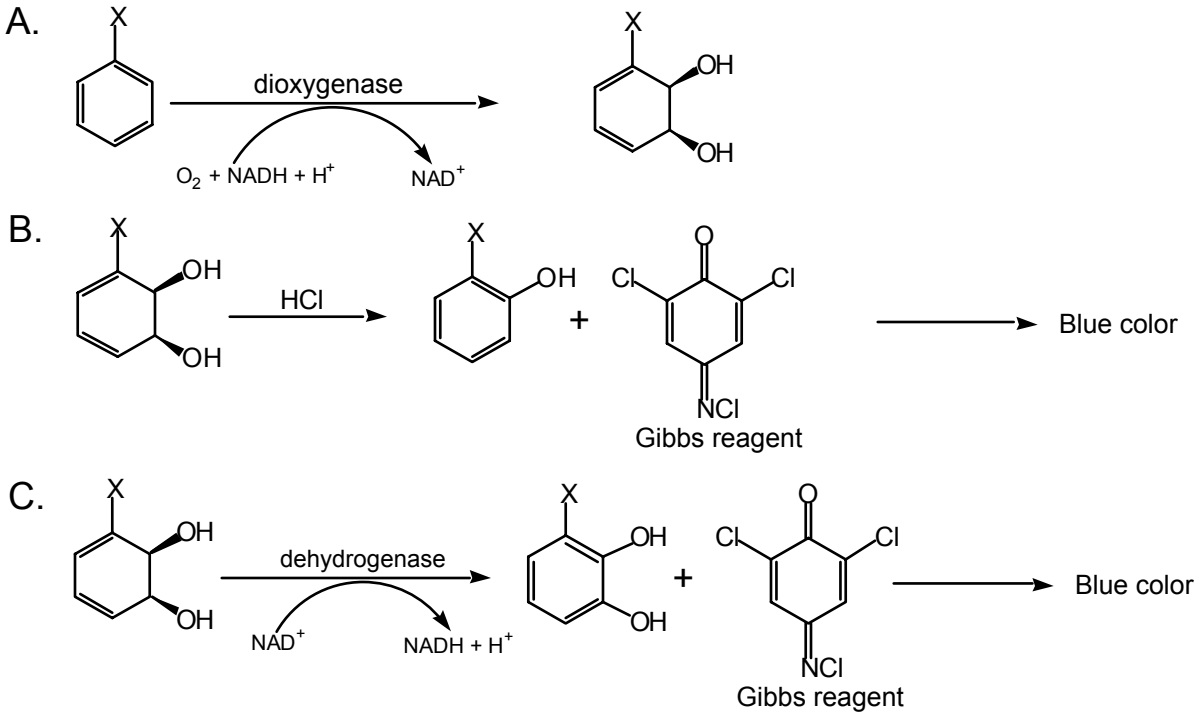


Figure 1. Chemistry used to detect dioxygenase products. A. Bioconversion catalyzed by dioxygenase enzymes, shown generically for a monocyclic, monosubstituted aromatic compound. B. Acidification method for quantitation of dihydrodiol. At low pH, dihydrodiol undergoes dehydration to either an *o*- or *m*-phenol, which is detectable with Gibbs reagent. C. Dehydrogenase method for quantitation of *cis*-dihydrodiol. *Cis*-dihydrodiol dehydrogenase converts *cis*-dihydrodiol to a catechol, which reacts with Gibbs reagent.

Because of their broad substrate range [4] and high enantioselectivity [5], dioxygenases are excellent candidates for applications in bioremediation [6], synthetic chemistry [7] and combinatorial biocatalysis [8]. Optically pure arene *cis*-diols formed by recombinant organisms lacking *cis*-dihydrodiol dehydrogenase have several proposed applications as starting materials in the synthesis of chiral drugs and specialty chemicals [7]. Unfortunately, the process engineer considering implementing a dioxygenase bioconversion is discouraged by their low activity toward unnatural substrates, their low stability, especially *in vitro,* the low solubility and toxicity of their substrates, the NADH cofactor

requirement, and by product inhibition and toxicity [9-12]. For many prospective applications, it will be necessary to modify the catalyst. This can be done using directed evolution, in which enzymes are improved by cycles of mutagenesis, recombination, and screening [18,19]. A reliable, high-throughput activity assay is crucial to any directed evolution effort. Rapid screens usually employ a whole-cell bioconversion, with little or no purification of products. Because these biological samples are so chemically complex, the detection chemistry usually must be specific to the analyte of interest rather than to a generic chemical functionality.

Here we present a high-throughput digital imaging screen for dioxygenase activity. In this method (See Figure 1), arene *cis*-diol products are converted to phenolic or catecholic compounds through acidification or further reaction with the dehydrogenase. This step is followed by colorimetric detection with 2,6-dichloro-*p*-benzoquinone (Gibbs reagent) [13]. This versatile method can be applied to numerous aromatic dioxygenase substrates and can be used as a general screen for aromatic oxidation reactions. High-throughput quantitation of activity using digital imaging and image analysis is sensitive and reproducible, making this screen ideal for directed evolution and catalyst discovery.

Materials and methods

Terrific Broth, ampicillin, chlorobenzene, and Gibbs reagent were purchased from Sigma (St. Louis, MO). Isopropyl-β-D-thiogalactopyranoside was purchased from ICN Biomedicals, Inc. (Aurora, OH). ME25 0.45 μm nitrocellulose membranes were purchased from Schleicher & Schuell (Keene, NH). V-bottom microtiter plates were purchased from Corning (Corning, NY). BL21(DE3) cells were provided by Stratagene (La Jolla, CA). Taq buffer, $MgCl_2$, and AmpliTaq DNA polymerase were supplied by Perkin Elmer (Norwalk, CN). *Cis*-(1S,2S)-chloro-3,5-cyclohexadiene-1,2-diol (chlorobenzene *cis*-dihydrodiol) was purchased from QuChem (Belfast). The genes encoding toluene dioxygenase were kindly provided by D.T. Gibson on the plasmid pDTG602 [14].

Liquid-phase screening for activity toward chlorobenzene

Colonies of *E. coli* BL21(DE3) expressing pJMJ8 were inoculated into the wells of a 96-well plate containing 100μL of LB supplemented with 100mg/L ampicillin. Cells were grown for 18 hours in a shaking incubator set to 37°C. 5μL of each culture was transferred to a V-bottom microtiter plate containing 95μL of M9 media [17] supplemented with 100mg/L ampicillin, 1mM IPTG, 1.6% D-glucose, and 80 mg/L $FeSO_4$-$7H_2O$ (M9-GIA). These cultures were incubated at 30°C for 5 hours without shaking. To start the biotransformation, 50μL of M9-GIA also containing 15 mM chlorobenzene was added, and the plate was wrapped in

Saran wrap and incubated at 30°C for 90 minutes. The 96-well plate was then centrifuged at 1700 x g for 10 minutes. 100µL of supernatant was transferred to a flat-bottom, transparent microtiter plate containing 100µL of 0.1M HCl. This plate was incubated at 37°C for 30 minutes, and then 20µL of 1M Tris-HCl, pH 8.5, was added to raise the pH. At this point, 25µL of 0.4% Gibbs reagent in ethanol was added to each well. The absorbance at 652nm was read after 40 minutes at room temperature.

Solid-phase screening for activity toward chlorobenzene

Plasmid pJMJ2 was transformed into BL21(DE3) competent cells and plated on terrific broth (TB) agar plates containing 100mg/L ampicillin and 0.5mM IPTG. Plates were incubated for 6 hours at 37°C, and then at 30°C for 12-14 hours. Colonies were lifted with a nitrocellulose membrane and transferred to M9 media [17] containing 4% agar, 100mg/L ampicillin, 1.6% D-glucose, and 80 mg/L $FeSO_4$-$7H_2O$. The colonies were then incubated for 20 minutes in an airtight container at 30°C containing an open dish of chlorobenzene. The membrane was transferred to a 4% agarose plate also containing 0.025% Gibbs reagent (added as a 2% solution in ethanol).

Digital imaging and analysis

The final agarose plate was imaged using a Fluor-S MultiImager (Biorad, Hercules, CA) equipped with a Tamron SP AF20-40mm lens (Tamron Co., Ltd., Tokyo, Japan). Digital images (1300x1000 pixels) were imported to the image analysis tool Optimas (Optimas Corp., WA) for filtering and quantitation. A median filter was run, followed by Wallis filtering with a 5x5 grid size. A 5x5 averaging filter was then applied three times. For colony detection, a threshold intensity was set such that only active colonies were highlighted. Using Optimas, we calculated intensity, area, and circularity statistics for each colony. To determine the fraction of wild-type activity retained by each colony, the difference of the mean colony intensity and the threshold intensity is divided by the difference of the average wild-type mean colony intensity and the threshold intensity.

Error-prone PCR of gene coding for the large subunit of toluene dioxygenase and library construction

Two primers (5'-CGGAATTCTAGGAAACAGACCATG-3' and 5'-CCGGATCCAACCTGGGTCGAAGTCAAATG-3') were used to amplify the gene encoding the large subunit of toluene dioxygenase under error-prone conditions. A reaction volume of 100μL contained: 133pg of pJMJ8-like plasmid DNA, 40

pmoles of each primer, 1xTaq buffer, 0.2 μmoles of each dNTP, 0.7 μmoles of $MgCl_2$, 60 nmoles of $MnCl_2$, and 2.5U of AmpliTaq DNA polymerase. PCR was carried out in a MJ Research PTC-200 thermal cycler (Watertown, MA) under the following conditions: 3 minutes at 94°C, 30 cycles of (30 seconds at 94°C, 30 seconds at 50°C, 1 minute at 72°C), and 3 minutes at 72°C. PCR product was cloned by restriction digestion and ligation into an appropriately digested vector.

Results and Discussion

Detecting the products of dioxygenase-catalyzed dihydroxylation in liquid media

The chemistry by which *cis*-dihydrodiol products are detected in liquid-phase biotransformations is shown in Figure 1a. After biotransformation of substrate by *E. coli* expressing a pJMJ8-type plasmid (See Table 1), the cells are removed by centrifugation, and the pH of the supernatant is lowered to 2.0 for dehydration of *cis*-dihydrodiol to the corresponding phenol. Addition of buffer to the supernatant raises the pH to 8.0. At this pH, phenols couple with Gibbs reagent rapidly to yield colored compounds that absorb between 500 and 700 nm (See Figure 2). Applying this method to wild-type clones arrayed in 96-well plates results in activity measurements with a standard deviation of only 5-9%, depending on the substrate used.

Plasmid name	Gene insert/vector	Promoter	Description
pJMJ2	todC1C2BAD/ptrc99A	Ptrc	Expresses toluene dioxygenase (todC1C2), electron transfer proteins (todBA) and dehydrogenase (todD)
pJMJ8	todC1C2BA/ptrc99A	Ptrc	Expresses toluene dioxygenase (todC1C2) and its electron transfer proteins (todBA)

Table 1. Plasmids used for expression of toluene dioxygenase cistron

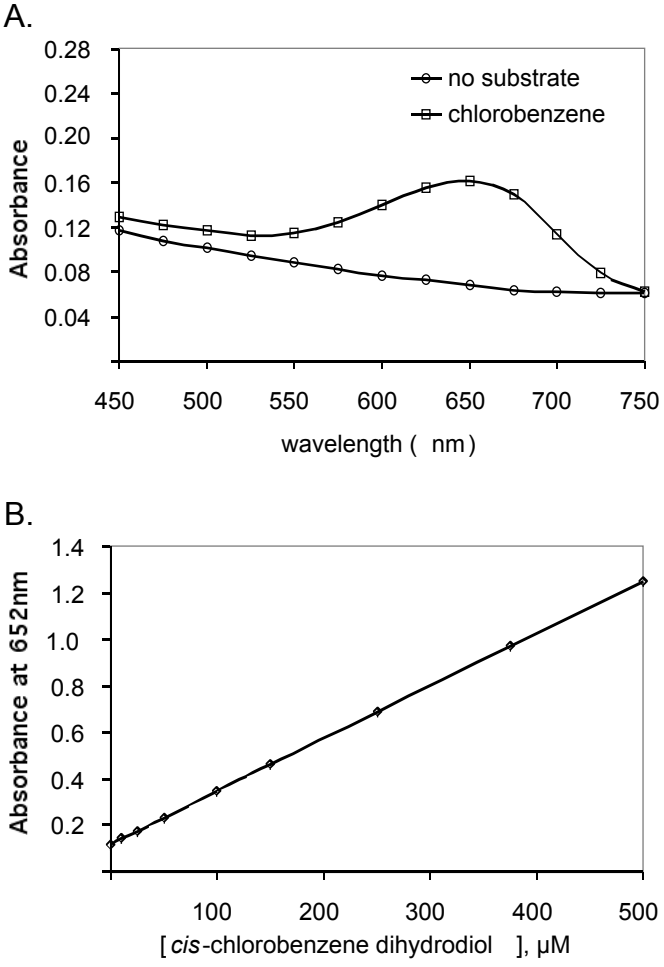


Figure 2. Validation of spectroscopic detection method. A. UV spectrum of colored products resulting from liquid-phase screening for toluene dioxygenase activity toward chlorobenzene. B. Plot showing linear correlation between chlorobenzene *cis*-dihydrodiol concentration and its absorbance at 652 nm after acidification and coupling with Gibbs reagent in minimal medium.

Solid-phase determination of dioxygenase activity

Dioxygenase activity can also be determined by screening freshly transformed colonies directly in the solid phase. This method reduces reagent usage and, more importantly, eliminates the time-consuming step of inoculating colonies into microtiter plates. This increases the throughput of the screen to about 10,000 clones/day. Because pH changes are difficult to effect on the solid phase, a different pathway, shown in Figure 1B, was used. The enzyme *cis*-dihydrodiol dehydrogenase (todD) was used to convert the *cis*-dihydrodiols to catechols that react readily with Gibbs reagent to yield colored compounds.

To screen for dioxygenase activity on the solid phase, *E. coli* are transformed with the appropriate plasmids and grown overnight under inducing conditions. Colonies are replicated on a nitrocellulose membrane and transferred to a minimal medium plate. This plate is then exposed to chlorobenzene vapor to allow the bioconversion to occur. The membrane is transferred to another agar plate containing Gibbs reagent, where a blue color quickly develops on the membrane under active colonies. About 500 colonies can be screened on one 15 cm plate.

Quantitation of dioxygenase activity

For directed evolution applications it is essential to quantitate the activity of individual clones. To this end, we have implemented a digital imaging and image analysis strategy for quantitation of dioxygenase activity on the solid phase. As shown in Figure 3, the final agar is imaged, and the image is filtered using local averaging to eliminate salt and pepper noise and Wallis filtering to remove any global contrast. Using a feature detection algorithm in the software package Optimas, individual colonies are selected and characterized based on their mean intensity, size and circularity. Since colonies not expressing dioxygenase do not yield any visible color, the difference between the colony mean pixel intensity and the intensity of the surrounding area is used as an activity metric. If desired, size and circularity statistics can be used to eliminate overlapping colonies.

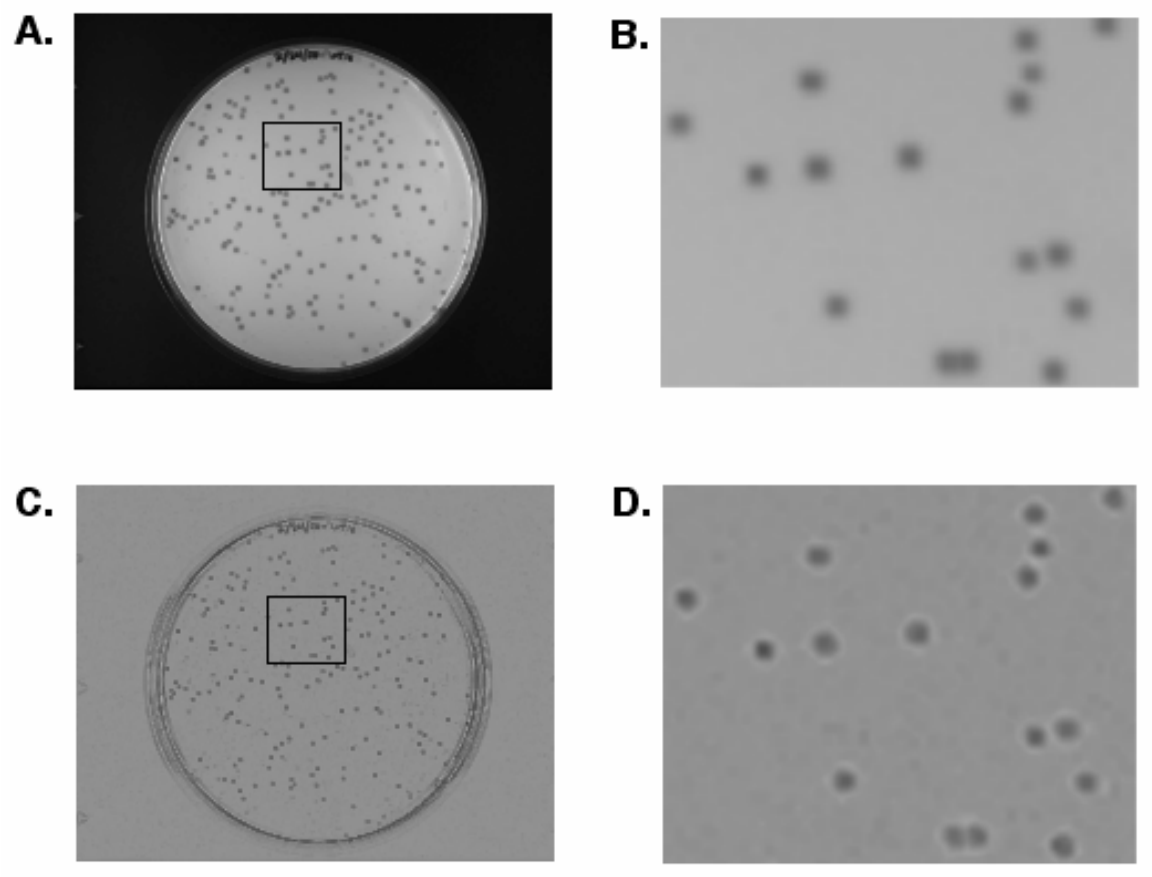
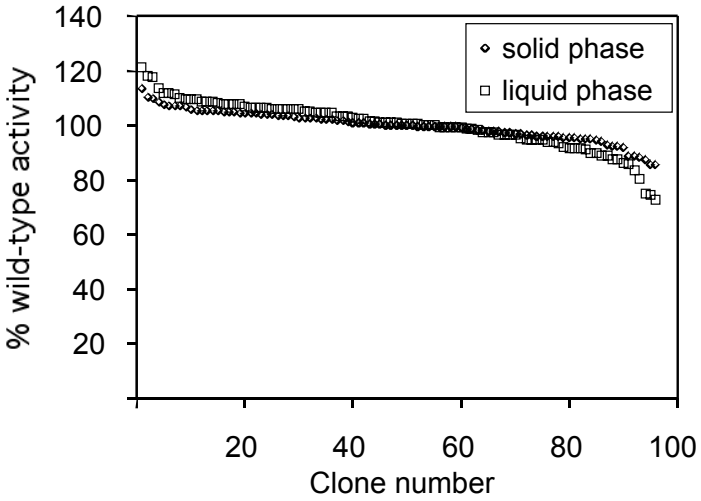


Figure 3. Image analysis. A. Unmodified digital image showing a screening result for colonies expressing wild-type pJMJ2 (toluene dioxygenase). B. Enlargement of squared region in A. C. Digital image after processing, as described in the text. D. Enlargement of squared region in C. The petri dish is 15cm in diameter.

Comparative evaluation of liquid- and solid-phase methods

BL21(DE3) cells were transformed with either wild-type pJMJ8 or pJMJ2, and 96 clones were screened for activity toward chlorobenzene using the liquid- or solid-phase method, respectively. The distributions of activity measurements for these two experiments are shown in Figure 4A. The standard deviation of activity measurements was 9.0% with the liquid-phase assay, and only 5.3% with the solid-phase method.

A.



B.

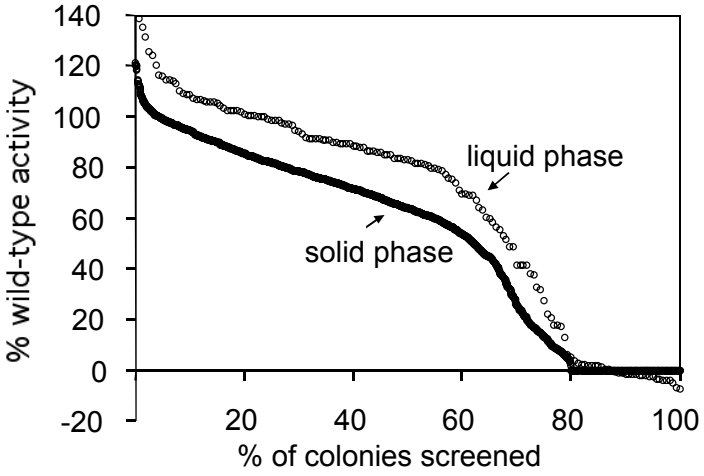


Figure 4. Validation of screening methods. A. Comparison of wild-type activity measurements from solid- and liquid-phase methods. Ninety-six colonies expressing wild-type toluene dioxygenase were screened using both methods. Activities are plotted in descending order. The standard deviation of activity measurements was 9.0% with the liquid-phase method and 5.3% with the solid-phase method. B. Comparison of mutant activity measurements generated by both the solid- and liquid-phase methods. Mutants were created by error-prone PCR applied to the gene coding for the large subunit of toluene dioxygenase. One hundred sixty mutants were screened with the liquid-phase method, and 1899 were screened with the solid-phase method.

As an additional validation of the solid-phase approach, a mutant library of toluene dioxygenase was screened using both methods. The gene for the large subunit of toluene dioxygenase was subjected to error-prone PCR and cloned into both pJMJ2 and pJMJ8. One hundred sixty mutant pJMJ8 clones and 1899 mutant pJMJ2 clones were screened for activity toward chlorobenzene using the liquid- and solid-phase methods, respectively. Figure 4B shows the activities of the clones plotted in descending order. Similar bulk library characteristics were found using the two methods (fraction inactive clones, fraction with wild-type-like activity). We thus conclude that the relative activities in a mutant library do not depend significantly on either the cell-growth method (on agar or liquid medium) or the specific method of detection (chemical or enzymatic, as shown in Figure 1B and 1C, respectively).

The screening methods described above are applicable to dioxygenase bioconversions of various aromatic substrates. These methods may also find application in screening for other types of aromatic oxidation that result in phenolic products, such as monooxidation, dealkylation, and oxidative dehalogenation. Quintana et al. [15] describe 14 phenols and catechols that react with Gibbs reagent to yield colored products. In general, most phenols with a good leaving group at the *para* position (-H, $-OCH_3$, halogens) will couple readily with Gibbs reagent [16].

Having high sensitivity and very low inherent variability, this screening method is ideal for directed evolution experiments. Often with directed evolution small improvements in function must be selected from a library and recombined to generate larger improvements [20,21]. With this screen, phenol or arene *cis*-diol concentrations of 30 μM can be reliably quantitated. This high sensitivity suggests that Gibbs reagent-based screening could be used to discover novel genes coding for oxidative enzymes in environmental samples. Also, the screen should be suitable for discovering gain-of-function variants created by random mutagenesis or recombination.

References

1. Butler, C.S., Mason, J.R. (1997). Structure-function analysis of the bacterial aromatic ring-hydroxylating dioxygenases. Adv. Micr. Phys. 38:47-84.
2. Gibson, D.T., Resnick, S.M. (1996). Diverse reactions catalyzed by naphthalene dioxygenase from *Pseudomonas* sp. strain NCIB 9816. J. Ind. Microbiol. 17:438-457.
3. Lange, C.C., Wackett , L.P. (1997). Oxidation of aliphatic olefins by toluene dioxygenase: enzyme rates and product identification. J. Bac. 179:3858-3865.
4. Gibson, D.T., Parales, R.E. (2000). Aromatic hydrocarbon dioxygenases in environmental biotechnology. Current Opinion in Biotechnology. 11:236-243.
5. Boyd, D.R., Sharma, N.D., Byrne, B., Hand, M.V., Malone J.F., Sheldrake, G.N., Blacker, J., Dalton, H. (1998). Enzymatic and chemoenzymatic synthesis and stereochemical assignment of *cis*-dihydrodiol derivatives of monosubstituted benzenes. J. Chem. Soc. Perkin Trans. 1:1935-1943.
6. Wackett, L.P. (1995). Recruitment of co-metabolic enzymes for environmental detoxification of organohalides. Environ. Health Perspect. 103:45-48.
7. Sheldrake, G.N. (1992). Biologically derived arene *cis*-dihydrodiols as synthetic building blocks. In *Chirality in Industry*. A.N. Collins, G.N. Sheldrake, and J. Crosby (eds.), pp.127-166. John Wiley & Sons Ltd, New York.

8. Wendeborn, S., De Mesmaeker, A., Brill, W.K.-D. (1998). Polymer bound 3,5-cyclohexadiene-1,2-diols as core structures for the development of small molecule libraries. Synlett. 8:865-868.
9. Jenkins, R.O., Stephens, G.M., Dalton, H. (1986). Production of toluene *cis*-glycol by *Pseudomonas putida* in glucose fed-batch culture. Biotech. Bioeng. 29:873-883.
10. Wahbi, L.P., Gokhale, D., Minter, S., Stephens, G.M. (1996). Construction and use of recombinant *E. coli* strains for the synthesis of toluene *cis*-glycol. Enz. Micr. Tech. 19:297-306.
11. Wilkinson, D., Ward, J.M., Woodley, J.M. (1996). Choice of microbial host for the napthalene dioxygenase bioconversion. J. Ind. Micr. 16:274-279.
12. Harrop, A.J., Woodley, J.M., Lily, M.D. (1992). Production of naphthalene-*cis*-glycol by *Pseudomonas putida* in the presence of organic solvents. Enz. Micr. Tech. 14:725-730.
13. Gibbs, H.D. 1927. Phenol tests III, the indophenol test. J. Biol. Chem. 72:649-664.
14. Zylstra, G.J., Gibson, D.T. (1989). Toluene degradation by *Pseudomonas putida* F1. J. Biol. Chem. 264:14940-14946.
15. Quintana, M. G., Didion, C., Dalton, H. (1997). Colorimetric method for a rapid detection of oxygenated aromatic biotransformation products. Biotechnol. Tech. 11:585-587.
16. Josephy, P.D., Van Damme, A. (1984). Reaction of gibbs reagent with *para*-substituted phenols. Anal. Chem. 56:813-814.

17. Sambrook, J., Fritsch, E.F., Maniatis, T. (1989). *Molecular Cloning*. Cold Spring Harbor Laboratory Press, United States of America.
18. Kumamaru, T., Suenaga, H., Mitsuoka, M., Watanabe, T., Furukawa, K. (1998). Enhanced degradation of PCB's by directed evolution of biphenyl dioxygenase. Nat. Biotech. 16:663-666.
19. Bruhlman, F., Chen, W. (1998). Tuning biphenyl dioxygenase for extended substrate specificity. Biotech. Bioeng. 63(5):544-551.
20. Moore, J.C., Arnold, F.H. (1996). Directed evolution of a *para*-nitrobenzyl esterase for aqueous-organic solvents. Nat. Biotech. 14:458-467.
21. Miyazaki, K., Wintrode, P.L., Grayling, R.A., Rubingh, D.N., Arnold, F.H. (2000). Directed evolution study of temperature adaptation in a psychrophilic enzyme. J. Mol. Biol. 297:1015-1026.

Chapter 6

A protocol for high-efficiency DNA shuffling

Preface

This chapter is to be published in the book **Methods in Molecular Biology** by Humana Press (F. H. Arnold and G. Georgiou, Eds.) under the title “DNA Shuffling.” The goal of this series is to publish protocols that, unlike the primary literature, provide enough detail to be followed with little difficulty the first time. Thus the protocol described here is written with frequent annotation and discussion of steps that may require optimization upon application to a new set of parent genes. I have constructed 14 chimeric libraries using different experimental conditions and dioxygenase parent combinations, and seven of these were characterized by Peter Meinhold and Lillian Pierce using the probe hybridization assay described in Chapter 7. This chapter is an attempt to encapsulate the knowledge gained from these experiments into a consensus protocol complete with useful discussion of each step.

Introduction

DNA shuffling is a method for *in-vitro* recombination of homologous genes invented by W.P.C Stemmer **(1)**. The genes to be recombined are randomly fragmented by DNaseI, and fragments of the desired size are purified from an agarose gel. These fragments are then reassembled using cycles of denaturation, annealing, and extension by a polymerase (See Figure 1). Recombination occurs when fragments from different parents anneal at a region of high sequence identity. Following this reassembly reaction, PCR amplification with primers is used to generate full-length chimeras suitable for cloning into an expression vector.

In several instances, chimeric enzymes with improved activity and stability have been isolated from libraries constructed using DNA shuffling **(2,3,4,5)**. In other cases, the method resulted in libraries with either too many mutations **(6)** or too few crossovers **(7)** to be useful. The DNA shuffling method we describe in this chapter is a hybrid of various published methods that has yielded highly chimeric libraries (as many as 3.7 crossovers per 2.1kb gene) with a low mutagenesis rate **(8)**. Fragments are made in much the same way as in the first Stemmer method **(1)**, the reassembly protocol is borrowed from Abècassis *et. al*. **(9)**, and *Pfu* polymerase is used throughout, as suggested by Zhao *et. al.* **(6)**. We have used this method successfully to recombine parents with only 63% DNA sequence

identity; however, more crossovers occur (and the library is more diverse) when the parent genes are more similar **(8)**.

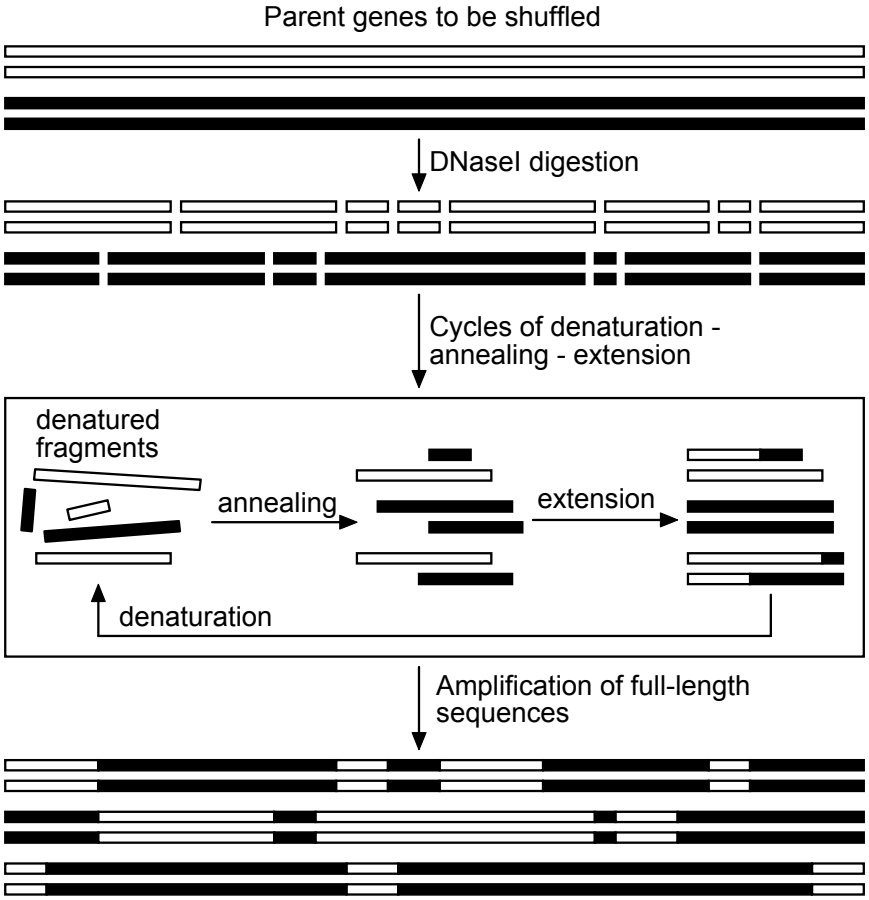


Figure 1. Schematic of DNA shuffling method. Parental genes are cleaved randomly using DNaseI to generate a pool of fragments. These fragments are recombined using PCR with a specialized thermocycling protocol. Fragments are denatured at high temperature, then allowed to anneal to other fragments. Some of these annealing events result in heteroduplexes of fragments from two homologous parents. Annealed 3' ends are then extended by polymerase. After 20-50 cycles of assembly, a PCR amplification with primers is used to selectively amplify full-length sequences.

Materials

1. Cloned *Pfu* polymerase and 10x buffer (Stratagene: La Jolla, CA)

2. PCR nucleotide mix, 10mM each (Promega: Madison, WI)

3. Dimethyl sulfoxide
4. MJ Research PTC-200 thermal cycler
5. 0.5M Tris-HCl, pH 7.4
6. 0.5M EDTA, pH 8.0
7. 0.2M manganese chloride
8. DNaseI, Type II, from bovine pancreas (Sigma: St. Louis, MO)
9. QIAquick gel extraction kit (QIAGEN: Valencia, CA) or equivalent
10. G25 Columns (Amersham Pharmacia Biotech, Inc.: Piscataway, NJ)
11. Two sets of primers. (See Note 1)
12. Parent DNA. The parent DNA should have large regions flanking the gene of interest so that "nested" primers can be used (see Note 1). A plasmid containing the gene of interest is ideal.

Methods

The procedures outlined below detail (1) obtaining DNA fragments from a DNaseI digestion, (2) reassembly of those fragments and (3) amplification of full-length sequences from the reassembly reaction.

Obtaining DNA fragments for shuffling

1. To get parent DNA for shuffling, mix in a PCR tube: 10μL 10x*Pfu* buffer, 2μL of PCR nucleotide mix, 40 pmol of each outer primer (See Note 1), 5U of *Pfu*

polymerase, 3μL DMSO, 0.08 pmol of template, and 75μL of water. Thermocycle using an annealing temperature appropriate for the outer primers. Extension should occur at 72°C for 2-3 minutes per kilobase of DNA amplified. 20-25 cycles are generally required.

2. Using a QIAquick gel extraction kit or similar spin column system, purify the PCR reactions. The DNA concentration after purification should be at least 40 μg/mL.
3. For the DNaseI fragmentation, prepare a solution of 0.167M Tris-HCl buffer, 0.0833M manganese chloride and 1.67U/mL DNaseI (See Note 2). In a separate tube, prepare 70μL of an equimolar mix of parent DNA with a concentration of 50-125μg/mL. Bring these solutions to 15°C in a thermocycler. At the same time, put 6μL of EDTA solution on ice in a microcentrifuge tube. Add 30μL of the buffered DNaseI solution into the parent DNA mix, and mix by pipetting several times. Incubate at 15°C for 0.5 to 10 minutes (See Note 3). To stop the reaction, transfer the solution to the tube containing EDTA and mix thoroughly.
4. Run the DNA fragments on an agarose gel containing ethidium bromide, and excise the desired size range (See Note 4). Purify the selected fragments using a QIAquick gel extraction kit or similar spin column system. The effluent should be further purified using a G25 column.

Reassembly of DNaseI fragments

1. To 42μL of purified fragment DNA, add 5μL of 10x*Pfu* buffer, 2μL of dNTP solution and 1μL of *Pfu*.
2. Cycle according to the following protocol: 96°C, 90 sec.;35cycles of(94°C,30 sec.;65°C, 90 sec.; 62°C, 90 sec.; 59°C, 90 sec.; 56°C, 90 sec.; 53°C, 90 sec.; 50°C, 90 sec.; 47°C, 90 sec.; 44°C, 90 sec.; 41°C, 90 sec.; 72°C, 4 min.); 72°C, 7min.; 4°C thereafter. (See Note 5)
3. Run 5μL of this reaction on an agarose/ethidium bromide gel. A smear of reassembled DNA that extends above the molecular weight of the parent genes should be visible.

Amplification of full-length sequences

1. Combine 10μL of 10x*Pfu* buffer, 2μL of dNTP solution, 40pmol of each inner primer, 3μL of DMSO, an aliquot (10-1000nL) of unpurified reassembly reaction (See Note 6), 5U of *Pfu*, and water to a final volume of 100μL.
2. Thermocycle using an annealing temperature appropriate for the inner primers. Extension should occur at 72°C for 2-3 minutes per kilobase of DNA amplified. 20-25 cycles are generally required.
3. Run 5μL of this reaction on an agarose/ethidium bromide gel. A band should be observed at the molecular weight of the parent gene. DNA should be purified by gel extraction prior to cloning into an expression vector.

Notes

1. Design of primer sets. Two sets of 18-25 bp primers with GC content ~50% should be designed in a “nested” configuration, i.e., the inner primers close to the gene of interest, and the outer primers ~150bp outside of the inner primers. The outer primers are used to amplify DNA for the fragmentation reaction, and the inner primers are used to amplify full-length sequences following the assembly reaction. Generally, when only one primer set is used, the amplification step to regenerate full-length sequences will fail. This might result from digestion or degradation of priming sites during the reassembly due to residual exonuclease activity from the polymerase.
2. Handling of DNaseI. DNaseI was dissolved in sterile water to a concentration of 10U/μL and stored at -20°C. An aliquot from a fresh 1:200 dilution was used to carry out the DNaseI digestion protocol.
3. Incubation with DNaseI. The incubation time with DNaseI is a critical parameter for generating fragments of the desired size. Before attempting this step, prepare enough parent DNA to digest small aliquots with varying incubation times (30 seconds to 10 minutes). Then select an optimal condition for a larger-scale digestion. In our hands, digestion of a 2.6kb gene for two minutes gave a size distribution of fragments centered at ~0.7 kb.
4. Selecting an appropriate fragment size. The first account of DNA shuffling reported selecting fragments in the range of 10-50bp **(1)**. This size range can be difficult to reassemble. Using this method, we have had success using fragments of 0.4-1kb to reassemble a 2.1kb gene and create a chimeric

library with 3.7 crossovers per gene **(8)**. Thus, if 10-50bp fragments do not reassemble successfully, using larger fragments may get the reassembly to go while still generating a sufficiently diverse library.

5. Temperature cycle during reassembly. Alternatively, a constant annealing temperature can be used for the reassembly. We had success annealing at constant temperatures ranging from 42°C to 58°C for 5min. For small fragments (~100-500bp) a higher annealing temperature (58°C) was required to eventually obtain a full-length product (2.1kb), but a set of large fragments (~200-1500bp) reassembled readily using either a 42, 50, or 58°C annealing temperature (unpublished results). Theoretically, more crossovers should occur when a lower annealing temperature is used, and we have in fact observed this experimentally (unpublished results).
6. Amplification of full-length sequences from reassembly reaction. In our experience, this step requires the most optimization. The amount of assembly reaction and the number of cycles are critical variables. We suggest varying the number of cycles from 20 to 25 cycles, and the amount of reassembly reaction from 1μL to 10nL per 100μL reaction. We have observed the counterintuitive result that if too many cycles (28-32) are used, a significant decrease in yield occurs. If too much reassembly reaction was added, a smear was observed upon running the reaction on a gel.
7. Mutagenesis rate. Using this method, we shuffled three parent genes of 2.1kb and sequenced 8 active chimeras and 10 inactive chimeras. Only two spontaneously generated mutations were found for a nucleotide mutation rate

of 0.011%. If mutations are desired in addition to recombination, error-prone PCR can be used in the first step to amplify parent DNA for the DNaseI digestion.

References

1. Stemmer, W.P.C. (1994) DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *PNAS* **91**, 10747-10751.
2. Chang, C.J., Chen, T.T., Cox, B.W., Dawes, G.N., Stemmer, W.P.C., Punnonen, J. and Patten, P.A. (1999) Evolution of a cytokine using DNA family shuffling. *Nat. Biotech.* **17**, 793-797.
3. Ness, J.E., Welch, M., Giver, L., Bueno, M., Cherry, J.R., Borchert, T.V., Stemmer, W.P.C. and Minshull, J. (1999) DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotech.* **17**, 893-896.
4. Christians, F.C., Scapozza, L., Crameri, A., Folkers, G. and Stemmer, W.P.C. (1999) Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nat. Biotech.* **17**, 259-264.
5. Bruhlmann, F. and Chen, W. (1998) Tuning biphenyl dioxygenase for extended substrate specificity. *Biotech. Bioeng.* **63**, 544-551.
6. Zhao, H. and Arnold, F.H. (1997) Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Res*. **25**, 1307-1308.
7. Kikuchi, M., Ohnishi, K. and Harayama, S. (1999) Novel family shuffling methods for the *in vitro* evolution of enzymes. *Gene* **236**,159-167.
8. Joern, J.M., Meinhold, P. and Arnold, F.H. (2002) Analysis of shuffled gene libraries. *J. Mol. Biol.* **316**, 643-656.

9. Abècassis, V., Pompon, D. and Truan, G. (2000) High efficiency family shuffling based on multi-step PCR and *in vivo* DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. *Nucleic Acids Res*. **28**, e88.

Chapter 7

Analysis of shuffled gene libraries

(Joern, J. M., Meinhold, P. and Arnold, F. H. (2002). *J. Mol. Biol.* **316**, 643-656.)

Abstract

In vitro recombination of homologous genes (“family shuffling“) has been proposed as an effective search strategy for laboratory evolution of genes and proteins. Few data are available, however, on the composition of shuffled gene libraries, from which one could assess the efficiency of recombination and optimize protocols. Here probe hybridization is used in a macroarray format to analyze chimeric DNA libraries created by DNA shuffling. Characterization of hundreds of shuffled genes encoding dioxygenases has elucidated important biases in the shuffling reaction. As expected, crossovers are favored in regions of high sequence identity. A sequence-based model of homologous recombination that captures this observed bias was formulated using the experimental results. The chimeric genes were found to show biases in the incorporation of sequences from certain parents, even before selection. Statistically different patterns of parental incorporation in genes expressing functional proteins can help identify key sequence-function relationships.

Introduction

Recombination is an effective search strategy for optimization problems in fields as diverse as molecular evolution, animal breeding, computer programming and economics [1,2]. During the laboratory evolution of biological molecules, recombination has been used to generate novel sequences in a process known as “family shuffling” [3-6]. In family shuffling, homologous genes are recombined *in vitro* or *in vivo* using one of a number of methods which include Stemmer’s DNA shuffling reaction [7,8]; staggered extension (StEP) [9], heteroduplex [10], random priming [11], and RACHITT [12] recombination; as well as *in vivo* methods [13-15]. The product is a library of hybrid, or chimeric, genes that contain sequence information from one or more of the parents.

Family shuffling represents a potentially powerful approach to generating novel sequences that encode functionally interesting proteins. Even when the homologous parent proteins differ at a large number of amino acids (as much as 30 or 40%), a significant fraction of the resulting chimeric proteins retain some level of function [4,6,16]. Thus recombination explores regions of sequence space that are distant from the starting proteins yet encode folded and functional proteins [3]. In contrast, comparably large jumps in sequence space made by random mutagenesis generate non-functional genes almost exclusively, due to cumulative deleterious effects of mutation and creation of stop codons. Recombination therefore efficiently exploits information present in the parental

sequences to assemble new, functional sequences. The assumption for laboratory evolution is that some measurable fraction of these novel, shuffled genes will express proteins with specific desirable traits.

It is unclear, however, how recombination should be performed so as to create libraries containing the most novelty. To evaluate this, we need to relate large numbers of sequence changes to changes in function. With this information we will be able to optimize shuffling protocols and compare recombination to other evolutionary search strategies such as random point mutagenesis. The usual practice of sequencing a small number of chimeric genes (and usually only the ones that show desired properties) leaves the researcher ignorant of key features of the library. We need to know, for example, the numbers and positions of crossovers in a statistically significant sampling of the library, both before and after selection. We also need to determine the percentage of sequences that are not recombinant, biases in locations of crossovers, and biases in incorporation of different parents, as well as how all these parameters affect fitness. Recently, Truan and coworkers described a multiple macroarray system based on annealing of radioactive oligonucleotide probes to preselected gene positions which allows rapid assessment of many of these factors [16]. When combined with additional functional information obtained by screening, these data from libraries of chimeric sequences will guide us in the best use of recombination for molecular optimization.

Here we describe the analysis of shuffled gene libraries encoding dioxygenase enzymes using two tools developed for this purpose. The first is a modification of the previously mentioned probe hybridization method [16] in which a set of labeled probes that anneal to specific parental gene positions is used to determine where sequences corresponding to the different parents appear in the chimeric genes. From these data, we estimate crossover positions and frequencies based on data from hundreds of clones. The second tool is a sequence-based hybridization preference model that can be used to predict biases in the distribution of crossovers in a shuffled library. Finally we discuss interpretation of the data generated by the probe hybridization experiments and by high throughput screening for function in the context of optimizing laboratory evolution and investigating sequence-function relationships.

Results and Discussion

Creation of family shuffled libraries

Two libraries were created by recombining genes encoding the α and β subunits of toluene dioxygenase (*todC1C2*), tetrachlorobenzene dioxygenase (*tecA1A2*), and biphenyl dioxygenase (*bphA1A2*) using a modification of Stemmer's method [7,16]. *Tod* and *tec* are 84.9% identical overall. The *bph* gene is less similar, exhibiting 63.1% and 63.9% sequence identity with *tod* and *tec,* respectively.

All three parents were used to make one library; only *tod* and *tec* were recombined for the second.

DNA sequencing results

Screening the clones from the three-parent dioxygenase library for activity towards toluene allowed us to divide the library into a toluene-active group (55 clones) and a toluene-inactive group (319 clones). Ten inactive and eight active clones were selected at random and sequenced. The results are summarized in Figure 1. The inactive clones contained 4.2 $\pm$ 0.8 crossovers on average and a range of 0 to 7, while active clones contained 3.8 $\pm$ 0.8 crossovers with a range of 1 to 8. In the 18 sequenced clones (~34,900 bp), only four point mutations (all transitions) arose during shuffling, a mutation frequency of 0.011% ($\pm$ 0.005%) or about 0.2 base substitutions per gene. Others have reported much higher point mutagenic rates for shuffling (0.05% [17], 0.7% [8], and 0.9% [16]), which makes it almost impossible to separate the functional consequences of the crossover and point mutation operations.

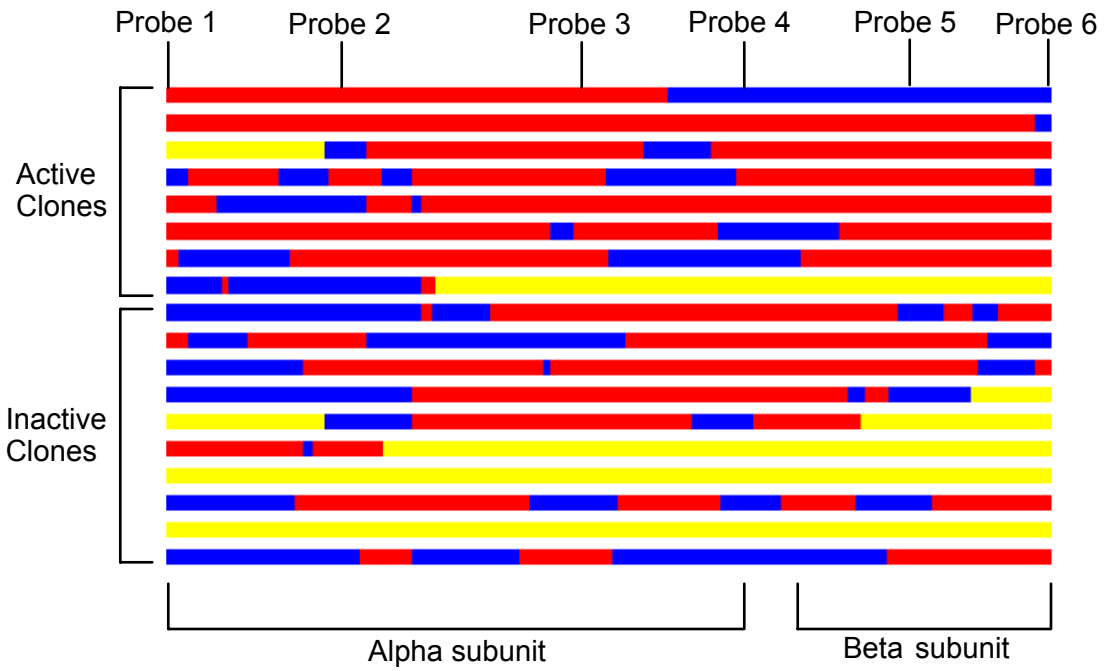


Figure 1. Sequencing results of 18 clones from the library made by shuffling genes encoding the α and β subunits of three dioxygenases. Horizontal colored bars represent the sequences of individual clones from the toluene-active or toluene-inactive subset of the library. Sequence elements from *todC1C2*, *tecA1A2* and *bphA1A2* are colored red, blue and yellow, respectively.

Because library construction relies on homologous recombination, crossovers are expected to occur preferentially where the parents share high sequence identity. Figure 2 compares the size distribution of regions of identity in the pairwise sequence alignments of the three parents to the size distribution of identical regions where crossovers occurred (See Figure 2a for an example of how these regions are defined). Figure 2b shows that while small regions of contiguous identity < 6bp are quite frequent in the sequence alignments (81%), the fraction of crossovers occurring in these regions is relatively low (21%). In contrast, while large regions of contiguous identity occur with relatively low frequency (7.3% for n > 10), a relatively high percentage (62%) of the crossovers take place in these regions.

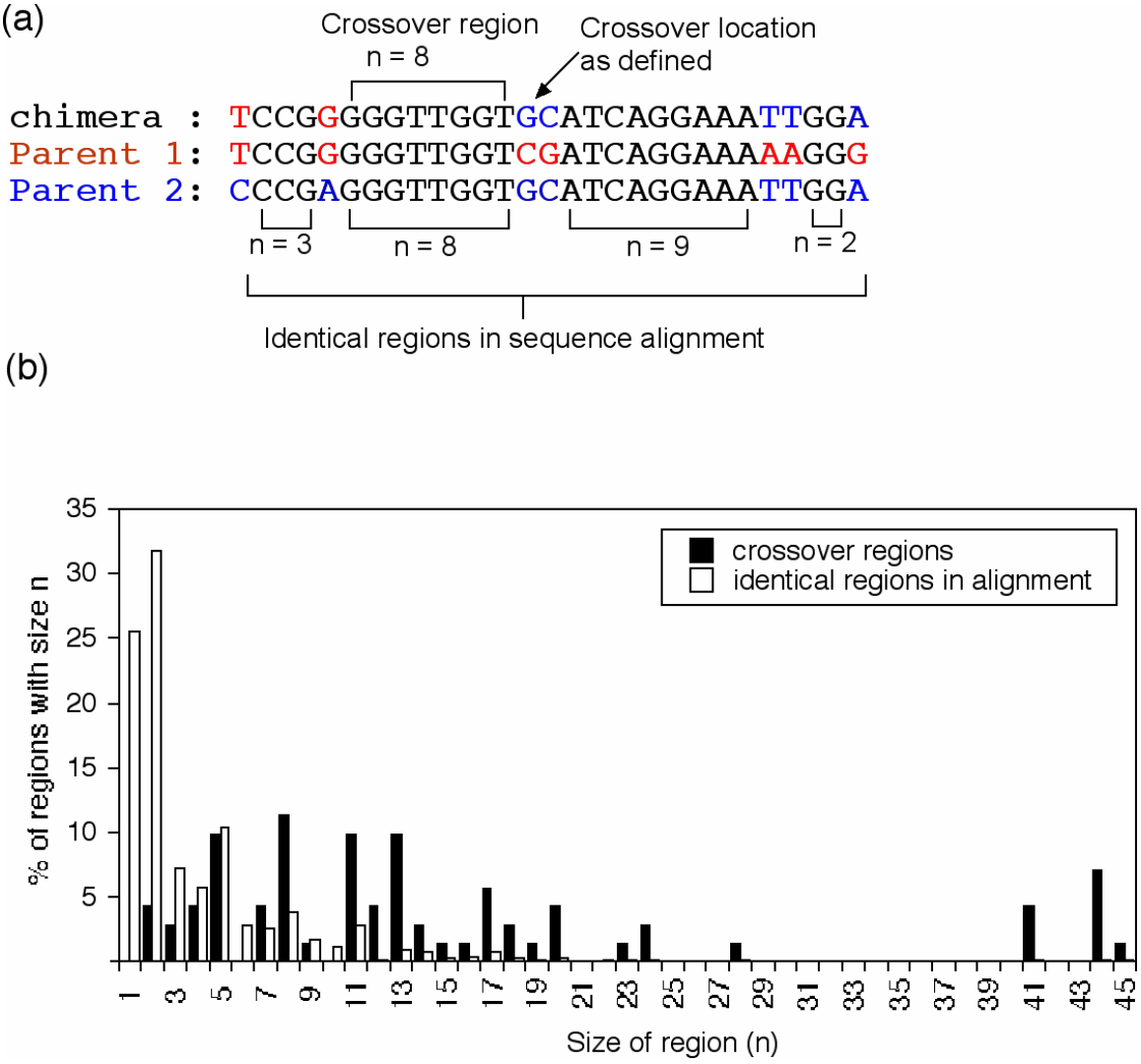
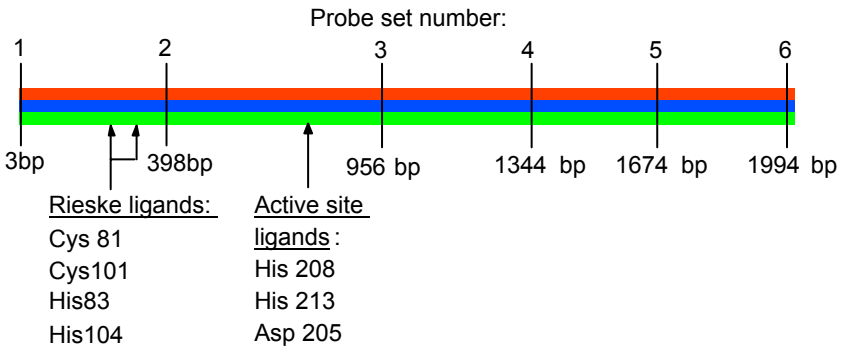


Figure 2. The size distribution of regions where crossovers occurred in 18 sequenced genes of the three-parent library, compared to the size distribution of identical regions in the sequence alignments. a) Sample section of a sequence alignment containing a crossover. A crossover has occurred between the second and third alleles shown, in a region of 8 bp. Because the exact crossover location can not be determined even by sequencing, it is defined as the first nonidentical base in the alignment of the upstream parent with the chimera. Identical regions in the sequence alignment are defined as the region between two alleles. b) Distribution of the lengths of the crossover regions for the 71 crossovers and the lengths of identical regions (1118 total) in the pairwise alignments of the three parent genes.

Probe hybridization characterization of shuffled gene libraries

To characterize the shuffled gene libraries, labeled oligonucleotide probes were designed to anneal specifically to one parent and thereby determine the identity of the parent at that position. Probes of 19-25 nt were roughly equally spaced over the ~ 2100 bp genes at six positions (Figure 3). The parent genes within the target annealing region differed at not less than three positions. Choosing probe positions with three or more mismatches simplifies optimization of the protocol, as does designing the probes such that their annealing temperatures at all positions are approximately equal. An antibody-alkaline phosphatase complex is used to detect bound label by display of chemiluminescence after free probe is washed away.



Probe set/parent	Probe sequence (5' to 3')
1 / *tod*	GAATCAGACCGACACATCACC
1 / *tec*	GAATCACACCGACACCTCC
1 / *bph*	GAGTTCAGCAATCAAAGAAGTGC
2 / *tod*	CTTACGAGGCCGAATCCTTCG
2 / *tec*	CCTTCGAGGCTGAATCCTTCC
2 / *bph*	CGTGCCGTTCGAGAAGGAAG
3 / *tod*	CCTTCCTCCCAGGTATCAATACG
3 / *tec*	CTTCCTTCTAGGCGCCAACAC
3 / *bph*	CATTCCTGCCCACCTTCAAC
4 / *tod*	GACACGCTGAATCCAGAGACAG
4 / *tec*	CACACGCTGAATCACGACAC
4 / *bph*	CCTGATCAAGACGCAATCGTTAG
5 / *tod*	GAATACTCAGGCTCCCGAGAG
5 / *tec*	CTGGAGTACTCGGGCACC
5 / *bph*	GAGCTGGAATATTCCGGCGAC
6 / *tod*	CATCCTGGCCAATAACCTCAGTTTC
6 / *tec*	TGGCGAACAACCTCAGCTTC
6 / *bph*	GCTGTCGAACAACCTGAGCATG

Figure 3. Positions of oligonucleotide probe sets. Cofactor ligands are also indicated, based on *todC1C2* sequence. Six sets of three oligonucleotide probes were designed such that each probe binds specifically to its parent gene and all probes bind with a calculated T_m of ~62°C.

For each shuffled dioxygenase library, two 384-well plates containing chimeric clones and the parents (6 wells) were analyzed. One plate contained clones picked randomly (unselected library) and the other contained only clones that showed activity toward indole (selected library), as determined by a colony assay for indigo formation (See Materials and Methods). Parental clones and empty

wells were used as controls. Clone-probe combinations that gave a chemiluminescent signal were assigned “true,” and ones that did not were assigned “false.” A position can generate an ambiguous result when there is partial probe mismatch due to PCR-induced point mutations, if a crossover occurs within the binding region of the probe, if the clone contains more than one plasmid, or if more than one colony is transferred into a single well of the 384-well plate. No result is obtained when single colonies do not grow on the membrane; it could also be the consequence of a point mutation or crossover within the probe-binding region. In our experience these problems are user- and system-dependent and can generally be resolved by altering colony growth conditions and by optimizing hybridization and wash temperatures. For the libraries analyzed in this study, 96.3% of the positions gave unambiguous results, 2.0% were ambiguous, and 1.2% gave no result.

Average number of crossovers

By counting the number of instances where neighboring probe sites are occupied by different parents, we measured 1.77 $\pm$ 0.07 crossovers/gene for the unselected three-parent library and 2.11 $\pm$ 0.07 for the unselected two-parent library. Because two or more crossovers can be hidden between probes, however, these numbers are significantly smaller than the number of crossovers found by sequencing. To better estimate the actual number of crossovers n_c, we developed an equation that relates the probability P^m_{abX} of observing parent *a* at

probe position *X* and parent *b* at probe position *X* + 1 to the probabilities P_{abX} that nucleotide *x+1* is from parent *b* given that nucleotide *x* is from parent *a* between probes *X* and *X*+1. (See Appendix for explanation and calculations.)

Table 3 and Figure 6 show the results of applying this method to calculating crossover frequencies for the two libraries. For the three-parent library, our estimate of 3.65 $\pm$ 0.25 crossovers/gene agrees with the sequencing results (4.20 $\pm$ 0.79 for unselected clones) and is considerably higher than the 1.77 $\pm$ 0.07 observed crossovers. For the two-parent library, the estimated number of crossovers is 5.04 $\pm$ 0.18, compared to only 2.11 $\pm$ 0.07 observed crossovers.

The probe hybridization data can provide an accurate estimate of the number and positional distribution of crossovers if a sufficient number of probes is used. When two or three parents are recombined, the required number of probes is roughly equal to 1.25 times the average crossover number. At average crossover numbers between two probes above about 1.25, the probe hybridization results will not change significantly even though there are more and more actual crossovers. To investigate the relationship between the actual number of crossovers and the number of observed crossovers, we simulated the construction of chimeras from different numbers of parents, assuming that each parent was incorporated to an equal extent and crossovers between different pairs of parents occurred with equal frequency. As shown in Figure 4, the observed number of crossovers saturates at the expected value of $(n_p - 1)/n_p$ (n_p

= number of parents). As the curve begins to saturate, small errors that result solely from clone sampling in the number of observed crossovers give rise to larger errors in the actual number of crossovers.

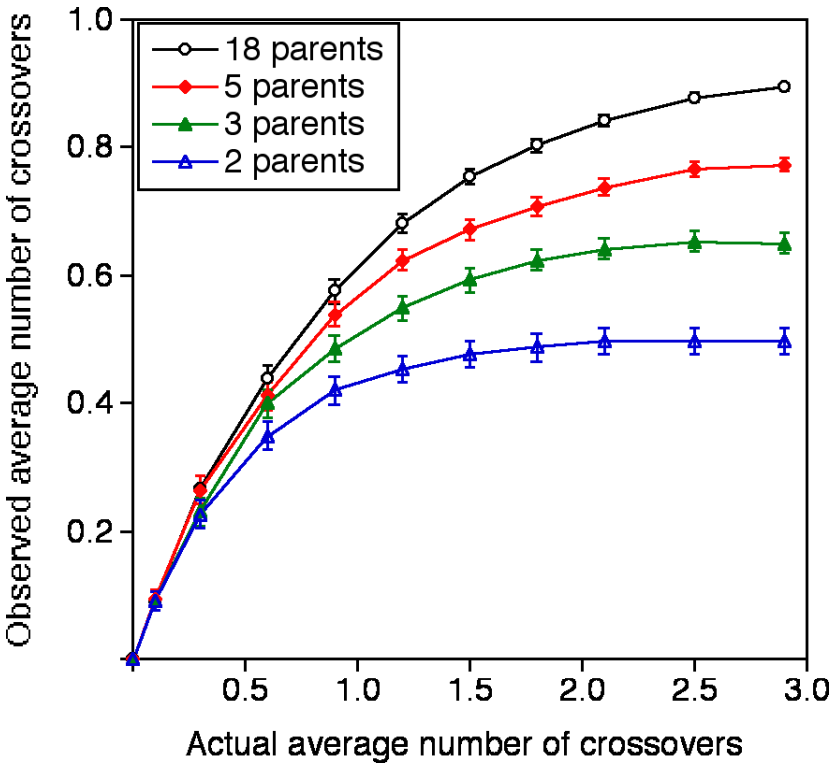


Figure 4. The observed average number of crossovers (the number directly apparent from probe hybridization data) between two probes plotted against the actual average number of crossovers for two probe sites separated by an arbitrary sequence length for different numbers of parents. For the purposes of this simulation, crossovers involving each pair of parents were assumed to occur with equal frequency, and parents were incorporated to an equal extent. Error bars are standard deviations assuming a sampling of 300 clones; errors are inversely proportional to the square root of the number of clones sampled.

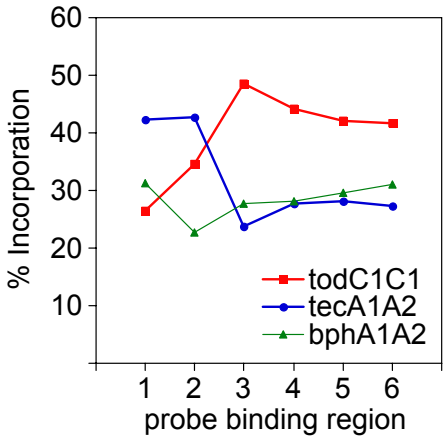
Figure 4 can be used as a guide for setting up a probe hybridization experiment when the actual number of crossovers is roughly known. As a hypothetical case, consider three parent genes of 1500 bp sharing a similar percent identity that are shuffled under conditions where the average crossover number could be as high

as five per gene. From Figure 4 we see that the actual number of crossovers can be determined with reasonable accuracy when fewer than 1.25 crossovers occur between neighboring probes. Thus four sets of neighboring probes (5 probes total) are required (4 = maximum average crossovers (5)/1.25). Therefore, if ~300 clones are assayed at five probe positions, the actual number of crossovers can be determined with good accuracy. When the parent genes do not have similar percent identity as in this hypothetical case, additional probe positions should be used.

Biases in parental incorporation

Of the detected probe signals, 39.6% were *todC1C2*, 32.0% were *tecA1A2* and 28.4% were *bphA1A2*. That the overall parental incorporations differ slightly from the expected 33% could be the result of unequal concentrations of the DNAseI-fragmented parental DNA fragments in the shuffling reassembly reaction. Interestingly, however, the parental incorporation also varied from region to region (Figure 5a). The hybrid library is heavily biased towards *tecA1A2* at the 5'-end and towards *todC1C2* at probe 3. This bias was even more pronounced in the two-parent (*todC1C2* and *tecA1A2*) library (Figure 5b), in which only 28.5% are *todC1C2* at position 1, even though the library is 57.9% *todC1C2* overall. In a similar library analysis, Abècassis *et al.* [16] reported the same frequency for all analyzed sequence segments, in contrast to our observations. Thus, biases in parental incorporation may depend strongly on the genes that are shuffled.

a.



b.

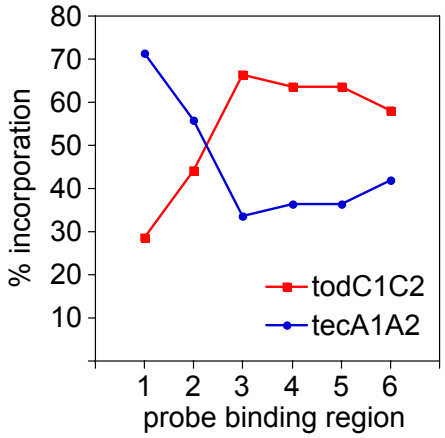


Figure 5. Incorporation of parent sequences at different probe positions in the (a) three-parent dioxygenase library (306 clones) and (b) two-parent library (317 clones) is biased towards *tecA1A2* at the 5'-end of the gene (probe positions 1 and 2) and towards *todC1C2* at the 3' half (positions 3 to 6). Sequences from *bphA1A2* are distributed homogeneously in the three-parent library.

Unequal amplification [18] or cloning efficiency as well as sequence-dependent variations in DNaseI digestion could bias the shuffling reaction towards one or more parents. In fact, when we amplified *todC1C2* and *tecA1A2* in a standard PCR, the *todC1C2* reaction gave a higher yield. This may be due to the fact that

tecA1A2's overall GC-content is 2.5% higher than that of *todC1C2*. This difference does not vary significantly along the genes, however, and therefore does not explain the observed positional bias. Also, the genes shuffled by Abècassis *et al.* [16] differed in their GC-content by 5.2%, but little bias in parental incorporation was observed.

Another source of the observed positional bias could be preferential elimination of genes encoding proteins with *tecA1A2* at the C-terminus during the cloning procedure [19]. A simple experiment supports this. *E. coli* BL21(DE3) were transformed with the same amounts of plasmids pJMJ2 (containing *todC1C2*) and pJMJ6 (containing *tecA1A2*) and plated out onto (a) LB-agar plates containing 100 μg/ml ampicillin and 1 mM IPTG to induce protein expression and (b) LB-agar plates only containing ampicillin. No pJMJ6 transformants were found on the IPTG plates, while the pJMJ2 transformation yielded approximately 1,000 transformants. The same transformation mixtures plated without IPTG yielded approximately 1,000 transformants in both cases. Thus the presence of tetrachlorobenzene dioxygenase encoded by *tecA1A2* seems to inhibit growth of *E. coli* BL21(DE3). Since leaky expression of the protein in the absence of IPTG does occur, a small amount of TCDO could bias parental incorporation, and this could be position dependent. In both shuffled libraries, *tecA1A2* is favored at the first and second probe positions and disfavored at positions 3-6. If the toxicity of the *tecA1A2* gene product is concentrated in the C-terminal portion, we could expect to see the observed parent incorporation pattern.

Small variations in the fraction of each parent in the initial DNAseI digestion could also bias parental incorporation. During reassembly, fragments from a parent present at relatively high concentration have greater opportunity to anneal and extend. Over dozens of cycles, the concentration of DNA increases several-fold, and since annealing events between fragments from the same parent are favored, this additional DNA will come preferentially from the parent with higher initial concentration. This autocatalytic mechanism has the effect of geometrically increasing the initial variation in the reassembly mixture. Furthermore, under conditions where many of the fragments do not grow to full length, the subset of full-length sequences will be more biased than the pool of fragments as a whole, due to preferential extension of fragments with the high-concentration parent at their 3' ends. The incorporation biases we observe probably result from some combination of these factors.

Frequency of wild-type genes in the shuffled library

In the three-parent dioxygenase library, 19.7% of the clones had hybridization patterns that did not reveal any crossovers. Of these, the majority were *bphA1A2* (76%), followed by *todC1C2* (19%) and *tecA1A2* (6%). *BphA1A2* has relatively low sequence identity with the other parents (Table 1) and experiences a disadvantage with respect to recombination with the other two parents *(vide infra).* Having fewer favorable recombination points with the other genes

promotes reassembly of wild-type *bphA1A2*. When *bphA1A2* was not included in the shuffling reaction, the frequency of parental hybridization patterns was reduced to 6.4%, of which 65% were *todC1C2* and 35% were *tecA1A2*.

	Probe interval					
Parent Pair	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**Overall**
tod - tec	84.9%	85.1%	87.6%	80.2%	87.5%	84.9%
tod - bph	66.9%	65.9%	67.0%	52.3%	63.2%	63.1%
tec - bph	66.9%	64.8%	68.6%	54.1%	66.0%	63.9%

Table 1. DNA sequence identity for parent genes used in this study.

Crossover biases in DNA shuffling

Significant biases in where crossovers occur or in which parents are involved can limit the accessible genetic diversity and affect the molecular evolution search process. We have observed biases in parental incorporation and in reassembly of parental sequences, as discussed above. We also expect bias in the crossover locations and in which parents are most likely to recombine. Because the *in vitro* recombination method reassembles the genes by overlap extension, it is expected that crossovers will occur preferentially between the most similar parents in regions of high sequence identity. Table 1 shows the sequence identity shared by the dioxygenase parents between the different sets of neighboring probes. From the probe hybridization data, we calculated the average number of crossovers between nearest-neighbor probes (Figure 6). For the three-parent library, crossovers between *bphA1A2* and the other two parents were highly disfavored, especially between probes 4 and 6, where sequence

identity is lowest (Table 1). For the library made from two parents, crossovers were approximately evenly distributed over all probed regions.

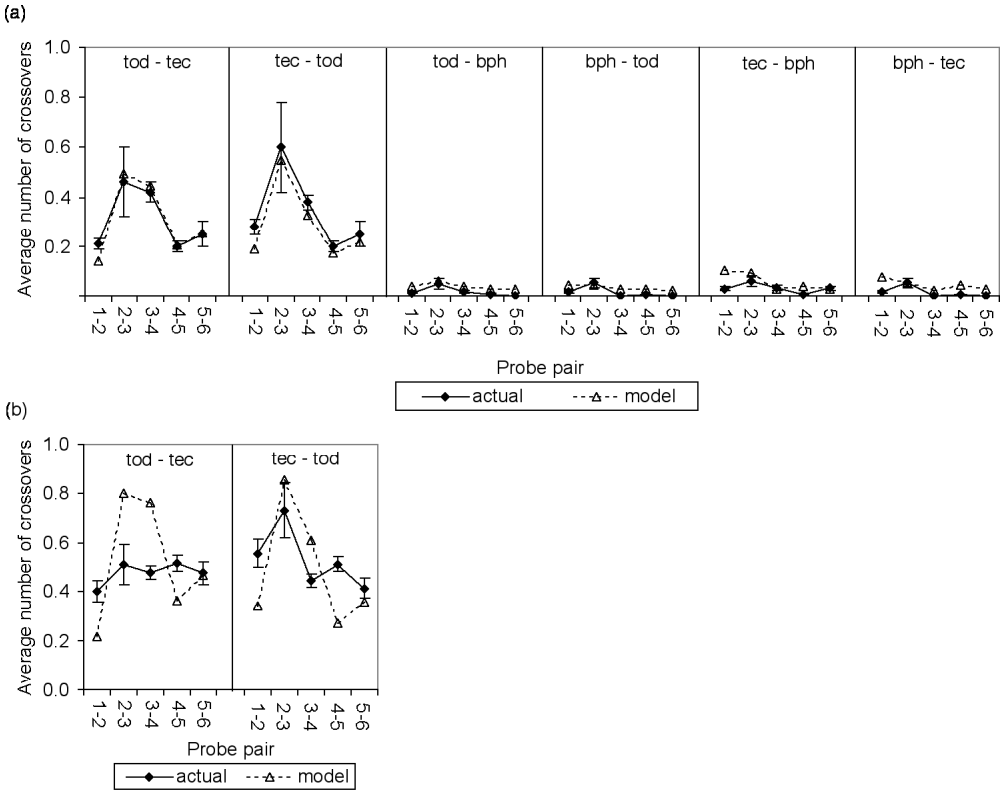


Figure 6. The number of crossovers N_{abX} after correction of the probe data ("actual") is compared to the N_{abX} predicted by the model using eqn. 1 ("model") for all types of crossovers. (a) Three-parent dioxygenase library. The average number of crossovers from *todC1C2* to *tecA1A2* within the region between the probes on the x-axis is plotted in the box labelled "tod – tec". The solid line represents values obtained after correction of the probe data and the dashed line shows the model prediction for a = 1.6 and T = 41°C. **(**eqn. 1**).** (b) Two-parent library. The error bars represent approximately one standard deviation and are based only on sampling error.

A sequence-based model for homology-dependent recombination

To formalize the apparent correlation between likelihood of crossover and sequence identity, we developed a simple sequence-based model to calculate the probability P_{abx} that a sequence corresponding to parent a will cross over to

parent *b* at nucleotide *x*[†]. We assume this probability is proportional to the Gibbs' free energy change upon duplex formation between nucleotides from parents *a* and *b* around position *x* (ΔG_{abx}) (eqn. 1). Because this proportionality need not be linear, the exponential parameter α (fit to a value of 1.6 for this study) is used to tune the model. To calculate the free energy, the model of Sugimoto *et al.* [20] (with no self-complementarity contribution) is applied to the region of maximal overlap without a mismatch on the upstream side of the position under consideration. To simulate the construction of a chimera, the first nucleotide is parent *a* with probability equal to the fraction of parent *a* in the library, and crossovers occur to other parents with probability P_{abx} at subsequent positions. The number of each type of crossover is averaged over a few thousand constructs. n_c is the average total number of actual crossovers that occur in *L* nucleotides. The parameter β is included to adjust the simulated total number of crossovers to n_c. When the parents are present at equimolar concentrations, β = 1; when they are not, β must be increased above 1. The number of parents (n_p) is also included.

$$(1)\quad P_{abx} = \frac{(\Delta G_{abx})^{\alpha}}{\sum_{a\neq b}\sum_{b\neq a}\sum_{k=1}^{L}(\Delta G_{abk})^{\alpha}} \bullet n_p \bullet n_c \bullet \beta$$

Crossover is not allowed (P_{abx} = 0) at positions where the region of overlap is 1-2 bp or the free energy change upon duplex formation is positive. To validate and

[†] P_{abx} is a function of sequence position x, whereas the P_{abX} variable discussed previously is constant over

tune the model, we compared the model prediction for the average number of crossovers to the values obtained by correcting the probe hybridization data from the two unselected libraries (N_{abX}). Simulated chimeras were constructed in sections corresponding to the regions between probe positions by taking the first nucleotide from parent *a* with probability P^m_{aX} for the upstream probe (the probability that a clone has parent *a* at probe position *X*), and allowing crossover to other parents at subsequent positions with probability P_{abx}. To capture the positional bias we observed for parental incorporation (Figure 5), we calculated the P^m_{aX} values from the probe hybridization data. For the simulation, we used the lowest annealing temperature from the actual reassembly (41°C) and constructed 4,000 chimeras *in silico* as described above. For n_c, we used values of 3.65 and 5.04 for the three- and two-parent libraries, respectively, as determined by correcting the probe hybridization data, and $\beta = 1$. Setting the parameter α to 1.6 optimized the fit to the available data.

Figure 6 compares the number of crossovers between neighboring probe pairs predicted by application of Eqn. 1 to the number found by correcting the probe data for multiple crossovers. For the three-parent library, the model predicts the bias against crossovers involving *bphA1A2* and fits the data remarkably well for crossovers involving only *todC1C2* and *tecA1A2* (Figure 6a). For the two-parent library, however, the correlation is much weaker (Figure 6b). We do not know the reason for this.

---

the region from probe position X to X+1. When P_{abx} is averaged over x, the value is similar to P_{abX}.

Figure 7 compares the actual crossover points determined from the 18 sequenced clones (Figure 7a) to the relative probabilities of crossover according to the model (Figure 7b). The crossover position is defined as the first base coming from a new parent when reading from 5' to 3', with 1 being the start of translation. Some sequence positions with high probability density according to the model correspond to positions with a high frequency of crossovers in the sequenced clones (e.g., positions 600, 621 and 2048). Thus, for the three-parent library, the model predictions are roughly consistent with both the sequence-level and probe-level results.

Overall, our results show that crossovers are strongly favored in regions of high sequence identity. Because crossovers occur frequently in regions of 5-8 bp of identity (Figure 2) where there is high variability in GC content and hence free energies of duplex formation, sequence identity itself is not useful for evaluating individual crossover sites. The free energy model allows us to treat the correlation between sequence identity and probability of crossover quantitatively. Our model should be useful for identifying preferred crossover sites and estimating relative frequencies of crossovers for particular regions. The more challenging problem of modeling the recombination of homologous genes with the goal of predicting the number and distribution of crossovers is an active area of research [21-23].

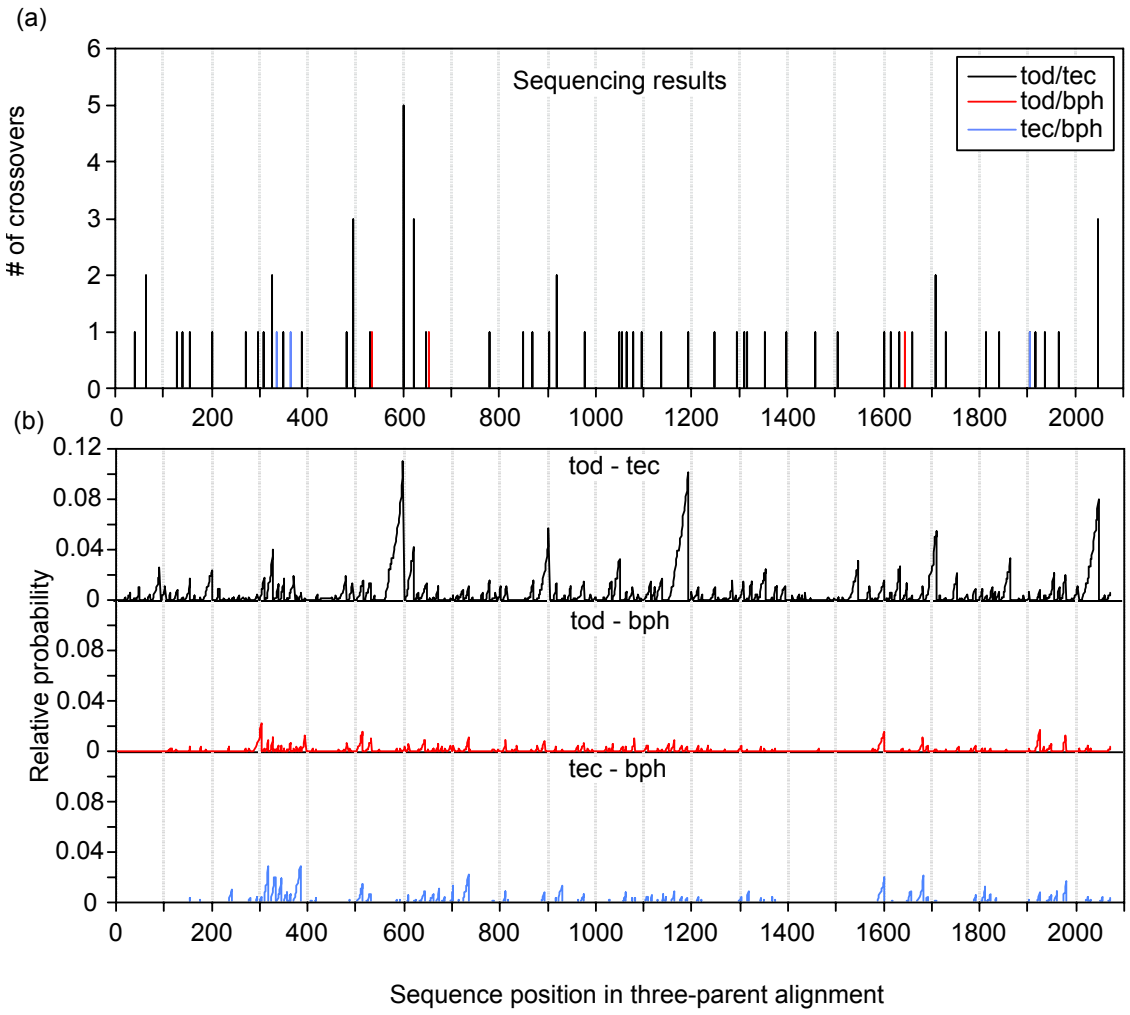


Figure 7. Model evaluation for predicting preferred crossover points. (a) Number of crossovers observed at each sequence position by DNA sequencing of genes from 18 clones. (b) Relative probabilities of crossover calculated according to eqn. 1, with *a* = 1.6, plotted against sequence position. The plots labelled *tod - tec*, *tod - bph*, and *tec - bph* show the expected relative probabilities of the three types of possible crossovers.

Recombination and protein function

Of the recombined dioxygenases, only TDO shows high activity on toluene; TCDO's activity is approximately 10% that of TDO, and BPO has no activity on this substrate (Table 2). Screening showed that 15% of the three-parent library

and 20% of the two-parent library retained at least 15% of wild-type TDO activity toward toluene. Less than 4% of the three-parent library is made up of TDO-like sequences, thus at least 11% of the shuffled dioxygenases are chimeras, primarily with TCDO, that are active towards toluene.

	Relative activity toward:	
Parent	**Indole**	**Toluene**
TDO	*****	*****
TCDO	*****	*
BPO	not active	not active

Table 2. Relative activities of three wild-type dioxygenase parents toward indole, as determined by indigo visualization, and toward toluene, as determined by solid-phase quantitative screening.

Both TDO and TCDO are positive in the indole assay based on indigo formation. This assay, while convenient, is less sensitive and less reproducible than the toluene assay. When the three- and two-parent libraries were screened for indole activity based on visible indigo formation, 16% and 25% of the colonies were indole-active, respectively.

For the two-parent library, neither an inactive parent nor point mutations can account for the 75% inactivation we observe. We considered two properties that could possibly correlate to loss of function in a library of recombined genes, i) the average number of crossovers and ii) the average fraction of sequence contributed by the parent with the highest representation in each clone (fraction of dominant parent). To examine how the number of crossovers affects function, we compared the crossover numbers for the unselected library to the numbers

for the subset showing activity toward indole (selected library). We found that clones from the selected library have the same number of crossovers on average as the library as a whole (see Table 3). At high point mutation rates (usually > 3 per gene), functional genes tend to have fewer point mutations than the library average [24]. Our data do not support a corresponding relationship between increasing crossover number and retention or loss of function, which indicates that increasing crossover frequency is not deleterious (or beneficial) to function, at least at the average crossover frequency characteristic of these libraries.

	3-parent library		2-parent library	
	Unselected	Selected	Unselected	Selected
Probe data				
Measured average number of crossovers	1.77±0.07	1.87±0.07	2.11±0.07	2.17±0.07
Corrected average number of crossovers	3.7±0.3	3.8±0.3	5.0±0.2	4.9±0.2
Sequencing				
Average number of crossovers	4.2±0.8	3.8±0.8	N/D	N/D

Table 3. Comparison of the average number of crossovers for unselected clones and clones selected for activity toward indole. The measured average number of crossovers is determined directly from the probe hybridization data and corrected for multiple crossovers between probes as described (see text). For the three-parent library, sequencing data provides a validation of this method. We observe no statistically significant difference in the average number of crossovers for the subset of the chimeric library that is functional.

For the two-parent library, the fraction of dominant parent was estimated for each clone by counting the number of probe positions occupied by the parent present at the most positions. As shown in Figure 8, for the unselected library the distribution of the number of probe positions (n) occupied by the most prevalent parent is close to n = 3 and 4, as would be expected for random incorporation of parental sequence. This distribution shifts, however, toward n = 5 and 6 for the

selected library. Thus, clones with a high percentage of sequence from a single parent are more likely to be active than clones with a more equal amount of information from both parents.

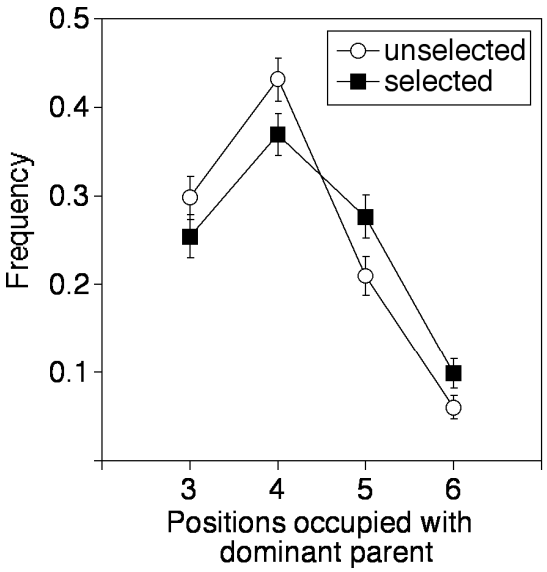


Figure 8. Distribution of the number of probe positions occupied by the dominant parent (the parent present at the most probe positions out of six total) for the two-parent unselected and selected libraries. Unselected and selected libraries comprise 317 and 318 clones, respectively. Error bars represent one standard deviation based solely on sampling error. In the selected library, the frequency of clones with 5 or 6 positions occupied by the dominant parent is significantly higher than it is for the unselected library.

We find it useful to think of chimeras as being inactivated by disruption of interactions that contribute to proper folding, stability or activity. The term schema disruption describes the extent to which a crossover disrupts beneficial sequences, analogous to its use in computer science and optimization by genetic algorithms [25]. Voigt *et al.* propose that schema disruption in proteins can be estimated from the 3-dimensional structure by counting the number of interactions jeopardized by a particular arrangement of crossovers [26]. This view is consistent with the observation that clones with a higher fraction of

dominant parent are less prone to inactivation, since such clones will, on average, conserve more interactions than the library as a whole, regardless of how the interactions are arranged or defined.

Identification of important functional regions

Although crossover number does not strongly influence function in the chimeric libraries, clones from the selected libraries have hybridization *patterns* that are markedly different from their unselected counterparts. The selected and unselected clones from the three-parent dioxygenase library were sorted by their relative activities toward toluene and plotted as a heat map in Figure 9. Two features emerge from this analysis. First, although fragments from *bphA1A2* made up 28.4% of the unselected library, only 7 of 266 active clones contained some *bphA1A2* sequence according to the probe analysis. This observation is consistent with the relative activities of the parents (the wild-type *bph* construct shows no activity toward toluene or indole, Table 2) and with the limited incorporation of *bphA1A2* into chimeric sequences.

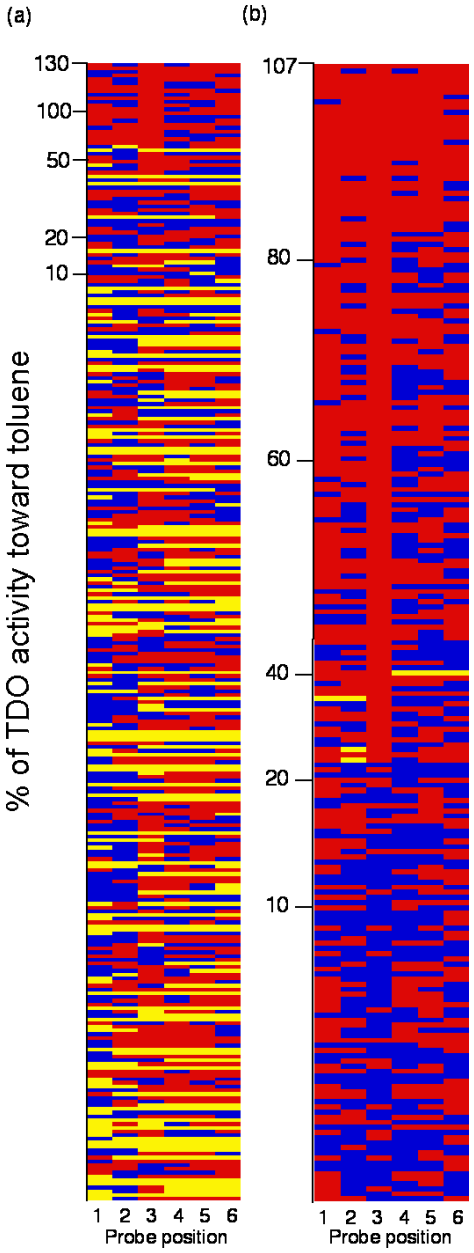


Figure 9. Probe hybridization patterns sorted by relative activity towards toluene and plotted as a heat map using the program Spotfire® (Spotfire, Inc., Cambridge, MA), wildtype *todC1C2* (toluene dioxygenase) has an activity of 100%. *TodC1C2* is colored red, *tecA1A2* blue and *bphA1A2* yellow. (a) The pattern of an unselected library (306 clones) shows a random distribution of all three parents below 20 % relative activity. (b) In the selected library (223 clones), probe position three is biased towards *todC1C2* and *bphA1A2* is essentially absent.

The second feature is that the *todC1C2* parent (which has relatively high activity toward toluene) is overwhelmingly favored at probe position 3 and slightly favored at position 1 in clones that are highly active toward toluene. Because the chimeric genes for the dioxygenase were coexpressed with the electron transfer proteins from toluene dioxygenase, we expected that active clones might be biased toward incorporation of the *todC1C2* parent to optimize interactions with the electron transfer proteins that are required for activity. The crystal structure of naphthalene dioxygenase [27], which shares 28% amino acid identity with toluene dioxygenase, suggests that probe 3 is located near the center of the β-sheet that makes up a large portion of the hydrophobic core of the α subunit. Also, probe 3 is close to the coding regions for the active site ligands (Figure 3). Thus the probe hybridization experiment performed on a shuffled library clearly identified a functionally important region. Random chimeragenesis experiments have in fact been used for the purpose of identifying functionally important sections of primary sequence in a number of enzymes [14,15,28]. In these studies, chimeric sequences are evaluated by restriction digestion, immunoblotting or sequencing. Because the probe hybridization method is a high throughput technique that can determine parental identity at several sites simultaneously, it may find application in locating functionally important sites and in identifying important interacting regions.

We chose to include parent *bphA1A2* in the shuffling experiment because we wished to determine to what extent a distantly related parent that is inactive on a

particular substrate would be incorporated into active, chimeric constructs. Such constructs make up only 2.6% of the active fraction of the three-parent dioxygenase library. For the unselected library, the *bphA1A2* parent was found in 28.4% of the positions probed. But at least one of the six probe positions was identified as *bphA1A2* for 47.1% of the clones, almost all of which were inactive. Because two crossovers involving *bphA1A2* on the same gene are highly improbable, only 11.8% of these *bphA1A2*-containing clones (17/306 total) did not have *bphA1A2* sequence at a terminus. Thus it appears that incorporation of sequence information from *bphA1A2,* at least in this way, is detrimental to forming enzymes that are active on toluene. Further study of the functional properties of the chimeric proteins will be necessary, however, to determine what role segments (especially small segments) from *bphA1A2* play in creating folded, chimeric proteins and in the acquisition of other properties, for example novel substrate specificities.

Relevance to laboratory evolution

The goal of laboratory protein evolution is usually to alter function towards some specific performance goal, such as increasing thermostability, binding affinity or enzyme activity on nonnatural substrates [29]. At this point, we have not yet assessed the evolutionary potential of the shuffled libraries. The laboratory evolution of enzymes with new substrate specificities, for example, may be accompanied by loss of activity towards substrates accepted by the parent

enzyme(s). Thus retention of activity on toluene may not be a good measure of whether the shuffled library contains dioxygenases that insert oxygen into new substrates not accepted by the parent enzymes. Catalytic function requires proper folding and activity on toluene or indigo does, however, indicate a lower limit on the fraction of chimeric sequences that can fold, and therefore on the fraction of the library that potentially contains new enzymes. A future goal of our work is to make an explicit connection between the library characteristics we measure here (crossover numbers, choice of parent sequences, activity) and the potential for acquisition of new properties.

Conclusions

Probe hybridization analysis allowed us to examine libraries made by DNA shuffling of dioxygenase genes. We found significant biases in where crossovers occur and in which parents are involved. These biases reduce the diversity of a library. In the context of a library of, say, 5,000 clones, this manifests itself as a small percentage of duplicate chimeras. This percentage should scale inversely (and the diversity should scale) with the number of combinations of good recombination sites (regions of sequence identity roughly **>** 7bp) taken n_c (the average number of crossovers) at a time.

If the parent pool contains parents with low sequence identity to others, few recombination sites will be available among the low-identity parents. Thus, clones containing sequence information from the low-identity parent are relatively

less diverse than the library as a whole. Fragments from a low-identity parent tend strongly to reassemble into full-length wildtype genes, which further reduces diversity. One useful strategy for avoiding reassembly of wildtype genes of a low-identity parent is to use only parts of this parent rather than a complete gene in the shuffling reaction.

Sequencing is expensive, and usually only a few clones from a library are completely sequenced. A limited probe hybridization analysis can determine the frequency of rarer events, such as crossovers between less similar parents and can accurately compare relative crossover frequencies in different regions. These data allowed us to draw conclusions that would not have been statistically significant or even evident from the complete sequences of a small sample of clones. From the probe hybridization data, we estimated the average number of crossovers in five regions of the dioxygenase genes with relatively high precision. The same data provided the basis for validating and tuning a model of the reassembly reaction. When functional information was coupled with the probe hybridization data, we were able to identify a critical region for enzyme activity and show that a low identity parent (*bphA1A2*) was incorporated into only 2.6% of active constructs. More extensive analysis, using larger numbers of probes to span the entire gene, will eventually provide data equivalent to complete DNA sequencing, at a fraction of the cost.

Materials and Methods

Construction of parent plasmids

Two libraries were created by recombining toluene dioxygenase (TDO), tetrachlorobenzene dioxygenase (TCDO) and biphenyl dioxygenase (BPO). Plasmid pJMJ4 was constructed by inserting the todBAD gene fragment from pDTG602 [30] between the BamHI and XbaI sites of ptrc99A (Amersham Pharmacia Biotech, Piscataway, NJ). Plasmids pJMJ2, pJMJ6, and pJMJ7 were constructed by cloning *todC1C2* from pDTG602 [30], *tecA1A2* from pSTE7 [31], and *bphA1A2* from LB400 [32], respectively, into the KpnI/BamHI sites of pJMJ4. Taq polymerase was used to amplify these genes prior to restriction digestion. In each case, several clones containing the target plasmid were tested for dioxygenase activity, and the most active clone was selected as the parent for DNA shuffling. Despite this effort to eliminate mutations introduced by cloning, several mutations were found two or more times in a pool of sequenced chimeras, and therefore probably were present on the parent plasmids. On *todC1C2*, the mutations G841A(Val to Ile), G1105A(Gly to Ser), and T1540C(Val toAla) occurred; on *tecA1A2*, the mutation G249A(Arg to Arg) occurred; and on *bphA1A2*, the mutations A1599G(Glu to Glu) and A1781G(Lys to Arg) were noted.

Creation of chimeric libraries using DNA shuffling

A hybrid of the DNA shuffling methods of Stemmer, *et al.* [8] and Abècassis, *et al.* [16] was used to create chimeric libraries. A forward primer (5' – GCATAATTCGTGTCGCTCAAGGC – 3') and a reverse primer (5' – GCCGAAATGCAACGTGCATTCG –3') were used to amplify a fragment (2.4-2.5kb) containing *todC1C2*, *tecA1A2*, and *bphA1A2* from pJMJ2, pJMJ6, and pJMJ7, respectively, using Pfu polymerase (Stratagene). A 100μl reaction mixture contained: 10μl 10xPfu buffer, 2μl of PCR nucleotide mix (10mM each), 40 pmol of each primer, 5U of Pfu polymerase, 3μl DMSO and 0.08 pmol of template plasmid. PCR was carried out on a MJ Research PTC-200 thermal cycler (Watertown, MA) under the following conditions: 94°C for 3 min, followed by 20 cycles of (94°C for 30 sec; 52°C for 30 sec; 72°C for 5 min), 72°C for 10 min, 4°C thereafter.

After purification and quantitation, equal amounts of parent DNA as determined by UV absorption at 260nm were mixed and subjected to DNAseI (Type II, from bovine pancreas, Sigma, St. Louis, MO) digestion. A 100μl digestion contained 70μl parent DNA mix, 10μl of 0.5M Tris-HCl (pH 7.4), 5μl of 0.2M manganese chloride, and 0.167U of DNAseI. After a 3-minute digestion at 15°C, the reaction was removed to 5μl of 1M EDTA, pH 8.0, on ice. Using the QIAquick gel-extraction kit (QIAGEN, Valencia, CA), fragments from 0.4-1.0kb were purified.

Fragments were reassembled in a 50μl reaction containing 42μl of fragment DNA, 5μl of 10xPfu buffer (Stratagene), 2μl of dNTP mix (10mM each, Promega, Madison, WI) and 1μl (2.5U) of Pfu polymerase (Stratagene). Cycling was according to the following protocol [16]: 96°C, 90 sec.; 35 cycles of (94°C, 30 sec.; 65°C, 90 sec.; 62°C, 90 sec.; 59°C, 90 sec.; 56°C, 90 sec.; 53°C, 90 sec.; 50°C, 90 sec.; 47°C, 90 sec.; 44°C, 90 sec.; 41°C, 90 sec.; 72°C, 4 min.); 72°C, 7min.; 4°C thereafter.

To amplify full-length (2.1kb) genes, this reassembly reaction was diluted 500x in the same PCR mixture used to acquire DNA for fragmentation. Forward and reverse primers internal to the first set of primers (5' – GGAATTCGAGCTCGGTACCAGGA – 3' and 5' – GTCATGACATCACCTAGGGATCC – 3') were used. Cycling was done as with the first reaction.

Library characterization

Unselected libraries: 374 wells of a 384-well plate were filled with 70 μl of M9-minimal media [33] containing 100 mg/L ampicillin and 0.4% glucose. Independent colonies were picked randomly using a QpixII colony picker (Genetix, New Milton, UK) and inoculated into the filled wells. The remaining 10 wells were then filled with 70 μl of M9-minimal media. 4 wells were left

uninoculated and 6 wells were inoculated with *E. coli* BL21(DE3) previously transformed with pJMJ2, pJMJ6 or pJMJ7.

Selected libraries: Following transformation and overnight incubation at 30°C on Luria-Bertani (LB) agar [33], indole crystals (Sigma, St. Louis, MO) were spread out onto the lid of the plate. The plate was incubated at 30°C for 3 hours and then stored overnight at 4°C. Oxidation of indole by the dioxygenase leads to the spontaneous formation of indigo, which is visible as a blue color. Blue colonies were gridded by hand into 374 wells of a 384-well plate which was filled in the same way and with the same controls as described for the unselected libraries.

Following overnight incubation at 275 rpm / 37°C in a New Brunswick Scientific Innova® incubator shaker (Edison, NJ) the plate was replicated onto Hybond-N+ 7.5 x 11.5 cm membranes (Amersham, Piscataway, NJ) placed on M9-minimal media [33] plates containing 1.5% bacto-agar using a 384-pin replicator (V&P Scientific, San Diego, CA). A separate membrane was used for each probe. After 17 hours of growth, cells were lysed and DNA was denatured and bound to the membrane by UV crosslinking according to the manufacturer's protocol (Amersham, Piscataway, NJ).

An oligonucleotide probe of about 22 nt was designed to specifically bind to each of the three parents in the initial pool at six gene positions at approximately the same temperature. The 18 probes (3 parents x 6 positions) for the dioxygenase

libraries were obtained from Gibco (Rockville, MD). They were labeled with fluorescein-11-dUTP using the terminal transferase reaction according to the Gene Images 3'-oligolabeling module protocol (Amersham, Piscataway, NJ).

Labelled probes were hybridized to chimeric clones according to the Gene Images protocol. Approximately 90 ng (11 μl of labeling reaction mixture) of labeled probe was added to prehybridized membranes in 18 ml of hybridization buffer and incubated for 2 to 3 hours at 61°C in a Model 400 Hybridization oven (Robbins Scientific, Sunnyvale, CA). Stringency washes were carried out twice in 1x SSC (15 mM Na_3citrate, 150 mM NaCl, pH7) for 15 min at 53°C. The Gene Images CDP-Star detection module (Amersham, Piscataway, NJ) was used according to manufacturer's instructions to obtain a chemiluminescent signal.

Data analysis

A digital image of the chemiluminescent signal was acquired using a Fluor-S MultiImager (Biorad, Hercules, CA) with a Nikkor 50mm f/1.4D AF lens (Nikon, Denver, CO). Peak signal intensity of each spot in the 24 by 16 array was quantified with the image analysis software Quantity One (Biorad, Hercules, CA) and exported to a Microsoft Excel spreadsheet. A signal intensity threshold was defined for each of the 18 blots. Intensities above this value were considered positive (true) while intensities below this value were considered negative (false). These data were used to determine the parent sequence present at each probed position for each clone in the array.

Solid-phase screening for activity toward toluene

Clones analyzed by probe hybridization were screened for activity toward toluene using the method described in Chapter 5 modified for 384-well use as described below. A 384-pin replicator (V&P Scientific, San Diego, CA) was used to transfer cells from a 384-well plate to Luria-Bertani (LB) agar [33] plates containing 100mg/L ampicillin. Colonies grew in this gridded format for 14 hours at 30°C and were transferred to M9 media [33] containing 4% bacto-agar, 0.5 mM IPTG, 100mg/L ampicillin, 1.6% D-glucose, and 80 mg/L $FeSO_4$-$7H_2O$ on a 132 mm diameter nitrocellulose membrane (Protran, 0.45μm, Schleicher & Schuell). The colonies were incubated for 12 min in an airtight container at 30°C containing an open dish of toluene. The membrane was transferred to a 3% agarose plate also containing 0.025% Gibbs reagent (added as a 2% solution in ethanol). After 8-9 min, a purple color developed under the active colonies and a digital image of the bottom of the plate was acquired using a Fluor-S MultiImager (Biorad, Hercules, CA) equipped with a Tamron SP AF20-40mm lens (Tamron Co., Ltd., Tokyo, Japan) and a 590±20nm bandpass filter (Omega Optical, Brattleboro, VT).

The image analysis tool Quantity One (Biorad, Hercules, CA) was used to quantitate the relative activities of the clones. A 24 x 16 array with a 2.5 x 2.5mm cell size was framed to the dimensions of the 384-well plate. The peak intensity for each cell was exported to an Excel spreadsheet. For inactive colonies, the peak intensity is equivalent to that recorded for areas with no colony present. The activity of each clone relative to toluene dioxygenase was determined by

dividing the difference of its peak intensity and the baseline peak intensity by the difference of the peak intensity of wild-type toluene dioxygenase and the baseline peak intensity for wild-type toluene dioxygenase. The baseline peak intensity varied slightly across the image, and was estimated for each clone by using the minimum peak intensity of the 8 nearest-neighbor cells. When the peak intensity of none of the nearest-neighbor cells was below a threshold value, the threshold value was used as the baseline intensity. The screening was done in duplicate to reduce uncertainty in the measurement.

References

1. Gilfillan, S. (1935). Inventing the Ship. Follett Publishing Co., Chicago, IL.

2. Nelson, R.S. & S. Winter. (1982). An Evolutionary Theory of Economic Change. Belknap Press, Cambridge, MA.

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

4. Christians,F.C., Scapozza,L., Crameri,A., Folkers,G. & Stemmer,W.P.C. (1999). Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nat. Biotechnol*. **17**, 259-264.

5. Schmidt-Dannert, C., Umeno, D., & Arnold, F.H. (2000). Molecular breeding of carotenoid biosynthetic pathways. *Nat. Biotechnol.* **18**,750-753.

6. Ness,J.E., Welch,M., Giver,L., Bueno,M., Cherry,J.R., Borchert,T.V., Stemmer,W.P.C. & Minshull,J. (1999). DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* **17**, 893-896.

7. Stemmer,W.P.C. (1994). Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389-391.

8. Stemmer, W.P.C. (1994). DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA* **91**, 10747-10751.

9. Zhao,H., Giver,L., Shao,Z., Affholter,A. & Arnold, F.H. (1998). Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol.* **16**, 258-261.

10. Volkov, A.A., Shao, Z. & Arnold F.H. (2000). Random chimeragenesis by heteroduplex recombination. *Methods in Enzymology* **328**,456-463.

11. Shao,Z.X., Zhao,H.M., Giver,L. & Arnold,F.H. (1998). Random priming *in vitro* recombination: an effective tool for directed evolution. *Nucleic Acids Res*. **26**, 681-683.

12. Coco, W.M., Levinson W.E., Crist M.J., Hektor H.J., Darzins A., Peinkos P.T., Squires C.H. & Monticello D.J. (2001). DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat. Biotechnol.* **19**,354-359.

13. Pompon, D. & Nicolas A. (1989). Protein engineering by cDNA recombination in yeasts:shuffling of mammalian cytochrome P-450 functions. *Gene* **83**,15-24.

14. Kim, J.-Y. & Devreotes P.N.. (1994). Random chimeragenesis of G-protein-coupled receptors. *J. Biol. Chem.* **269**,28724-28731.

15. Levin, L.R. & Reed R.R. (1995). Identification of functional domains of adenylyl-cyclase using *in vivo* chimeras. *J. Biol. Chem.* **270**,7573-7579.

16. Abècassis, V., Pompon, D. & Truan, G. (2000). High efficiency family shuffling based on multi-step PCR and *in vivo* DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. *Nucleic Acids Res*. **28**, e88.

17. Zhao, H. & Arnold, F.H. (1997). Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Res.* **25**,1307-1308.

18. Polz, M.F. & Cavanaugh, C.M. (1998). Bias in template-to-product ratios in

multitemplate PCR. *Appl. and Environ.Microb.* **64**, 3724-3730.

19. Forns,X., Bukh,J., Purcell,R.H. & Emerson,S.U. (1997). How *Escherichia coli* can bias the results of molecular cloning: Preferential selection of defective genomes of hepatitis C virus during the cloning procedure. *Proc. Natl. Acad. Sci. USA* **94**, 13909-13914.

20. Sugimoto,N., Nakano,S., Yoneyama, M. & Honda,K. (1996). Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res*. **24,** 4501-4505.

21. Moore, G.L., Maranas C.D., Lutz S. & Benkovic S.J. (2001). Predicting crossover generation in DNA shuffling. *PNAS* **98**,3226-3231.

22. Moore, G.L. & Maranas C.D. (2000). Modeling DNA Mutation and Recombination for Directed Evolution Experiments. *J. Theor. Biol.* **205**,483-503.

23. Sun, F. (1999). Modeling DNA Shuffling. *J. Comput. Biol.* **6**,77-90.

24. Suzuki, M., Christians F.C., Kim B., Skandalis A., Black M.E. & Loeb L.A. (1996). Tolerance of different proteins for amino acid diversity. *Molecular Diversity* **2**,111-118.

25. Voigt, C.A., Kauffman, S. & Wang, Z.G. (2001). Rational evolutionary design: The theory of *in vitro* protein evolution. *Advances in Protein Chemistry* **55**,79-160.

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

27. Kauppi,B., Lee,K., Carredano, E., Parales,R.E., Gibson,D.T., Eklund,H. & Ramaswamy, S. (1998). Structure of an aromatic-ring-hydroxylating dioxygenase-naphthalene 1,2-dioxygenase. *Structure* **6**,571-586.

28. Hansson, L.O. & Mannervik, B. (2000). Use of chimeras generated by dna shuffling: probing structure-function relationships among glutathione transferases. *Methods in Enzymology* **328**,463-477.

29. Petrounia, I.P. & Arnold F.H. (2000). Designed evolution of enzymatic properties. *Curr. Opin. Biotech.* **11**,325-330.

30. Zylstra,G.J. & Gibson,D.T. (1989). Toluene degradation by *Pseudomonas putida* F1 – Nucleotide sequence of the todC1C2BADE genes and their expression in *Escherichia-coli*. *J. Biol. Chem*. **264**, 14940-14946.

31. Beil, S., Happe,B., Timmis,K.N.& Pieper,D.H. (1997). Genetic and biochemical characterization of the broad spectrum chlorobenzene dioxygenase from *Burkholderia sp*. strain PS12-Dechlorination of 1,2,4,5-tetrachlorobenzene. *Eur. J. Biochem.* **247**,190-199.

32. Mondello,F.J. (1989). Cloning and expression in *Escherichia coli* of *Pseudomonas* strain LB400 genes encoding polychlorinated biphenyl degradation. *J. Bacteriol*. **171,** 1725-1732.

33. Sambrook,J., Fritsch,E.F. & Maniatis,T. (1989). *Molecular Cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Chapter 8

Functional genomics of a library of chimeric enzymes

Abstract

Recombination of homologous genes ("family shuffling") has been applied repeatedly to the laboratory evolution of enzyme function. In this study, we demonstrate the use of family shuffling to investigate sequence-function relationships through characterization of substrate specificity and sequence for libraries of chimeric dioxygenases. We screened hundreds of chimeras for activity toward ten substrates accepted by at least one of the parent enzymes and identified clones with altered substrate specificities. From a library of 7,900 chimeric dioxygenases, we isolated 13 that were active toward *n*-hexylbenzene, a substrate not metabolized to a measurable extent by the parent enzymes. Chimeric enzymes with altered or novel function generally had more crossovers than the library as a whole and were primarily derived from the most identical parent genes. A probe hybridization assay was used to determine the sequence of the screened chimeras at specified positions. By coupling the functional data with the sequence information, we assessed sequence-function relationships and identified a single region of sequence that to a large extent determined the substrate specificity. This approach should be generally useful for identifying sequence elements that affect protein characteristics.

Introduction

We can learn a great deal about sequence-function and structure-function relationships by applying functional genomics to our study of genes encoding evolutionarily related enzymes. Within a given enzyme family, distinct functionalities can arise from a small number of amino acid changes. As in functional genomics where genes are inserted or deleted to determine their function, the positions of function-altering polymorphisms within a single gene can be determined by swapping sequence elements between functionally distinct homologs. In fact, in several studies, small numbers of designed chimeras have provided useful insight into sequence-function relationships [1-4]. These chimeras typically contain limited numbers of crossovers (one or two), and thus noncontiguous sequence elements that contribute synergistically to a property are generally not encountered. Typically the number of sequences analyzed is small due to the difficulty of constructing particular chimeras.

In some well-characterized enzyme families, these same sequence-function relationships can be predicted using a bioinformatic approach that combines phylogenetic and structural information. With this approach, the task is to separate functionally important mutations that have occurred over millennia of evolution from those that are functionally neutral. A structure generally suggests positions that contribute to a particular enzyme characteristic; for example, residues surrounding the active site often contribute to substrate specificity.

Positions implicated by the structure that *are not* conserved in functionally distinct homologs (and *are* conserved in functionally similar ones) are predicted by this approach to play a role in determining the property of interest. Unfortunately, the sequences of functionally distinct homologous enzymes are often too nonidentical to allow this type of evaluation. In other cases, the analysis is complicated because the functional change under investigation either results from a combination of mutations or can arise from a multitude of different mutations.

In this study, we use family shuffling of functionally distinct homologs to create libraries of chimeric proteins that are analyzed to determine sequence-function relationships. Unlike earlier studies where a small number of chimeras with few crossovers were characterized, our chimeras have multiple crossovers, and we employ high-throughput screening and sequencing tools to allow consideration of hundreds of chimeras. This coupling of family shuffling with high-throughput analysis should allow us to extract more complex sequence-function relationships such as those involving a collection of noncontiguous sequence elements. We can use this family shuffling approach to investigate the genetic basis of *specific* characteristics (e.g., substrate specificity or stability) by choosing parents that differ in that property.

The dioxygenase enzymes have been the subject of laboratory evolution efforts targeted to prospective applications in bioremediation and biocatalysis [5,6].

Dioxygenases with altered substrate specificity, extended substrate range, and enhanced activity have been generated by family shuffling [7-10] as well as random mutagenesis [11-13]. Family shuffling experiments have targeted PCB bioremediation, and chimeras of various biphenyl dioxygenases have been created with broadened congener specificity [8-10] or enhanced activity toward alkylbenzenes [7].

In Chapter 7, we demonstrated sequence characterization of several hundred dioxygenase chimeras using a probe hybridization screen to determine the parent sequence incorporated at six specific gene positions. These dioxygenase libraries have a moderate number of crossovers (≤ five) and an extremely low (0.01%) point mutation rate. Using these libraries, we demonstrate here that recombination of these dioxygenase parents gives rise to a diverse array of functionalities, including altered substrate specificities and activity toward a substrate not accepted by the parents. Through analysis of substrate specificity data in concert with probe hybridization data, we identify a key sequence element important to substrate specificity and determine sets of sequence elements that correlate to certain functionalities. Finally we discuss the implications of our study for the appropriate choice of parents and the frequency of crossovers for laboratory evolution.

Results and Discussion

Chimeric library construction and characterization

Two libraries (C1and C2) were constructed by DNA shuffling the genes encoding the α and β subunits of three dioxygenase parents, toluene dioxygenase from *Pseudomonas putida* (TDO) [14], tetrachlorobenzene dioxygenase from *Burkholderia sp*. strain PS12 (TCDO) [15] and biphenyl dioxygenase from *Pseudomonas* strain LB400 (BPDO) [16]. For the α and β subunit genes and the ~100 bp region between them, TDO and TCDO are 85% identical at the nucleotide level while BPDO shares only 63-64% identity to the other two parents. To construct the parent plasmids and libraries we used the genes encoding the electron transfer proteins from TDO, thus it is possible that the TCDO and BPDO constructs have diminished activity compared to their native gene arrangement. This gene arrangement is not expected to influence substrate specificity.

We applied probe hybridization analysis to six gene positions to determine which parent contributed sequence at each position for hundreds of clones from the chimeric libraries. Probe sites were distributed uniformly across the ~2000 bp genes encoding the α and β dioxygenase subunits. Probe hybridization analysis of library C1 was described in Chapter 7; we are presenting those data here with more extensive functional evaluation. From the probe hybridization data and

using a statistical model that corrects for pairs of crossovers that might occur between probes, we calculated that library C1 has 3.7±0.3 crossovers per gene, while C2 has 2.0±0.2 (see Chapter 7). 83% of the chimeras from library C1 and 70% from C2 had unambiguous results for all six probe positions; only these clones were included in further functional analysis. ~35% of the constructs from the two libraries retained at least 20% of the activity of the most active parent enzyme toward toluene, bromobenzene, biphenyl or indole. The two libraries have a similar positional bias in the percent incorporation of each parent: TCDO is preferentially incorporated at probe position 2, and TDO sequence predominates at position 3 (see Chapter 7). Crossovers between TDO and TCDO occurred approximately 7 times more frequently than between other parent combinations, reflecting the relatively high (85% vs. 63-64%) sequence identity of these two parents. Sequencing of 8 active and 10 inactive clones from library C1 revealed only four nucleotide substitutions (see Chapter 7). Library C2 was constructed with the same proof-reading polymerase, so we expect a similarly low error rate.

Activities of chimeric dioxygenases toward ten different substrates

In order to assess the substrate specificities of our chimeras, we chose to evaluate the relative activity of each clone toward ten substrates. We chose the substrate set to represent a range of sizes and chemical functionalities while still containing some pairs that are quite similar in both respects (see Figure 1). A

consequence of this choice is that the parent dioxygenases are shown to have different substrate specificities (see Figure 1). TDO is highly active toward the monocyclic substrates tested, while TCDO and BPDO prefer larger substrates such as biphenyl and naphthalene. TCDO has a preference for benzenes with larger halogen substituents over fluorobenzene and toluene, which were not metabolized to a measurable extent.

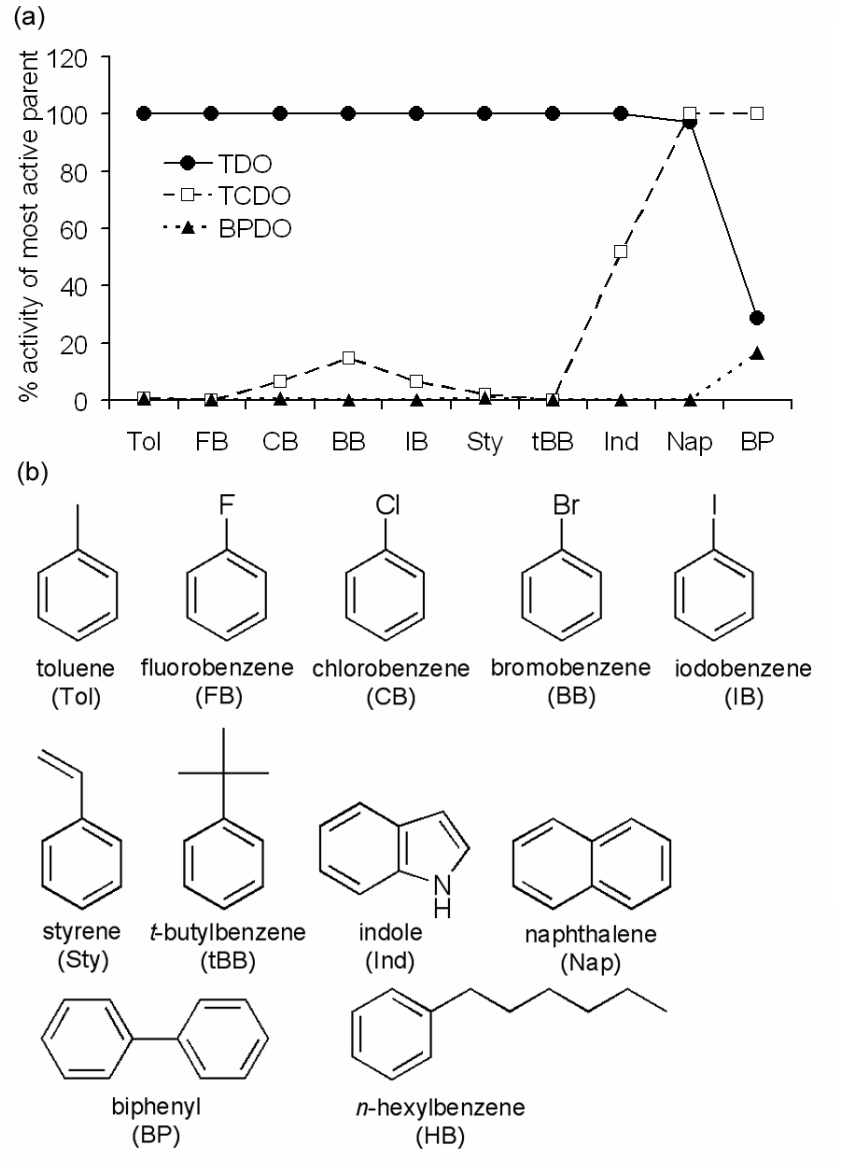


Figure 1. (a) Substrate specificity profiles for three wildtype dioxygenases. The relative activity shown is defined as the percentage of the activity of the most active wildtype strain (which was TDO for most of the tested substrates). Substrate abbreviations are shown in Figure 1(b). (b) Substrates used to assess dioxygenase specificities. *n*-Hexylbenzene is not a substrate of any of the parent dioxygenases.

Hundreds of clones from each library were first screened for activity toward toluene, bromobenzene, biphenyl, indole, naphthalene, *t*-butylbenzenene and styrene using a solid-phase method (see Materials and Methods). Clones with > 25% activity relative to the most active parent on at least one substrate were screened in the liquid phase for activity toward toluene, fluorobenzene, chlorobenzene, bromobenzene, iodobenzene and biphenyl (see Table 1). For substrates where activities were measured in both liquid and solid phase, we used the liquid phase data to compile activity profiles for each clone.

Library	# screened	# active*	# screened in liquid media
chimeric C1	364	124	97
chimeric C2	363	123	109
total →	727	247	206

* Clones were considered active if they retained at least 25% of the activity of the most active parent toward at least one substrate

Table 1. Summary of screening for activity toward toluene, fluorobenzene, chlorobenzene, bromobenzene, iodobenzene, styrene, *t*-butylbenzene, indole, naphthalene and biphenyl. Variants were screened to determine retention of activity. Activities of the active mutants and the subset of active chimeras with complete probe hybidization data were determined in liquid media, as described in the text.

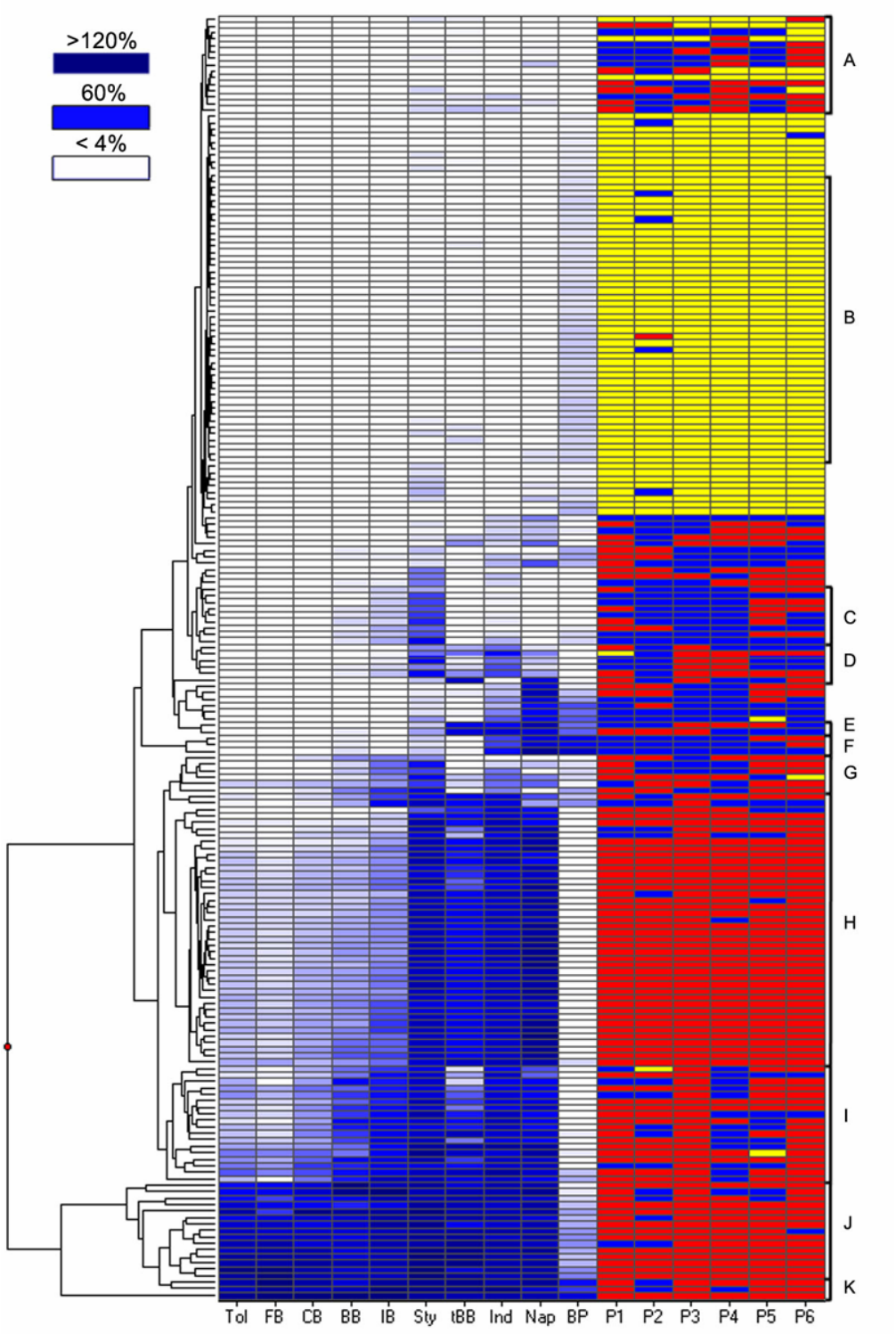
Based on these primary screening data, we selected 28 clones with altered substrate specificities for rescreening to check the reproducibility of the measurements. This time the *t*-butylbenzene and styrene measurements were made in liquid phase, while the indole and naphthalene measurements were repeated on the solid phase. There is a nonlinear correspondence between the solid and liquid methods, such that clones with low relative activity (10-30%) in

the liquid assay generally have higher relative activity (40-60%) in the solid-phase screen. For clones active toward a particular substrate (>25% of wildtype activity), the duplicate liquid-phase measurements had an average relative deviation of 12-24%, depending on the substrate. Despite the significant uncertainty observed for some substrates, 24/28 of the chimeras showed the expected altered specificity after rescreening.

Analysis of sequence-function relationships

Data from functional screening and probe hybridization analysis were combined into a single dataset comprising 206 active chimeras (see Table 1). We found it useful to visualize this 16-dimensional dataset using hierarchical clustering as performed by the Spotfire® software package (Spotfire, Somerville, MA). Clustering of the chimeras was based only on their substrate specificities. Hierarchical clustering was conducted using Euclidean distance as a similarity metric and the sum of the scaled relative activities for each clone as the ordering function. To reduce the effect of the high uncertainty associated with measurements of activity toward naphthalene, indole, styrene and *t*-butylbenzene, we scaled the relative activity data for these substrates by a factor of one-third. The activity and genotype data were then projected graphically using a heatmap representation, as shown in Figure 2. To present the result as a heatmap, the scaled relative activity values were replaced with the true values and then represented on a blue scale ranging from white for no activity to dark

blue for high activity. The sequence data from the probe hybridization experiments are shown in separate columns.



(legend on following page)

Figure 2. Heatmap representation of hierarchical clustering results for the two chimeric libraries. Substrate specificity profiles for active clones were scaled and clustered as described in the text. For each substrate, the relative activity of each clone is represented using shades of blue (see below), where white is low activity and blue is high activity. Columns representing relative activity toward a particular substrate are abbreviated as shown in Figure 1(b). Columns P1 - P6 represent the sequence identified at each of the six probe positions. Sequence from TDO, TCDO and BPDO is represented by red, blue and yellow, respectively. The clustering was based only on the functional data. Distances calculated during clustering are represented in the tree to the left of the heatmap (see Figure 2) as a horizontal branch length. Functionally similar clones are associated with a low distance and thus are connected by short branches. From the tree representation, one can identify clusters of varying size by selecting either small branches that contain only a few clones or large branches that subsume a large fraction of the library.

From the clustering analysis, we can readily distinguish sequence-function relationships. Clones in regions J, F, and B of Figure 2 display the TDO, TCDO, and BPDO specificities, respectively, and the probe hybridization analysis shows a prevalence of sequence from the expected parent (e.g., TDO sequence for TDO specificity) in these regions. To a large extent, the sequence at probe 3 determines the substrate specificity: active clones with TDO sequence at this position have TDO-like specificity (regions D,E,H-K), while clones with TCDO sequence have varied functionalities (regions C, F and G), but in general have no activity toward toluene, *t*-butylbenzene or fluorobenzene. Only three clones (region F) from the chimeric libraries maintained the wildtype TCDO function, and these all had TCDO sequence at probe positions 1-4. Clones in region H have TDO specificity but decreased activity, yet many of these have TDO sequence at

every probe position. We believe that these clones are an artifact of the cloning process used to create the libraries[‡].

The clones with BPDO-like specificity shown in region B of Figure 2 generally conserve BPDO sequence at all six positions, although substitution at position 2 is allowed. The absence of examples of other gene configurations that maintain the BPDO specificity reflects the general incompatibility of BPDO sequence with the other parents (87.8% of constructs containing sequence from BPDO and at least one of the other parents were inactive) and from the lack of sequence diversity in the library. Due to the low sequence identity, there are relatively few sites for recombination between BPDO and the other parents (see Chapter 7).

Clones with similar changes in substrate specificity often have similar sequences, according to probe hybridization results. Clones in region C have little activity toward biphenyl and indole, no activity toward toluene, but slight activity toward benzenes with large halogen substituents and styrene; for these clones we observe a great deal of TCDO sequence, especially at probe positions 2, 3 and 4. Clones in regions D and E have high selectivity for *t*-butylbenzene over toluene, although both of these substrates are accepted by TDO and neither is accepted by TCDO or BPDO. In region D, TCDO sequence at position 2 is

[‡] The probe data and function observed for the wildtype-like clones in region H are consistent with that of the pJMJ2 plasmid used as a cloning vector for the libraries. We believe that this plasmid came through the cloning procedure without accepting a shuffled DNA insert. Due to a spurious 30bp insertion, clones carrying pJMJ2 have less total activity than with the wildtype-TDO plasmid pJMJ11 (See Materials and Methods). Sequencing of twelve plasmids from library C1 found none with the pJMJ2-like construction.

followed by TDO sequence at position 3. The two clones in region E also have TDO sequence at position 3, though position 2 is occupied by TCDO sequence in one case. Other altered specificities based on the common theme of high activity toward large halobenzenes and no activity toward toluene are lumped into region G. Region I contains clones with a preference for large halobenzenes over fluorobenzene and toluene; most of these clones have TCDO sequence at position 4. In region J we observe TDO-like clones with either improved selectivity toward biphenyl or improved overall activity.

In Figure 3 we show activity profiles for chimeras that exhibited altered specificity in the primary screen and after rescreening as described above (data shown are from the rescreening). K-means clustering with Spotfire® was used as a guide to group these clones into the five clusters shown in Figure 3. Clones shown in Figure 3(a) are not active toward toluene or fluorobenzene but have low activity toward other substrates accepted by TDO or TCDO. Clones shown in Figure 3(b) are similar to those in (a) except they have some activity toward toluene and in general are more active. In Figure 3(c) we show 3 TDO-like clones with improved activity and in at least two cases a marked specificity shift. Figure 3(d) shows three chimeras with a preference for *t*-butylbenzene over toluene and other TDO substrates, while still maintaining activity toward substrates of TCDO. Shown in Figure 3(e) is a variant with high activity toward indole and naphthalene but no measurable activity toward smaller substrates or biphenyl.

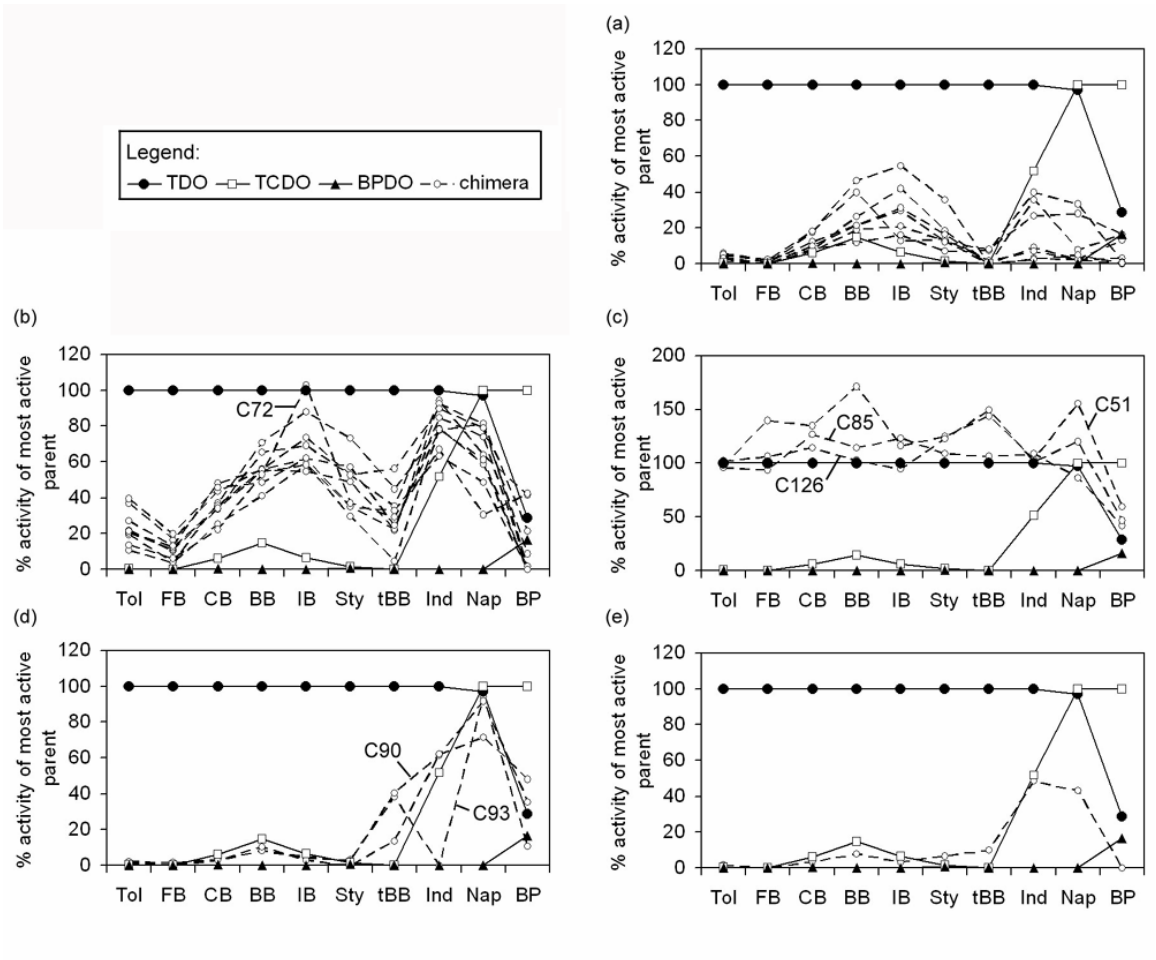


Figure 3. Substrate specificity profiles of selected chimeras with altered substrate specificity are compared to wildtype TDO, TCDO and BPDO. Activities are represented as a percentage of the activity of the most active parent. Grouping of similar clones into 5 clusters (a)-(e) was informed by K-means clustering. Sequenced clones C14 and C72 are in cluster (b); C51 and C126 are in (c); C90 and C93 are in (d). Profiles are labeled when space allows.

Sequences of seven of the variants having altered specificity are shown in Figure 4(b). For these seven clones, the probe hybridization data are consistent with the sequences. Four nucleotide-level mutations were observed (Figure 4(b)), and two of these gave rise to amino acid substitutions: Pro270Ser on the α subunit of C72, and Ala123Thr on the β subunit of C85. Thus for at least five variants, the

observed changes in substrate specificity can be attributed solely to recombination.

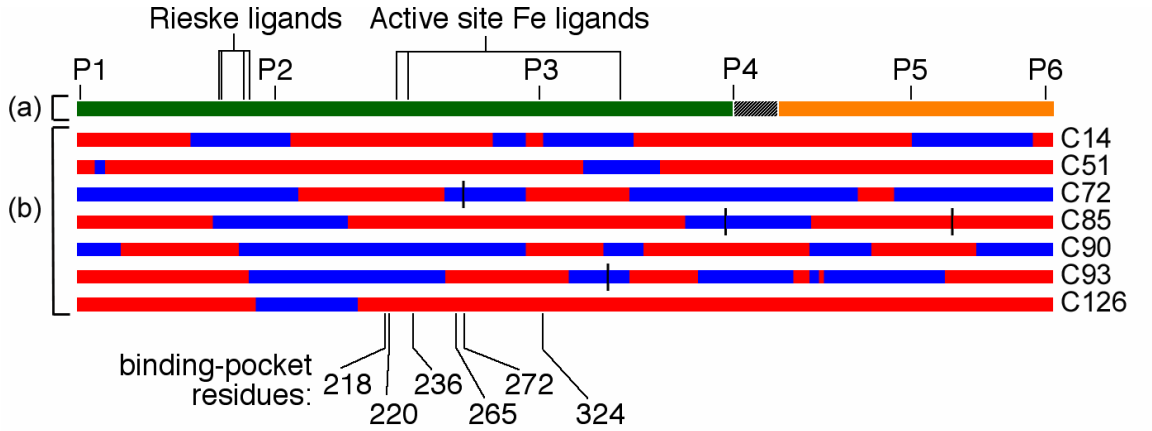


Figure 4. Sequence of selected chimeric dioxygenases. (a) Dioxygenase sequence landmarks and relative position of six probe sites (P1 through P6) used to characterize the chimeric libraries. Sequence encoding the α and β subunits are green and orange, respectively, and the noncoding region is shown in black. TDO residues that align to amino acids in the substrate binding pocket of naphthalene dioxygenase [18,28] are shown when not conserved among the three parents used in this study. Sequence encoding binding-pocket residue 324 lies within the binding region for probe 3. (b) Sequences of seven chimeras with altered substrate specificity. Sequences from TDO, TCDO and BPDO are represented by red, blue and yellow, respectively. Point mutations are shown as vertical black lines.

Structural interactions implied by activity conservation

Chimeric proteins which conserve important amino acid interactions are more likely to fold correctly and preserve function [17]. Thus we can infer that amino acids near two probe sites interact in the protein structure if those sites are occupied by the same parent among a high fraction (relative to other probe site pairs) of active chimeras. The *absence* of interaction can not be determined, since important interactions are often conserved among homologous genes. To

quantify the extent of enrichment in functional clones for library subsets defined by identifying the same parent at two probe sites, we define the Functional Enrichment Statistic (FES) as the number of *active* clones with the same parent at both positions divided by the *total* number of clones with the same parent at both positions, normalized by the fraction of clones that retain activity. Thus probe positions that are either noninteracting or where interactions are conserved among the parents should have an FES of approximately one, while probe positions connected in the structure by nonconserved, functionally important interactions are expected to have an FES > 1.

For the libraries investigated here, certain subpopulations defined by genetic conservation at two probe positions are enriched in active clones as determined by calculation of FES values[§] (see Table 2). For example, probe site 1, located at the N-terminus of the α subunit, is highly coupled to sites 3-6, as evidenced by high FES values (FES = 1.39 to 1.52). The first 80 amino acids of naphthalene dioxygenase span both the catalytic and Rieske domains and are in close proximity to β subunits [18]. Thus it is understandable that this region would interact with most other probe positions. The highest FES values are seen for noncontiguous probe pairs, while incorporation of the same parent at neighboring

[§] Clones were considered active if they they retained at least 20% of the activity of the most active wildtype strain toward toluene, bromobenzene, t-butylbenzene, styrene, indole, naphthalene, or biphenyl, as determined by solid-phase screening as described. Only the 556 chimeras with complete probe hybridization results were included in the calculation; of these, 40.2% were active. Depending on the probe pair, 51-78% of chimeras had the same parent at both positions.

probe pairs (see diagonal of Table 2) does not strongly correlate to retention of function, with the exception of probe pair 2-3.

probe position x	probe position y				
	2	3	4	5	6
1	1.16	1.52	1.45	1.39	1.46
2		1.37	1.28	1.25	1.29
3			1.16	1.18	1.18
4				1.04	1.09
5					1.09

Table 2. Functional Enrichment Statistic (FES) calculated for libraries C1 and C2, as described in the text. A high FES value implies a functionally important interaction between residues near the probe sites under consideration.

Acquisition of activity toward hexylbenzene

Seven thousand nine hundred previously unscreened variants from chimeric library C1 were screened for activity toward *n*-hexylbenzene using a solid-phase assay, as described in Materials and Methods. Briefly, colonies expressing enzyme variants are exposed to hexylbenzene and then contacted with Gibbs' reagent, which reacts with phenols to yield colored compounds. Using this assay, no color was observed for any of the parent enzymes, but 16 chimeric-dioxygenase-expressing colonies turned bright purple after exposure to Gibbs' reagent. These were rescreened in liquid media to ensure that the bright purple coloration from the solid-phase screen was dependent on the presence of hexylbenzene and could be reproduced; 13 of the 16 clones showed reproducible coloration.

To confirm that the oxidation product was indeed a dihydroxyhexylbenzene, supernatant from three chimeras and the three wildtype strains was extracted with ethyl acetate and analyzed by GC-MS. For the chimeras, the major peak in the chromatogram corresponded to a mass spectrum with a molecular ion at *m/z* = 194 (the expected molecular weight of dihydroxyhexylbenzene) and a base peak characteristic of alkyl-substituted catechols at *m/z* = 123. For TDO and TCDO, no quantifiable peak was observed in the chromatogram, but the mass spectrometer recorded some density at *m/z* = 194 and *m/z* = 123 at the elution time of the dihydroxyhexylbenzene species. No indication of hexylbenzene oxidation was recorded for the BPDO parent. Based on peak integration, we estimate that the best evolved chimera (6X7) is at least 100-fold more active toward hexylbenzene than either TDO or TCDO and produces dihydroxyhexylbenzene at a rate of approximately 3.5 μM/(hr•OD)[†] (see Materials and Methods for biotransformation conditions). Our wildtype TDO construct oxidizes toluene at a rate of 180 μM/(hr•OD); thus even the best gain-of-function variant metabolizes hexylbenzene to a relatively small extent.

We determined activity profiles for all 13 of the hexylbenzene-active clones using the ten test substrates. Hierarchical clustering was used to sort these clones according to this functional data. Four groups of functionally-similar clones were apparent from the clustering results (labeled I. through IV. in Figure 5). Activity

[†] Based on extraction of the cell pellet with ethanol, approximately 50% of the dihydroxyhexylbenzene product remains associated with the cells.

toward hexylbenzene was often acquired concomitantly with increased activity toward *t*-butylbenzene or biphenyl (see Figure 5(a), groups III. and IV.). The clone with the highest activity toward hexylbenzene, 6X7, is inactive toward nearly all of the other substrates (see Figure 5(a), group I.). The fact that we found several distinct substrate specificities demonstrates that recombination gave rise to diverse solutions to the problem of acquisition of activity toward hexylbenzene.

We have sequenced all thirteen of the hexylbenzene-active chimeras (Figure 5(b)). The chimera with the highest activity toward hexylbenzene, 6X7, is composed almost completely of BPDO sequence, with just a small section of TDO sequence around the probe 3 binding site. Clones from groups II., III. and IV. are composed solely of sequence from TDO and TCDO. For these clones, all residues thought to form the substrate binding pocket are identical to those from TDO (TDO and TCDO are identical at positions 236, 265 and 272). With the exception of 6X2, clones from groups II., III. and IV. have TCDO sequence at the C-terminus of the α subunit. Clone 6X14 contains sequence from all three parents and has low activity toward some of the tested substrates. Its substrate specificity is most similar to TCDO, perhaps reflecting incorporation of TCDO sequence at probe position 3. Of the hexylbenzene-active clones, only 6X14 has a spontaneous mutation in a coding region (I31T on the α subunit).

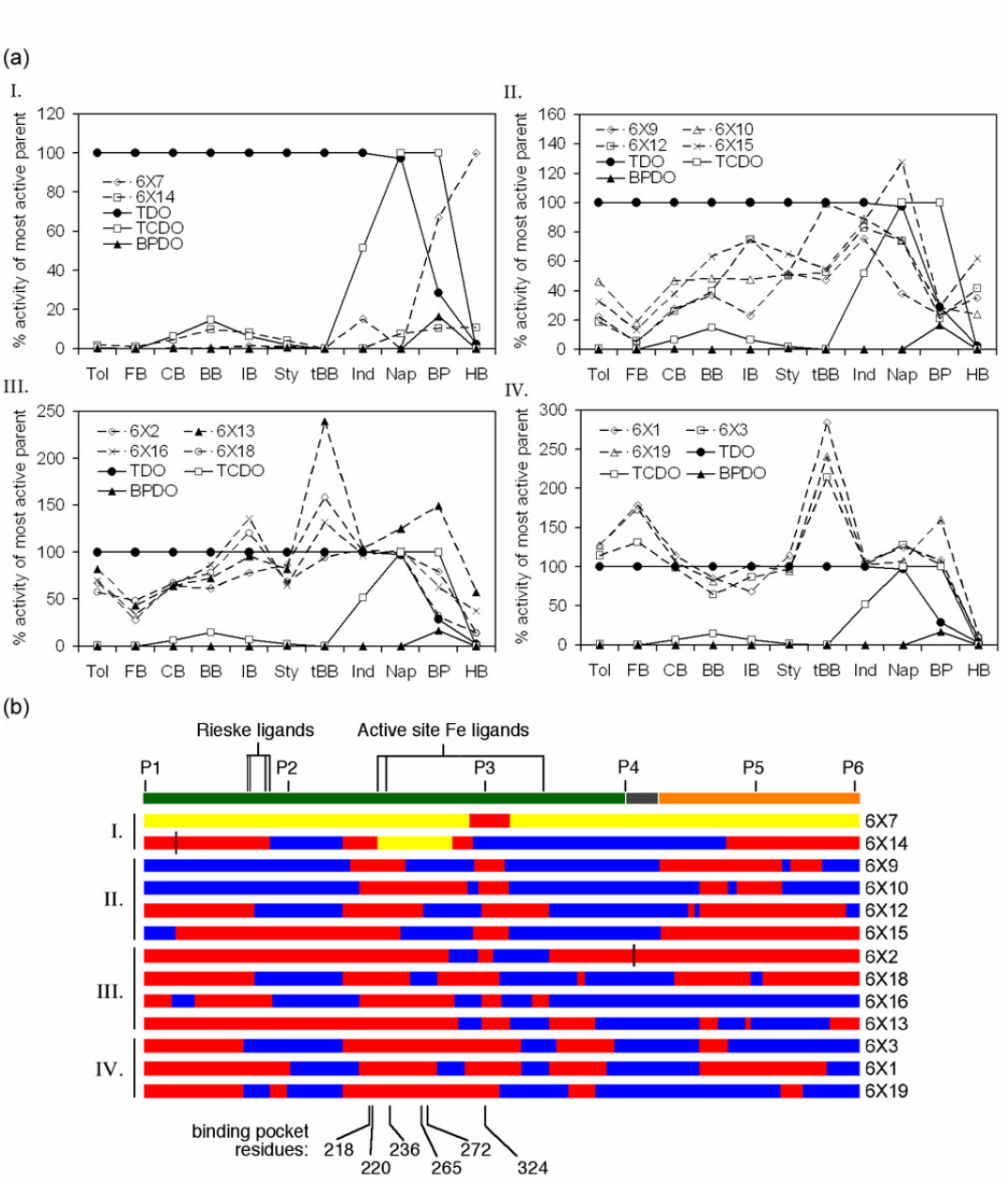


Figure 5. Substrate specificity profiles and sequences of chimeras that acquired activity toward hexylbenzene. (a) Hierarchical clustering was used to group chimeras with similar activity profiles into four groups labeled I. through IV. Activities are represented as a percentage of the activity of the most active parent. In the case of hexylbenzene (HB), the most active variant, 6X7, is used as a basis for comparison. All variants have at least 5% of the activity of 6X7 toward hexylbenzene. (b) Sequences of chimeras that acquired activity toward hexylbenzene are represented as described in Figure 4. Binding pocket residues (based on alignment to naphthalene dioxygenase [18,28]) are shown when not conserved among the three parents used in this study.

Statistical comparison of evolved clones to the naïve libraries

In the evolved clones, we see a relatively high number of crossovers. Nineteen out of the 24 chimeras with altered specificity were from the high-crossover-number library C1 (3.7±0.3 crossovers per gene), and only five were from library C2 (2.0±0.2 crossovers per gene). We screened an equal number of clones from the two libraries, and both libraries contained ~35% active variants. The seven sequenced clones with altered specificity had 6.0±1.0 crossovers per gene, while the 13 clones that were active toward hexylbenzene had 7.4±0.7 crossovers per gene, thus both groups had more crossovers than either library average. Several factors probably contribute to this effect: 1) the evolved clones can not be the wildtype with, of course, zero crossovers, 2) the prevalence in the evolved clones of sequence from TDO and TCDO, the most identical and therefore most crossover-prone parents (most chimeras containing BPDO sequence were inactive), and 3) a correlation between high numbers of crossovers and functional diversity. Correction of the average number of crossovers in library C1 for the first two considerations above results in a value of 5.0±0.7, which is still considerably lower than the average for the evolved clones.

No consensus on an optimal number of crossovers has emerged from family shuffling experiments. Examples of evolved variants with only one crossover are extremely rare [19], except in special cases [20], and were not found in this study, despite their prevalence in library C2. For the TDO and TCDO parents,

we have found that a library average of 4-6 crossovers per gene gives rise to variants with altered function; with a higher crossover rate (e.g. 15-30 per gene), we could have accessed more complex chimeric genes that may have been more enriched in functional diversity. Such high crossover rates are possible with "synthetic shuffling" methods which require synthesis and assembly of a set of degenerate oligonucleotides [21,22]. One such library has been shown to contain greater functional diversity than a DNA-shuffled library constructed from the same parents [22].

From sequencing of 18 clones from library C1, we found 16 chimeras (see Chapter 7), and these have an average Hamming distance of 38±7 (range 0-110) from the closest parent. In comparison, the seven clones with altered specificity and 13 hexylbenzene-active chimeras have an average Hamming distance of only 24±3 (range 3-49). Thus solutions found in a small library (~10^3 or 10^4) were relatively close to the starting parents in sequence space. As shown in Chapter 7 for probe-level sequence data, active chimeras tend to have sequence from a single parent incorporated at most of the probed positions; this implies that active chimeras on average have lower Hamming distance than inactive ones. As shown here, low-Hamming distance chimeras are more likely to exhibit altered or novel functions, perhaps simply because they are more likely to retain activity. In this case, it was unnecessary to access high-Hamming distance variants in order to generate functionally-diverse libraries.

Implications for selecting parents for DNA shuffling

As shown in Figure 1(a), the parents selected for this study have distinctly different substrate specificities. In addition, the parents have varied sequence identity: TDO and TCDO share 85% nucleotide identity, while BPDO is approximately 63% identical to the other parents. All but two of the sequenced, functionally novel clones contained sequence only from TDO and TCDO (see Figure 4(b) and Figure 5(b)), and, in fact, very few of the active chimeras contained sequence from both BPDO and either of the other parents (see Figure 2). This may result as much from the difficulty of obtaining diverse BPDO-containing chimeras by DNA shuffling as from incompatibility at the protein level. This prevalence of sequence from the most identical parents and the low Hamming distance of our evolved clones suggest that parents with high sequence identity (> 75%) should be shuffled when only relatively small libraries can be assayed and a homologous recombination method is to be used.

In only a few studies, parents with identity < 70% have been shuffled to yield improved variants. In one case success was enabled both by selection from a large library (~ 50,000) and perhaps also from an extremely high level of point mutagenesis [23]. In another study, a section of a protease gene was replaced with a shuffled library of natural isolates with identity between 56.4 and 99.5% [24]. After screening 10,000 chimeras, several variants with altered function were isolated and found to have at most 15 amino acid changes relative to the most

similar parent in the shuffled region. Our results are consistent with these studies in that evolved chimeras of low-identity parents are scarce and often have few substitutions relative to the most-similar parent.

Most reported studies used functionally distinct parents [7-10,25], and it is reasonable to hypothesize that such parents will beget more functional diversity than shuffling homologs that have simply diverged neutrally. However, since high sequence identity often comes at the cost of functional diversity, it can be difficult to identify parents that satisfy both criteria. Thus protein families with no characterized members that are both highly identical and functionally distinct may be difficult to evolve by DNA shuffling. Various homology-independent recombination techniques could be used to shuffle lower-identity sequences and circumvent this problem [20,26,27]. However, the currently available techniques either produce only one random crossover [20,27], require that each crossover in a multiple-crossover variant be nondisruptive [26], or require specified crossover points. Furthermore, random shuffling of low-identity sequences is likely to lead to a large fraction of inactive clones.

Functional role of the sequence surrounding probe 3

Figure 2 dramatically demonstrates that the sequence present at probe site 3 to a large extent determines the substrate specificity of the chimeric clones. Furthermore, all but one of the hexylbenzene-active clones have TDO sequence

at probe 3, and the most active of these toward hexylbenzene (6X7) consists solely of BPDO sequence except for a small segment of TDO sequence at probe 3 (see Figure 5(b)). These results strongly indicate a key functional determinant in this region.

Based on alignment to naphthalene dioxygenase (NDO), for which a crystal structure is available (PDB ID: 1NDO) [18,28], the sequence surrounding probe 3 folds into the core of the α subunit and forms part of the substrate binding pocket. Recent studies using biphenyl dioxygenases have confirmed an important functional role for this region [4,8,29]. One study employed family shuffling to create biphenyl dioxygenase chimeras with broadened specificity [8]. One of the chimeras found (BphA-II-9) is the BPDO strain used in this study with only 7 amino acids from another BPDO from *Comamonas testosteroni* B-356 substituted in the region around probe 3 (TDO residues 323-329). This clone is strikingly similar to our best hexylbenzene-active chimera, 6X7, which is BPDO except for TDO sequence from 314-345. Another study found that the Ile336Phe mutant of biphenyl dioxygenase from *Pseudomonas pseudoalcaligenes* KF707 was active toward 2,5,2',5'-tetrachlorobiphenyl, which is not accepted by the wildtype enzyme [29]. Position 336 in KF707 aligns both to Ile324 in TDO and the binding-pocket residue Ser310 in the NDO structure, yet it is not conserved among the three parents in this study (Ile324 occurs as Ala324 in TCDO and Phe335 in BPDO, see Figure 6). Sequence encoding this residue lies within

probe binding region 3 (see Figure 6), thus it may be that this position strongly contributes to the specificity of our enzymes.

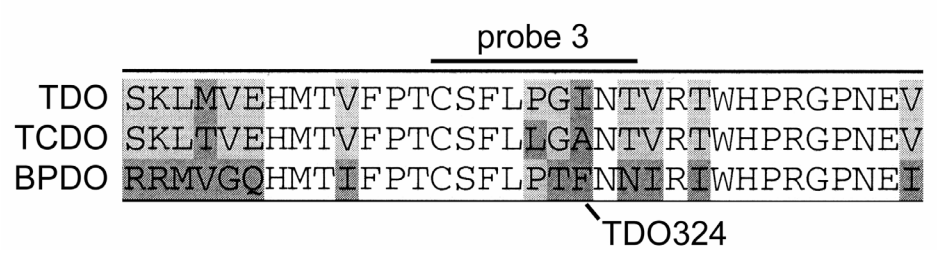


Figure 6. Alignment of the three parent dioxygenases centered around probe site 3. Nonidentical residues are in dark gray and the consensus residue at nonconserved positions is in light gray. Position 324 lies within probe binding region 3, and other nonconserved positions occur around this probe site.

Conclusions

Through analysis of functional and sequence data for hundreds of chimeric dioxygenases, we discovered correlations between function and the inheritance of certain sequence elements. One region of sequence around TDO residue 324 (probe 3) essentially determined the substrate specificity and was conserved among chimeras with novel activity toward hexylbenzene. We identified particular combinations of sequence elements that correlate with retention of wildtype function or altered substrate specificities.

This work is the first study of protein sequence-function relationships by parallel analysis of hundreds of chimeric proteins. This approach should be generally useful for identifying sequence elements (or combinations of elements) that determine any measurable protein characteristic (e.g., activity, solubility, stability or substrate specificity). These loci of functional divergence are generally

shrouded by accumulation of functionally neutral mutations, and thus this approach can enable the systematic study of natural molecular evolution of protein families. Insight gained from this type of analysis can be used to inform laboratory evolution experiments and suggest protein engineering approaches.

Materials and Methods

Construction of parent plasmids

Plasmids pJMJ2, pJMJ6 and pJMJ7 (see Chapter 3) contain genes encoding toluene dioxygenase (TDO) [14], tetrachlorobenzene dioxygenase (TCDO) [15] and biphenyl dioxygenase (LB400, referred to here as BPDO) [16], respectively, along with the ferredoxin, reductase and *cis*-dihydrodiol dehydrogenase from the toluene dioxygenase cistron. These plasmids were created by inserting the appropriate genes into the cloning site of the ptrc99A vector by PCR, restriction digestion and ligation. pJMJ2 has since been found to have an insertion of 34 bp located upstream of the cloned genes that presumably arose during PCR of the dioxygenase genes. The inserted sequence contains a KpnI restriction site in addition to the one present already on the vector. Digestion of pJMJ2 with KpnI followed by ligation gave the proper construct that we have named pJMJ11.

Creating chimeric libraries by DNA shuffling

Two chimeric libraries C1 and C2 were constructed for this study. Construction of library C1 was described previously (see Chapter 7). For library C2, DNA was PCR amplified from a plasmid and digested with DNaseI as described in Chapter 7. Ten minutes were allowed for the DNaseI digestion, and gel-extracted fragments used for the reassembly ranged in size from 0.2 to 0.8 kb. Fragments were reassembled in a 20 μl reaction containing 49 ng fragment DNA, 2 μl of 10xPfu buffer (Stratagene, La Jolla, CA), 0.8 μl of dNTP mix (10 mM each, Promega, Madison, WI) and 0.4 μl (1.0 U) of cloned Pfu polymerase (Stratagene, La Jolla, CA). Cycling was according to the following protocol: 96°C, 1.5 minutes, followed by 35 cycles of (94°C for 30 seconds; 58°C for 5 minutes; 72°C for 4 minutes), 72°C for 7 minutes, 4°C thereafter. Full-length chimeric genes were amplified by diluting the reassembly reaction 50x in PCR mixture containing primers and cycling as described in Chapter 7.

Cloning of DNA libraries into a bacterial expression system

DNA libraries were cloned into the KpnI/BamHI sites of pJMJ2. Ligated plasmid was transformed into XL10-Gold competent cells (Stratagene, La Jolla, CA). Transformants were collected and plasmid was purified from 50,000 colonies for

library C1 and 6,000 colonies for library C2. Purified plasmid was used to transform BL21(DE3) competent cells (Stratagene, La Jolla, CA).

Solid-phase screening for activity toward toluene, bromobenzene, styrene, t-butylbenzene, indole and biphenyl

BL21(DE3) transformants were inoculated into 384-well plates and allowed to grow to saturation by shaking at 37°C overnight. A solid-phase screen was performed, in which colonies replicated from these 384-well plates are induced with IPTG and biotransform substrate supplied in the gas phase (see Chapters 5 and 7). Toluene, bromobenzene, styrene, and *t*-butylbenzene were supplied by incubating the colonies with a dish of the substance for 6 minutes, 15 minutes, 15 minutes, and 1 hour, respectively, at 30°C in an airtight container. For the indole and biphenyl activity assays, crystals of the substrate were spread over the lid of the Petri dish containing the colonies and incubated for 1 hour and 2 hours, respectively. For substrates besides indole, Gibbs reagent was used to assay for catechol production (see Chapters 5 and 7). Indole oxidation by dioxygenase leads to indigo and possibly other colored substances. Quantification of the relative activity of each clone (based on relative coloration) was done using the Quantity One software package (Biorad, Hercules, CA) as described in Chapters 5 and 7.

Solid-phase screen for activity toward naphthalene

Colonies grown and induced as described above were exposed to naphthalene crystals for 80 minutes. Active colonies turn bright yellow from accumulation of 1,2-naphthoquinone, which forms spontaneously from 1,2-dihydroxynaphthalene in aqueous media [30]. The bottom of the membrane supporting the colonies was imaged using a desktop scanner, and yellow regions were selected and used to create a new image using Adobe Photoshop (San Jose, CA). This new image was converted to grayscale and exported in .tiff format to Quantity One to quantitate the relative activity of each clone.

Liquid-phase screen for activity toward toluene, fluorobenzene, chlorobenzene, bromobenzene, iodobenzenene, styrene, t-butylbenzene, and biphenyl

Clones active in the solid-phase screen were inoculated into 400μl of LB containing 100 mg/l ampicillin in a 2 ml, square-well, polypropylene 96-well plate (Corning) and grown to saturation. Glycerol was added to 20% (w/v) and the microtiter plates were frozen at -80°C. 5 μl of this culture was used to inoculate a similar plate with each well holding 400 μl of LB containing 100 mg/l ampicillin and 0.4% (w/v) D-glucose. This culture was incubated at 37°C for 14-16 hours in a New Brunswick Scientific Innova® incubator shaker (Edison, NJ) set to 250 rpm. 81 μl of these cultures was then transferred to 1.2 ml of the same LB

medium and incubated for 3 hours at 30°C in a similar incubator. IPTG was then added to 1 mM and the incubation was continued for another 2.5 hours. At this point, the cells were pelleted and resuspended in 1.3 ml of M9-minimal media [31] containing 1.0 mM IPTG, 100 mg/L ampicillin, 1.6% (w/v) D-glucose and 80 mg/L $FeSO_4 \cdot 7H_2O$. 170 μl aliquots were then combined with 8 μl of 0.1 M substrate in ethanol and incubated at 30°C for 1 hour for toluene, fluorobenzene, chlorobenzene, bromobenzene, iodobenzenene, and styrene, 1.5 hours for biphenyl and 2 hours for *t*-butylbenzene. The cells were then pelleted and 100 μl of supernatant was combined with 20 μl of 0.4% (w/v) Gibbs' reagent in ethanol. After 2-4 minutes, the absorbance was recorded at either 550 or 600nm, depending on which wavelength gave the larger absorbance for the particular substrate used.

Solid-phase screen for activity toward hexylbenzene

Screening for activity toward hexylbenzene was done using a method similar to one described in Chapter 5. BL21(DE3) competent cells were transformed with plasmids containing chimeric dioxygenase genes. Transformants were spread on a 22x22cm Luria-Bertani (LB) medium [31] plate containing 1.5% (w/v) bacto-agar, 100 mg/l ampicillin and 0.4% (w/v) D-glucose. Colonies were allowed to grow for 12 hours at 37°C and then transferred to M9-minimal medium [31] containing 4% (w/v) bacto-agar, 0.5 mM IPTG, 100 mg/l ampicillin, 1.6% (w/v) D-glucose, and 80 mg/L $FeSO_4 \cdot 7H_2O$ on a 21x21cm nitrocellulose membrane

(Protran, 0.45μm, Schleicher & Schuell) with the colonies facing up. After 4 hours at 30°C, the colonies were transferred to a similar 22x22 cm M9-minimal medium plate spread the night before with 0.6 ml of 0.5 M hexylbenzene in ethanol. After 90 minutes, the membrane was transferred to a 3% (w/v) agarose plate also containing 0.025% (w/v) Gibbs reagent (added as a 2% solution in ethanol). Active colonies turned bright purple approximately 30 minutes after contact with Gibbs reagent and were recovered by streaking from the original plate used to grow the colonies.

Liquid-phase assay for activity toward hexylbenzene

Colonies that appeared active in solid-phase screening on hexylbenzene were inoculated into LB containing 100 mg/l ampicillin and 0.4% (w/v) D-glucose and grown overnight to saturation. 0.3 ml of this culture was diluted with 4.4 ml of the same medium and incubated at 30°C for 3 hours in a New Brunswick Scientific Innova® incubator shaker (Edison, NJ) set to 250 rpm. IPTG was added to 1 mM and this 3-hour incubation was repeated. The cells were then pelleted and resuspended in 4ml of M9-minimal medium [31] containing 1.0 mM IPTG, 100 mg/L ampicillin, 1.6% (w/v) D-glucose, 80 mg/L $FeSO_4 \cdot 7H_2O$ and 2.5 mM hexylbenzene. After 7.5 hours, the cells were again pelleted and a 100 μl aliquot of supernatant was combined with 20 μl of 0.4% (w/v) Gibbs reagent in ethanol and the absorbance at 550 nm was recorded. A larger aliquot of supernatant was retained for GC-MS analysis.

GC-MS analysis of hexylbenzene oxidation product

250 μl of supernatant from the liquid-phase assay was extracted with 107 μl of ethyl acetate by vigorous mixing followed by phase separation by centrifugation. 50 μl of the ethyl acetate phase was mixed with sodium sulfate crystals, and 40 μl of dried extract was evaporated by vacumn centrifugation. The sample was redissolved in 10 μl of acetone and analyzed by gas chromatography-mass spectrometry (GC-MS, model 5890 gas chromatograph, model 5970 mass selective detector, Hewlett Packard) using a C-18 column (BPX5, 30 meter length, 0.25mm ID, 0.25 micron film, SGE International Pty Ltd.). Temperature was ramped from 70 to 270°C at 12°C/minute.

References

1. Kim, J.-Y. & Devreotes P.N.. (1994). Random chimeragenesis of G-protein-coupled receptors. *J. Biol. Chem.* **269**, 28724-28731.
2. Levin, L.R. & Reed R.R. (1995). Identification of functional domains of adenylyl-cyclase using *in vivo* chimeras. *J. Biol. Chem.* **270**, 7573-7579.
3. Beil, S., Mason, J.R., Timmis, K.N. & Pieper, D.H. (1998). Identification of chlorobenzene dioxygenase sequence elements involved in dechlorination of 1,2,4,5-tetrachlorobenzene. *J. of Bacteriol.* **180**, 5520-5528.
4. Zielinski, M., Backhaus, S. & Hofer, B. (2002). The principal determinants for the structure of the substrate-binding pocket are located within a central core of a biphenyl dioxygenase alpha subunit. *Microbiology* **148**, 2439-2448.
5. Gibson, D.T. & Parales, R.E. (2000). Aromatic hydrocarbon dioxygenases in environmental biotechnology. *Curr. Opin. Biotech.* **11**, 236-243.
6. Sheldrake, G.N. (1992). Biologically derived arene *cis*-dihydrodiols as synthetic building blocks. *Chirality in Industry.* John Wiley and Sons Ltd., New York, New York.
7. Suenaga, H., Mitsuoka, M., Ura, Y., Watanabe, T. & Furukawa, K. (2001). Directed evolution of biphenyl dioxygenase: Emergence of enhanced degradation capacity for benzene, toluene, and alkylbenzenes. *J. Bacteriol.* **183**, 5441-5444.
8. Barriault, D., Plante, M.M. & Sylvestre, M. (2002). Family shuffling of a targeted bphA region to engineer biphenyl dioxygenase. *J. Bacteriol.* **184**, 3794-3800.

9. Kumamaru, T., Suenaga, H., Mitsuoka, M., Watanabe, T. & Furukawa, K. (1998). Enhanced degradation of polychlorinated biphenyls by directed evolution of biphenyl dioxygenase. *Nat. Biotech.* **16**, 663-666.
10. Bruhlmann, F. & Chen, W. (1999). Tuning biphenyl dioxygenase for extended substrate specificity. *Biotechnology and Bioengineering* **63**, 544-551.
11. Zhang, N., Stewart, B.G., Moore, J.C., Greasham, R.L., Robinson, D.K., Buckland, B.C. & Lee, C. (2000). Directed evolution of toluene dioxygenase from *Pseudomonas putida* for improved selectivity toward cis-indandiol during indene bioconversion. *Metabolic Engineering* **2**, 339-348.
12. Sakamoto, T., Joern, J.M., Arisawa, A. & Arnold, F.H. (2001). Laboratory evolution of toluene dioxygenase to accept 4-picoline as a substrate. *Appl. and Environ. Microbiol.* **67**, 3882-3887.
13. Suenaga, H., Goto, M. & Furukawa, K. (2001). Emergence of multifunctional oxygenase activities by random priming recombination. *J. Biol. Chem.* **276**, 22500-22506.
14. Zylstra,G.J. & Gibson,D.T. (1989). Toluene degradation by *Pseudomonas putida* F1 – Nucleotide sequence of the todC1C2BADE genes and their expression in *Escherichia-coli*. *J. Biol. Chem*. **264**, 14940-14946.
15. Beil, S., Happe,B., Timmis,K.N.& Pieper,D.H. (1997). Genetic and biochemical characterization of the broad spectrum chlorobenzene dioxygenase from *Burkholderia sp*. strain PS12-Dechlorination of 1,2,4,5-tetrachlorobenzene. *Eur. J. Biochem.* **247**,190-199.

16. Mondello,F.J. (1989). Cloning and expression in *Escherichia coli* of *Pseudomonas* strain LB400 genes encoding polychlorinated biphenyl degradation. *J. Bacteriol*. **171,** 1725-1732.

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

18. Kauppi,B., Lee,K., Carredano, E., Parales,R.E., Gibson,D.T., Eklund,H. & Ramaswamy, S. (1998). Structure of an aromatic-ring-hydroxylating dioxygenase-naphthalene 1,2-dioxygenase. *Structure* **6**, 571-586.

19. Kikuchi, M., Ohnishi, K. & Harayama, S. (1999). Novel family shuffling methods for the *in vitro* evolution of enzymes. *Gene* **236**, 159-167.

20. Sieber, V., Martinez, C.A. & Arnold, F.H. (2001). Libraries of hybrid proteins from distantly related sequences. *Nat. Biotech.* **19**, 456-460.

21. Coco, W.M., Encell, L.P., Levinson, W.E., Crist, M.J., Loomis, A.K., Licato, L.L., Arensdorf, J.J., Sica, N., Pienkos, P.T. & Monticello, D.J. (2002). Growth factor engineering by degenerate homoduplex gene family recombination. *Nat. Biotech.* **20**,1246-1250.

22. Ness, J.E., Kim, S., Gottman, A., Pak, R., Krebber, A., Borchert, T.V., Govindarajan, S., Mundorff, E.C. & Minshull, J. (2002). Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nat. Biotech.* **20**, 1251-1255.

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

24. Ness, J.E, Welch, M., Giver, L., Bueno, M., Cherry, J.R., Borchert, T.V., Stemmer, W.P.C. & Minshull, J. (1999). DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* **17**, 893-896.

25. Raillard, S., Krebber, A., Chen, Y.C., Ness, J.E., Bermudez, E., Trinidad, R., Fullem, R., Davis, C., Welch, M., Seffernick, J., Wackett, L.P., Stemmer, W.P.C. & Minshull, J. (2001). Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. *Chem. & Biol.* **8**, 891-898.

26. Lutz, S., Ostermeier, M., Moore, G.L., Maranas, C.D. & Benkovic, S.J. (2001). Creating multiple-crossover DNA libraries independent of sequence identity. *PNAS* **98**,11248-11253.

27. Ostermeier, M., Shim, J.H. & Benkovic, S.J. (1999). A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotech.* **17**,1205-1209.

28. Carredano, E., Karlsson, A., Kauppi, B., Choudhury, D., Parales, R.E., Parales, J.V., Lee, K., Gibson, D.T., Eklund, H. & Ramaswamy, S. (2000). Substrate binding site of naphthalene 1,2-dioxygenase: Functional implications of indole binding. *J. Mol. Biol.* **296**, 701-712.

29. Suenaga, H., Watanabe, T., Sato, M., Ngadiman & Furukawa, K. (2002). Alteration of regiospecificity in biphenyl dioxygenase by active-site engineering. *J. Bacteriol.* **184**, 3682-3688.

30. Eaton, R.W. & Chapman, P.J. (1992). Bacterial metabolism of naphthalene: Construction and use of recombinant bacteria to study ring cleavage of 1,2-dihydroxynaphthalene and subsequent reactions. *J. Bacteriol.* **174**, 7542-7554.

31. Sambrook,J., Fritsch,E.F. & Maniatis,T. (1989). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Chapter 9

Empirical comparison of recombination and random mutagenesis as search strategies for enzyme engineering

Abstract

Successful search strategies for laboratory evolution to improve or alter protein function include recombination of homologous genes and recursive random mutagenesis. Here, we compare these two strategies empirically with respect to their ability to produce dioxygenase enzymes with new substrate specificities in one generation. We constructed libraries of both chimeras and point mutants of three dioxygenases and evaluated the libraries on two evolutionary tasks. The first task was to create variants with altered substrate specificity, for which we determined the relative activities of hundreds of clones toward ten substrates accepted by at least one of the three parents. Both recombination and mutagenesis yielded variants with altered specificity, but those from the chimeric libraries appeared with higher frequency and were more functionally diverse. The second task was to generate a variant that could efficiently dihydroxylate *n*-hexylbenzene, a substrate not accepted by any of the parents. Such variants were found only in the chimeric libraries. These two cases show that recombination can be more effective than random mutagenesis in accessing diverse functionalities.

Introduction

Evolutionary search strategies employing recombination and mutation have been effective in various contexts, whether applied explicitly, as in genetic algorithm theory and sexual reproduction, or implicitly as in the process of human invention [1,2]. Effective optimization strategies are built by recursive application of recombination and/or mutation in some intelligent way. In recent years, these evolutionary strategies have been employed in the laboratory to evolve proteins. In these studies, recombination is usually done by DNA shuffling of homologous genes [3-10], while mutation is accomplished by random mutagenesis of a single gene of interest [11,12]**.** Because the effects of mutations are sometimes additive, genes containing beneficial point mutations can be recombined in an effort to generate further improvements [13-15]. Due to the frequently high cost of screening mutant proteins for desired functions, one would like to know *a priori* which strategy is likely to work best for a particular problem.

It has been argued that recombination of homologous proteins has an advantage over mutagenesis in that all the amino acids have already been successful (in the evolutionary sense) in the context of at least one of the parents. Thus, when parents sharing high sequence identity (≥70%) are recombined, a significant percentage of the resulting variants fold and are active [4,5,9,16]. Similar jumps in sequence space made by random mutation (often tens of amino acid substitutions) would yield inactive variants almost exclusively, due to the creation

of stop codons and other cumulative deleterious effects of point mutation. Thus recombination is informed by natural evolution in that many mutations that are catastrophic to protein function do not appear in the library because they have been selected against. Still this insight leaves us ignorant of how best to evolve functionally interesting proteins in the laboratory. Although recombination allows us to make large but conservative jumps in *sequence* space, there is only limited evidence that *functional diversity* is more easily attained with recombination than with random mutagenesis.

Recombination is an effective strategy on fitness landscapes where peaks cluster together, since it allows for exploration of functionally rich regions of sequence space between local optima [17]. Such landscapes are also conducive to mutagenic exploration, and various computational studies have demonstrated synergy between the mutation and recombination operators using model protein fitness landscapes. Using an HP model to simulate sequence-structure relationships, Chan *et al.* found that recombinatoric search finds novel model protein structures more effectively than recursive point mutagenesis: recombination was shown to "tunnel" through areas of low fitness that impede a mutational walk [18]. Xiu and Levitt employed a similar model to demonstrate that when single-crossover recombination is allowed during a mutational walk, more highly optimized sequences are accessed than by mutation alone [19]. Studies of smooth landscapes based on the NK model suggest that recursive

mutagenesis is outperformed by strategies combining recombination and mutagenesis [17,20].

In these studies, the simplicity of the underlying models for mapping fitness to sequence prevents facile application to problems commonly addressed in the laboratory. For instance, in the case of the HP model, novelty is gauged by creation of new structures, whereas functional change in laboratory-evolved proteins is rarely, if ever, accompanied by a significant structural change (e.g., to a different fold). Application of these computational studies to real protein engineering problems is partially an act of faith that the fitness landscape of the model adequately represents the corresponding landscape of the property of interest, and additionally, that one can specify where in the model landscape the natural starting proteins reside.

In the preceding chapter, “Functional genomics of a library of chimeric enzymes,” we characterized two libraries made by recombination of three homologous dioxygenases with respect to 1) substrate specificity and 2) acquisition of activity toward a substrate not measurably accepted by the parent enzymes. In this study, we performed the same tests on three additional libraries, this time made by random point mutagenesis of each of the parent genes. Using these functional data, we can compare recombination (by family shuffling) and point mutagenesis as strategies for the laboratory evolution of functional diversity. In using a single generation, we are conducting an *exploration* of conceivable

functions rather than an *optimization* of a particular function. We discuss how this approach limits our conclusions and to what extent our results apply to a multi-generation search.

Results

Library construction and characterization

Three mutant libraries were constructed by error-prone PCR (see Materials and Methods) of the genes encoding the α and β subunits of three dioxygenase parents, toluene dioxygenase from *Pseudomonas putida* (TDO) [21], tetrachlorobenzene dioxygenase from *Burkholderia sp*. strain PS12 (TCDO) [22], and biphenyl dioxygenase from *Pseudomonas* strain LB400 (BPDO) [23]. The substrate specificities of these three enzymes and the plasmid-based expression system used to produce them in *E. coli* is described in Chapter 8.

Table 1 summarizes some important characteristics of these point mutation libraries and the chimeric libraries to which they will be compared (see Chapter 8). Approximately 40-50% of the clones from the three random point mutagenesis libraries retained function. Two active and two inactive clones were randomly selected from each of the three mutant libraries and sequenced. A total of 28 nucleotide point mutations (23 resulting in amino acid mutations) were observed in these twelve clones (2023-2064 nt from start of α subunit to end of β

subunit), implying a nucleotide mutation rate of 0.11%, or 2.3 per gene on average. In the sequence of one inactive enzyme, we also found a deletion of one base pair. Three mutant clones selected as described later in this study were found to contain a total of ten point mutations. Thus we have observed a total of 38 point mutations from which we can assess biases in base pair changes (see Table 2). Consistent with the findings of Shafikhani *et al.* [24], we found no mutations from A→C or T→G or from C→G or G→C. Of the observed mutations, 61% were from AT base pairs; Shafikhani *et al.* [24] report a similar bias, with 75.6% of mutations from AT base pairs. Transitions accounted for 74% of our observed mutations, while Shafikhani *et al.* [24] report only 43.2%.

library	method	% active*	mutation rate (%)	crossovers/gene
C1	DNA shuffling	35.5%	0.011±0.005	3.7±0.3
C2	DNA shuffling	35.3%	n/d	2.0±0.2
TDO mutants	Error-prone PCR	48.8%	0.11±0.03**	n/a
TCDO mutants	Error-prone PCR	51.9%	0.11±0.03**	n/a
BPDO mutants	Error-prone PCR	39.0%	0.11±0.03**	n/a

* Percent retaining at least 20% of the activity of the most active parent toward either toluene, bromobenzene, biphenyl or indole, as determined by solid-phase screening. (n > 240)

** Mutation rate is averaged over a total of 12 mutants. Two active and two inactive clones were selected from each of the three mutant libraries.

Table 1. Summary of library statistics. The mutant libraries have approximately 10-fold more point mutations than chimeric library C1. On average, 2.3 nucleotide and 1.9 aa mutations per clone occurred in the mutant libraries. Assuming a Poisson distribution, we expect that only 15% of clones from the mutant libraries are wildtype at the amino acid level, while 86% of the chimeras from C1 have no point mutations.

Mutation type	# of instances
A→T & T→A	6
A→C & T→G	0
A→G & T→C	17
G→A & C→T	11
G→C & C→G	0
G→T & C→A	4
Total →	38
A→N & T→N	23
G→N & C→N	15

Table 2. Mutations observed in 15 sequenced mutant clones.

Screening for altered substrate specificity and data analysis

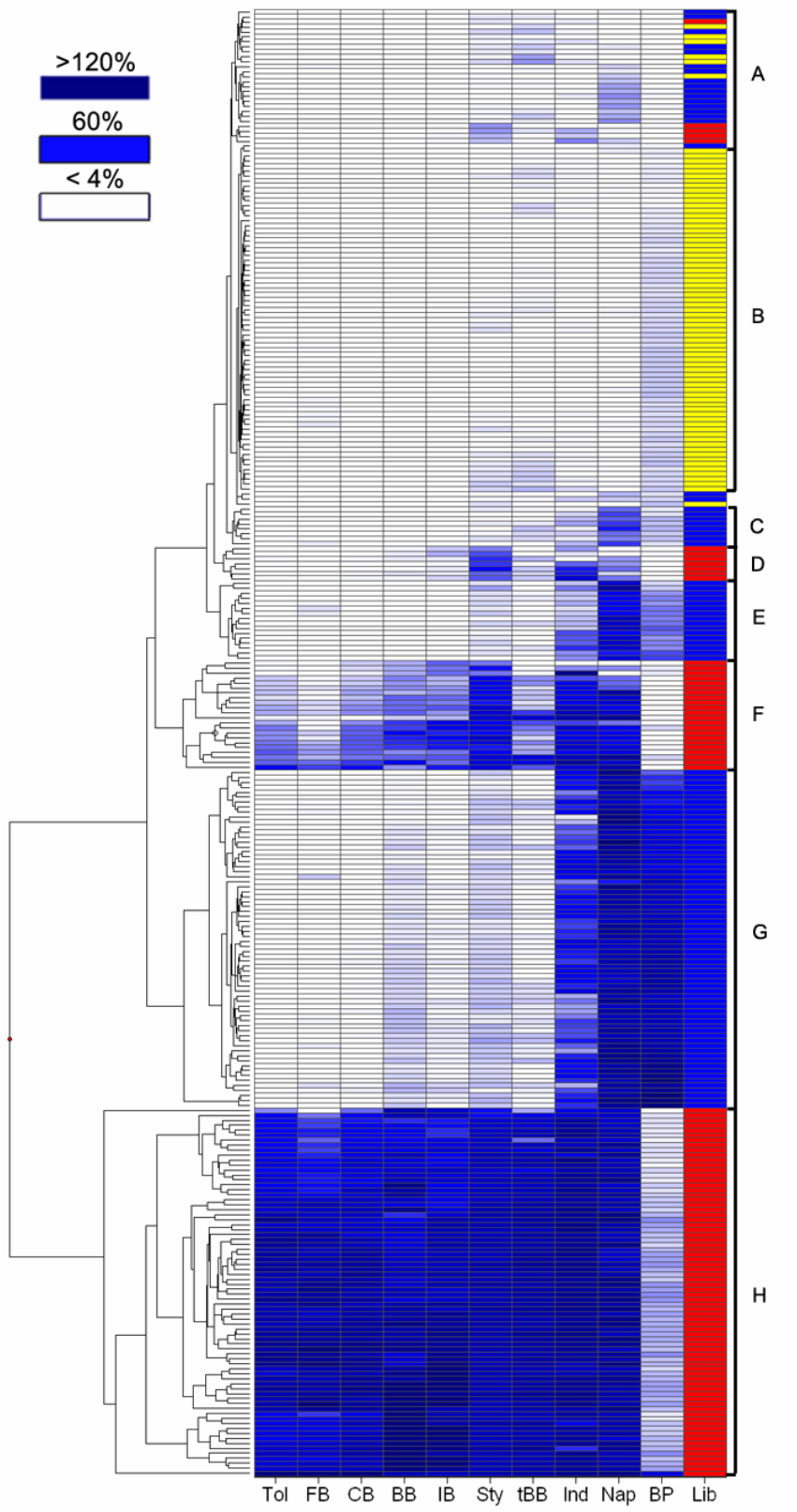
We screened more than 200 clones from each mutant library for activity toward toluene, fluorobenzene, chlorobenzene, bromobenzene, iodobenzene, styrene, *t*-butylbenzenene, indole, naphthalene and biphenyl, as described in Chapter 8 (see Table 3). Based on these data, we selected 15 clones from the mutant libraries with altered substrate specificities and rescreened them contemporaneously with selected chimeric enzyme variants, as described in Chapter 8. Eleven out of the 15 mutants showed the expected altered specificity after rescreening.

Library	# screened	# active*	# screened in liquid media
chimeric C1	364	124	97
chimeric C2	363	123	109
total →	727	247	206
TDO mutant	248	116	116
TCDO mutant	229	111	111
BPDO mutant	241	78	78
total →	718	305	305

* Clones were considered active if they retained at least 25% of the activity of the most active parent

Table 3. Summary of screening for activity toward toluene, fluorobenzene, chlorobenzene, bromobenzene, iodobenzene, styrene, *t*-butylbenzene, indole, naphthalene and biphenyl. Variants were solid-phase screened to determine retention of activity; then, all the active mutants and the subset of active chimeras with complete probe hybidization data were screened in liquid media, as described in the text.

The primary activity data for each mutant clone on ten substrates were organized using hierarchical clustering as done in Chapter 8 for the chimeric libraries. The functional profiles of the mutant clones were clustered based on their relative activities toward the ten substrates tested and then projected graphically using a heatmap representation, as shown in Figure 1. A similar chart for two chimeric libraries is presented in Chapter 8.



(Legend on following page)

Figure 1. Heatmap representation of hierarchical clustering results for the three mutant libraries. Substrate specificity profiles for active clones were scaled and clustered as described in Chapter 8. For each substrate, the relative activity of each clone is represented using shades of blue (see below), where white is low activity and blue is high activity. Columns representing relative activity toward a particular substrate are labeled with the following abbreviations: toluene (Tol), fluorobenzene (FB), chlorobenzene (CB), bromobenzene (BB), iodobenzene (IB), styrene (Sty), *t*-butylbenzene (tBB), indole (Ind), naphthalene (Nap) and biphenyl (BP). The parentage of the mutants is shown in the "Lib" column; TDO, TCDO and BPDO are represented by red, blue and yellow, respectively. This figure was created using Spotfire® (Cambridge, MA).

From Figure 1, we see that the substrate specificities of the mutant clones are highly correlated to the starting parent shown in the column labeled "Lib." Clones functionally similar to wildtype TDO, TCDO, and BPDO appear in regions H, G and B, respectively. Although the percentage of active clones is similar for the chimeric and mutant libraries (see Table 1), the percentage of active clones with wildtype specificity and activity is significantly higher for the mutants than for the chimeras (approximately 69% for the mutant libraries and 42% for the chimeric libraries). Mutants in regions C and E correspond to clones with TCDO-like specificity but decreased activity. In regions D and F we observe TDO mutants with a preference for styrene and large halobenzenes over fluorobenzene; these clones have several functional counterparts in the chimeric libraries.

In Figure 2 we show activity profiles for the 24 chimeras and 11 point mutants that exhibited a confirmed altered specificity (data shown are from the rescreening). The clones are grouped as described in the companion paper, but here we show both mutants and chimeras. The specificity shown in Figure 2(a) (no activity toward toluene, fluorobenzene or *t*-butylbenzene; low activity toward

other TDO or TCDO substrates) was found several times in the chimeric libraries, but not among the point mutants. Clones shown in Figure 2(b) are similar to those in (a) except they have some activity toward toluene and in general are more active; this specificity was common in both point mutant and chimeric libraries. One mutant (M55) was found with improved overall activity and high activity toward *t*-butylbenzene (Figure 2(c)); three chimeras had similar functional profiles. At least two distinct specificities are shown in Figure 2(d) (clones with a preference for *t*-butylbenzene over toluene), and these three clones are all chimeras. Both a chimera and a point mutant were found with high activity toward indole and naphthalene, but little or no activity toward smaller substrates or biphenyl. In this case, the mutant M112 had the more divergent specificity.

Mutants M38, M55 and M112 (see Figure 2) were sequenced. All are mutants of TDO. M38 contained one mutation, leading to amino acid substitution N168S, on its α subunit. M55 contained seven nucleotide mutations, two of which result in amino acid changes on the α subunit: T249A and S369G. G369S occurred during cloning of the TDO parent (see Chapter 3), so S369G represents a reversion to the original amino acid. M112 contained two mutations in DNA encoding the β subunit, resulting in R66W and I129T. To our knowledge, none of these positions has previously been shown to have an effect on dioxygenase function.

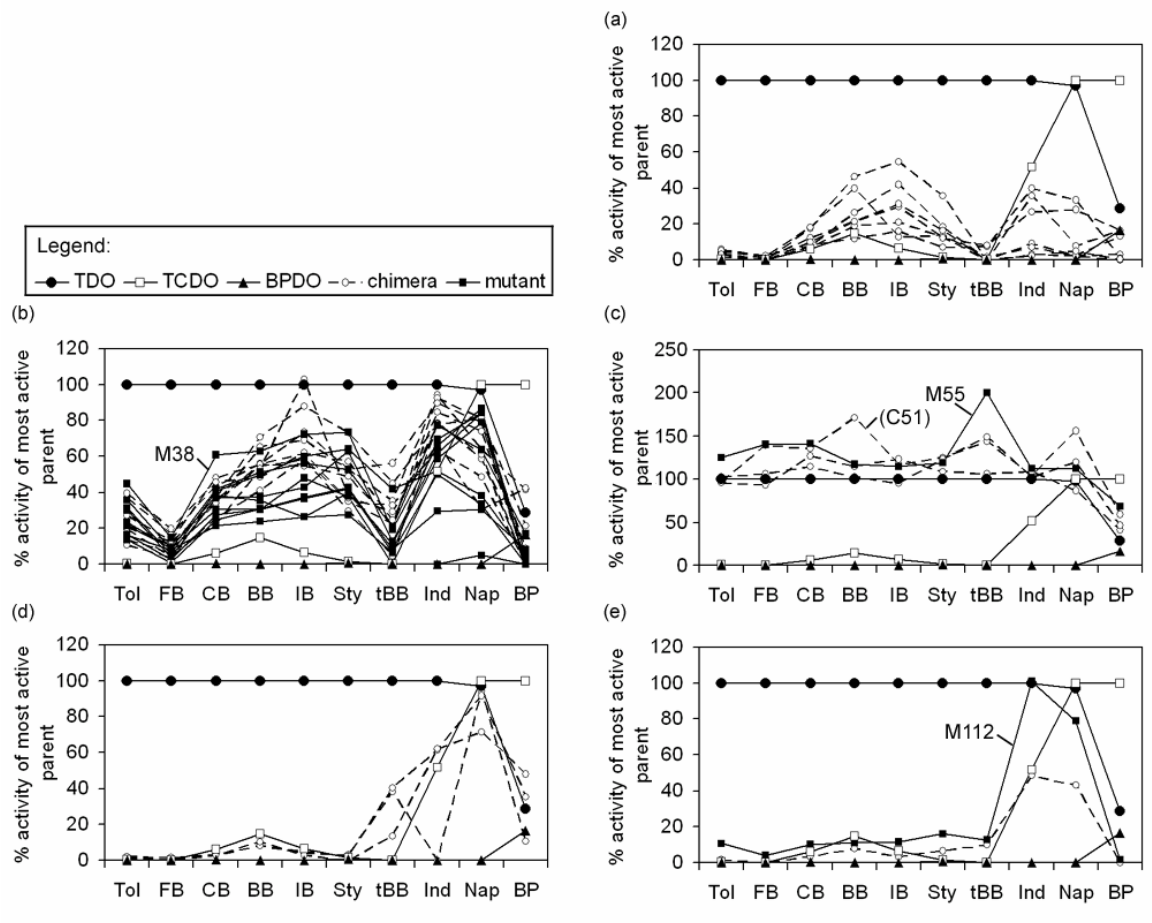


Figure 2. Comparison of functional profiles for chimeras and mutants with altered substrate specificity. Selected mutants and chimeras with altered substrate specificity are compared to wildtype TDO, TCDO and BPDO. Grouping of similar clones into five clusters (a)-(e) was informed by K-means clustering. Substrates are abbreviated as follows: toluene (Tol), fluorobenzene (FB), chlorobenzene (CB), bromobenzene (BB), iodobenzene (IB), styrene (Sty), *t*-butylbenzene (tBB), indole (Ind), naphthalene (Nap) and biphenyl (BP).

Principal component analysis

To view and compare data for the chimeric and mutant enzymes, we have used principal component analysis as performed by the data analysis package

Spotfire® (Cambridge, MA) to reduce the dimensionality of the substrate specificity data. In fact, the ten-dimensional dataset composed of relative activities for the active chimeras and mutants can be reduced to two dimensions (two principal components) with loss of only 9.9% of the information contained in the original dataset. In Figure 3, these two principal components are plotted for the active chimeras and mutants. Groups of clones with similar function as suggested by K-means clustering are assigned different colors: the wildtype TDO, TCDO and BPDO parents are found in the red, blue, and yellow clusters, respectively. A large fraction of the mutants fall in these regions, especially in the TDO and TCDO clusters. Regions of Figure 3 containing mostly clones with altered specificity are shaded, and approximately 50% more chimeras (circles) than mutants (triangles) are found in these regions.

We find that both mutation and recombination give rise to dioxygenase variants with altered substrate specificity. Such clones were approximately twice as frequent in the chimeric libraries, and overall more new specificities were observed in the chimeric libraries than in the mutant libraries. All of the mutants with altered specificity were mutants of TDO, which has a broad specificity based on the substrates tested. In Figure 2, we see that, with the exception of the mutant M55, the evolved mutants are more specific and less active than this parent. Mutation of TCDO and BPDO did not expand their specificity to substrates not accepted by these parents.

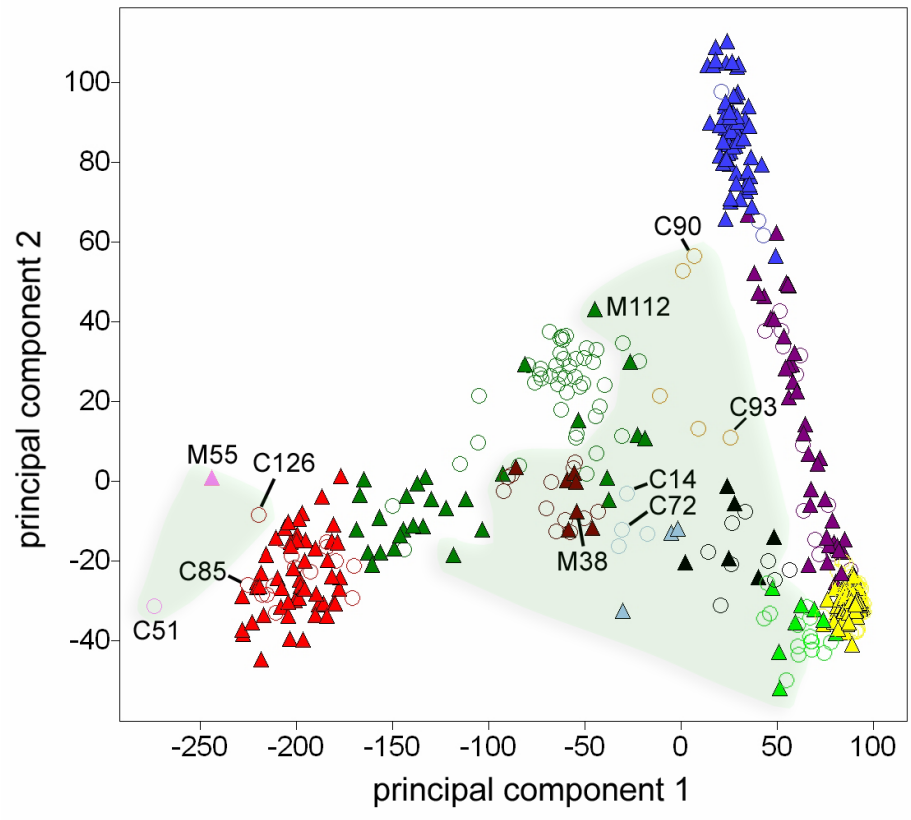


Figure 3. Projection of 10-D substrate specificity data for active chimeras (○) and mutants (▲) into two dimensions defined by principal components 1 and 2. Groups of functionally similar clones (as suggested by K-means clustering) are represented as different colors. Red, blue and yellow represent the wildtype TDO, TCDO and BPDO functionalities; clones with specificity similar to TDO but with decreased activity are colored dark green; low-activity, TCDO-like clones are colored purple. Regions containing a prevalence of clones with altered specificity are shaded. Principal components analysis, K-means clustering and plotting were performed using Spotfire® (Cambridge, MA).

Evolution of improved total activity

In the primary screening of both chimeras and mutants, several clones were found to have modest increases in total activity toward particular substrates relative to the most active parent clone. To assess the reproducibility of these apparent changes, four chimeras and seven mutants exhibiting a >35% increase

in total activity toward toluene, fluorobenzene, bromobenzene or iodobenzene were rescreened using liquid cultures as described in Chapter 8. No clones with improvements for the other tested substrates (styrene, *t*-butylbenzene, indole and naphthalene) were found. This is perhaps due to nonlinearities associated with the solid-phase screen used to obtain these data[**].

Two of the chimeras (C47 and C51) and one of the mutants (M55) were found to have reproducible improvements of at least 35% in total activity. C47 is 44% more active toward iodobenzene than TDO. C51 and M55 are both 40% more active toward fluorobenzene than TDO, and C51 is 71% more active than TDO toward bromobenzene. Figure 2(c) shows complete functional profiles for M55 and C51. According to the primary screening data, C47 prefers benzenes with large halogen substituents and has significantly less activity than TDO toward toluene, fluorobenzene and biphenyl. In these three cases, improvements in total activity were accompanied by a change in substrate specificity; this suggests that we made changes affecting specific enzyme function rather than expression level.

[**] For clones with decreased total activity, the solid-phase screen yields higher relative activity values than the liquid-phase screen, as discussed [19]. For clones with increased total activity, the solid-phase screen generally underestimates the extent of improvement. For example, clone M55 showed no improved activity toward t-butylbenzene in the initial solid-phase screen, but showed a 2-fold improvement upon liquid-phase screening toward this substrate. Solid-phase screening for activity toward toluene ranked M55 as the second most active mutant toward this substrate (10% improvement over TDO); this was prescriptive of a 32% increase in activity toward toluene observed during subsequent liquid-phase screening.

Acquisition of activity toward n-hexylbenzene

Several thousand variants from chimeric library C1 and each of the three mutant libaries were screened for activity toward *n*-hexylbenzene using a solid-phase Gibbs' assay, as described in Chapter 8. The results are summarized in Table 4. None of the mutants was confirmed to have activity toward this substrate. In contrast, 13 chimeras were active toward *n*-hexylbenzene and had varied sequences and substrate specificities (see Chapter 8). Twelve of these 13 chimeras contained no nucleotide point mutations in coding regions; thus the acquisition of hexylbenzene activity can be accomplished solely by recombination.

Library	# screened	# positive for activity toward hexylbenzene	# with confirmed activity
chimeras (C1)	7900	16	13
TDO mutants	9600	1	0
TCDO mutants	5100	0	0
BPDO mutants	6100	0	0

Table 4. Summary of screening for activity toward hexylbenzene.

Our characterization of the mutant libraries can be used to assess whether screening additional mutants for activity toward hexylbenzene would have been successful. Considering the degeneracy of the genetic code and the fact that codons mutated by random mutagenesis generally contain a single nucleotide substitution, we calculate that approximately 3,600 amino acid mutations of each

parental dioxygenase were possible. Because of biases inherent to error-prone PCR with Taq polymerase (see Table 2 and ref. [24]), perhaps 35-50% of these possible mutations will occur with 3- to 8-fold less frequency than those that are preferred. Our 12 sequenced clones from the mutant libraries have 1.9 aa mutations on average, so we estimate that ~28% of the mutants had just one aa mutation and ~57% had multiple mutations (assuming a Poisson distribution). Since 5,100-9,600 mutants (representing ~9,700-18,200 aa mutations) from each library were screened for hexylbenzene activity, most of the preferred mutations have been sampled independently, and almost all (perhaps 90%) have at least been sampled along with other mutations. Thus further screening might be expected to meet with diminishing returns, since new mutations would be scarce.

Discussion

By screening similarly sized libraries of chimeras and mutants for activity toward ten substrates, we found that, although both techniques gave rise to enzymes with altered specificities, the chimeric library was more functionally diverse and had a higher frequency of enzymes with altered specificity. Only recombination was successful on the task of evolving enzymes with activity toward hexylbenzene, a substrate not accepted by the parent enzymes.

Significance of results: important considerations

These results show that recombination can outperform mutagenesis as a strategy for accessing functional diversity, but caveats should be mentioned that may limit the applicability of these results. Biases in both recombination and mutagenesis as they are applied experimentally should be considered. With random mutagenesis, different mutations often occur with different frequencies, e.g., interconversion between G and C is disfavored [24,25]. In our libraries, we see evidence for such biases (see discussion under *Library construction and characterization* and Table 2*)*. These biases serve to reduce the diversity accessible by random mutagenesis. Other methods such as codon mutagenesis [26], high mutation rates [27], mutated polymerases [25] and mutator strains [28] may or may not have increased the functional diversity of our mutants. The crossovers in our chimeric libraries were biased to locations with high sequence identity, preventing the low-identity BPDO parent from forming complex constructs with the other two parents, and, in addition, certain parents were preferred at particular positions (see Chapters 7 and 8). These biases limit sequence diversity and thus affect the evolutionary search for functional diversity.

Another issue is the functional plasticity of the dioxygenase enzyme family. The role of these enzymes in nature is in creating nutrients for the cell from aromatic compounds. Since the identity and availability of these compounds is in constant flux, it may be that evolution selects for dioxygenases that adapt readily to new

substrates, and thus this family may be more “evolveable” than proteins under more constant selection pressure. The functional plasticity of the dioxygenases is evidenced by studies demonstrating impressive changes in enzyme properties through laboratory engineering [29-36] and has been noted by other authors [10]. The results of this study thus may not extend to proteins in general.

Implications for the selection of a search strategy for laboratory evolution

Because of the cost of experimental screening, a reasonable definition for an optimal search strategy is the one that results in the most-improved variant for a given number of screens conducted. In all cases, this optimal strategy will depend on the underlying fitness landscape [17]. Thus we expect that protein properties such as thermostability, tolerance to organic solvents, or solubility evolve on different landscapes than the functional landscapes investigated here. Another factor to consider when choosing an evolutionary search strategy is the size of the library that can reasonably be screened. There is no doubt that a mutant library with an average of twenty mutations per gene contains more functional diversity than one with two mutations per gene, but if 10^{10} screens must be performed to access a particular function of interest from the more diverse library, then it is useless for all practical purposes. Thus there should be a tradeoff between the number of screens performed and the quality of the best variants that emerge.

In our study, we used a single generation to compare the exploratory potential of random mutagenesis and recombination. For the optimization of a particular property, a multi-generation search is preferable on landscapes where mutations generally have additive effects. For a multi-generation search where small functional changes must be accumulated, an important practical consideration is the sensitivity of the screen, which must be high enough to reliably capture these small changes. For instance, it may be possible to evolve a dioxygenase with 6X7-like activity toward hexylbenzene through multiple generations of random mutagenesis, but in our case our screen was not sensitive enough to observe any improvements that might have occurred in the first generation of mutants.

When evolving improved activity toward a substrate that requires a difficult assay, one strategy for reducing library size is to first prescreen the library for retention of activity toward a substrate that allows for extremely high throughput, and then screen active clones for improved activity toward the substrate of interest. One example of how this approach can fail by eliminating interesting clones during prescreening is in our evolution of hexylbenzene-active dioxygenases. In this case, if our libraries were prescreened for activity toward any of the substrates besides biphenyl, the most active clone toward hexylbenzene, 6X7, would have been filtered out during prescreening. Thus evolved clones with narrow substrate specificity are likely to be lost using such a two-tiered approach.

A common goal of laboratory evolution experiments is to improve the rate of conversion of substrate by the cellular biocatalyst. Improvements in total activity can result from the enhancement of enzyme kinetics, altered expression level, increased enzyme stability, or decreased enzyme toxicity. To generate significant improvements, multiple generations of random mutagenesis are often required. In our single-generation study, both random mutagenesis and recombination gave rise to one or two variants with increased total activity that could have been used to parent a second generation. Other random mutagenesis experiments with TDO show improvements in total activity toward toluene and 4-picoline [30] as well as indene [29] in a single generation. Commonly, mutations that give rise to improvements in total activity are additive, and thus we expect further success with a multi-generation search. On the other hand, the advantage of recombining evolved *chimeras* has not been extensively investigated. The number of mutants we screened for improvements in total activity was small relative to other studies [29,30], and only 248 mutants of the broad-activity TDO parent were screened. As a result, few improved clones were isolated, and we could not determine whether recombination or mutagenesis was the better strategy for improving total activity.

Localization and "between-ness" in catalytic task space

The concept of catalytic task space [17] was introduced as a framework for understanding the evolution of novel catalytic activities. In this multidimensional

space, all chemically possible reactions are organized such that similar reactions neighbor each other, and thus an enzyme catalyzes a "ball" of tasks in the space. In its original conception, "similar" reactions (e.g., addition of oxygen into an alkylbenzene to yield a *cis*-dihydrodiol) are mapped to the same catalytic task. For our purposes, each possible reaction is represented as a catalytic task, and thus we can speak of a subset of the entire space that contains all dioxygenation reactions. These reactions are arranged such that chemically similar reactions (as gauged by reaction type and substrate structure) are said to be close, or localized, in the space. If we think of the "ball" of tasks catalyzed by a particular dioxygenase as a density distribution overlying dioxygenase substrates where activity was observed, we can visualize a shift in specificity as a shift in the distribution, and acquisition of a new activity as an extension of the distribution.

In this study, the substrate specificity data represent a partial mapping of this distribution of catalytic tasks for each of the enzyme variants. From the types of specificity that were observed, it is apparent that distribution shape is highly constrained. If the relative activity data were represented discretely (active or inactive) rather than as a continuous scale, we could specify 2^{10} = 1024 distinct specificities. But few of these were observed experimentally. In fact, from the principal components analysis, we find that the data can be represented with minimal information loss in only two dimensions, and, even within those two dimensions, certain areas were inaccessible to both search strategies, as they were applied here (see Figure 3). This implies that some of the reactions

investigated are localized in catalytic task space such that, due to chemical similarity, a single enzyme can not catalyze one but not the other. Furthermore, we expect that pairs of substrates that are easily discriminated represent reactions that are delocalized in the space.

Our functional data for clones with altered specificity reflect the relative positions of the different substrates in the space. For example, no evolved enzymes could convincingly discriminate styrene, bromobenzene or iodobenzene; these reactions should be localized in catalytic task space due to the similar size and polarity of the substrates. Relative activities toward chlorobenzene were often between those for fluorobenzene and bromobenzene, implying that chlorobenzene resides between these substrates in the space. That we found several variants with activity toward *t*-butylbenzene but not toluene (Figure 2(d)) (or increased specificity for *t*-butylbenzene as in Figure 2(c)) implies that these substrates are delocalized. Indole and naphthalene activities were generally correlated, although clone C93 could completely discriminate these substrates (Figure 2(d)). Since they are chemically dissimilar, biphenyl and the halogenated benzenes are expected to be distant in catalytic task space, and, in fact, these substrates are readily discriminated by the parent enzymes as well as many of the variants.

The chimeras isolated with novel activity toward hexylbenzene provide a further example of "between-ness" in catalytic task space. In many of these clones we

observed increased specificity toward benzenes with large, hydrophobic substituents such as *t*-butylbenzene and biphenyl. Thus we can think of these substrates as existing between the "natural" substrates and the new substrate, hexylbenzene. This suggests an alternate pathway to this novel specificity of first evolving increased activity toward a similar substrate within the range of the parent enzymes (e.g., *t*-butylbenzene or biphenyl), and then using further evolutionary tuning (if necessary) to finally acquire activity toward the new substrate. Using such an approach, it is possible that recursive mutation could be used to create enzymes active toward hexylbenzene.

Conclusions

Both mutation and recombination gave rise to dioxygenase variants with altered substrate specificities, but no mutants were found that were active toward a substrate not accepted by the parent enzymes. Due to the close chemical similarity of the substrates investigated, a great deal of functional similarity was observed among our functionally novel clones. We clearly see that similar substrates often were not discriminated by any of the enzyme variants. Evidently, such discrimination represents a particularly difficult evolutionary task.

Several important caveats plague any empirical comparison of search strategies, including biases inherent to random library construction and our inability to fully investigate all library construction parameters (e.g., mutation rate, crossover rate,

choice of parents). Despite these limitations, our findings support the idea that family shuffling is a powerful search strategy for evolution of enzymes with new functions, and is possibly more effective than random mutagenesis. This is the first comparison of these strategies carried out in the same laboratory on the same problem using well-characterized libraries. Our hope is that as more similar studies are carried out, our selection of a search strategy for a particular enzyme engineering problem will become more rational.

Materials and Methods

Parent plasmids encoding the wildtype dioxygenase genes were constructed as described in Chapter 8. All functional screening was done as described in Chapter 8.

Creating mutant libraries by error-prone PCR

Error-prone PCR was used to create a point mutant library from each dioxygenase parent. High-fidelity PCR was used to amplify a region of each of the parent plasmids containing the dioxygenase genes (e.g., *tod*C1C2) and the surrounding region. This PCR product was purified and used as a template for the error-prone reaction. A 100 µl reaction contained 10 ng of template DNA, 0.7 µl of 1M $MgCl_2$, 40 pmol each of a forward and reverse primer (5' – GGAATTCGAGCTCGGTACCAGGA – 3' and 5' –

GTCATGACATCACCTAGGGATCC – 3', respectively), 10 µl of Taq PCR buffer (Perkin Elmer), 1.5 µl of 5 mM $MnCl_2$, 2 µl of dNTP mix (10 mM each, Promega, Madison, WI) and 0.5 µl of Taq polymerase (Perkin Elmer). Cycling was carried out under the following conditions: 94°C, 3 minutes, followed by 30 cycles of (94°C for 30 seconds; 48°C for 30 seconds; 72°C for 2 minutes), 72°C for 10 minutes, 4°C thereafter. The PCR product was cloned into pJMJ11 as described in Chapter 8, and each library comprised at least 9,000 colonies.

References

1. Nelson, R.S. & Winter, S. (1982). *An Evolutionary Theory of Economic Change.* Belknap Press, Cambridge, MA.
2. Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science* **47**, 117-132.
3. Crameri, A., Raillard, S.A., Bermudez, E. & Stemmer, W. P. C. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288-291.
4. Christians, F.C., Scapozza, L., Crameri, A., Folkers, G. & Stemmer, W.P.C. (1999). Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nat. Biotechnol*. **17**, 259-264.
5. Ness, J.E., Welch, M., Giver, L., Bueno, M., Cherry, J.R., Borchert, T.V., Stemmer, W.P.C. & Minshull, J. (1999). DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* **17**, 893-896.
6. Stemmer, W.P.C. (1994). Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389-391.
7. Stemmer, W.P.C. (1994). DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA* **91**, 10747-10751.
8. Hansson, L.O. & Mannervik, B. (2000). Use of chimeras generated by dna shuffling: Probing structure-function relationships among glutathione transferases. *Methods in Enzymology* **328**,463-477.

9. Chang, C-C.J., Chen, T.T., Cox, B.W., Dawes, G.N., Stemmer, W.P.C., Punnonen, J. & Patten, P.A. (1999). Evolution of a cytokine using DNA family shuffling. *Nat. Biotechnol.* **17**, 793-797.

10. Raillard, S., Krebber, A., Chen, Y.C., Ness, J.E., Bermudez, E., Trinidad, R., Fullem, R., Davis, C., Welch, M., Seffernick, J., Wackett, L.P., Stemmer, W.P.C. & Minshull, J. (2001). Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. *Chem. & Biol.* **8**, 891-898.

11. Rai, G.P., Sakai, S., Florez, A.M., Mogollon, L. & Hager, L.P. (2001). Directed evolution of chloroperoxidase for improved epoxidation and chlorination catalysis. *Adv. Synth.Catal.* **343**, 638-645.

12. Bosma, T., Damborsky, J., Stucki, G. & Janssen, D.B. (2002). Biodegradation of 1,2,3-trichloropropane through directed evolution and heterologous expression of a haloalkane dehalogenase gene. *Appl. Environ. Microb.* **68**, 3582-3587.

13. Giver, L., Gershenson, A., Freskgard, P.O. & Arnold, F.H. (1998). Directed evolution of a thermostable esterase. *PNAS* **95**,12809-12813.

14. Wintrode, P.L., Miyazaki, K. & Arnold, F.H. (2000). Cold adaptation of a mesophilic subtilisin-like protease by laboratory evolution. *J. Biol. Chem.* **275**, 31635-31640.

15. Moore, J.C., Jin, H.M., Kuchner, O. & Arnold, F.H. (1997). Strategies for the *in vitro* evolution of protein function: Enzyme evolution by random recombination of improved sequences. *J. Mol. Biol*. **272**,336-347.

16. Abècassis, V., Pompon, D. & Truan, G. (2000). High efficiency family shuffling based on multi-step PCR and *in vivo* DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. *Nucleic Acids Res*. **28**, e88.
17. Kauffman, S. (1993). *The Origins of Order:Self-Organization and Selection in Evolution.* Oxford University Press, New York.
18. Cui, Y., Wong, W.H., Bornberg-Bauer, E. & Chan, H.S. (2002). Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *PNAS* **99**, 809-14.
19. Xia, Y. & Levitt, M. (2002). Roles of mutation and recombination in the evolution of protein thermodynamics. *PNAS* **99**, 10382-7.
20. Bogarad, L.D. & Deem, M.W. (1999). A hierarchical approach to protein molecular evolution. *PNAS* **96**, 2591-5.
21. Zylstra,G.J. & Gibson,D.T. (1989). Toluene degradation by *Pseudomonas putida* F1 – Nucleotide sequence of the todC1C2BADE genes and their expression in *Escherichia-coli*. *J. Biol. Chem*. **264**, 14940-14946.
22. Beil, S., Happe,B., Timmis,K.N.& Pieper,D.H. (1997). Genetic and biochemical characterization of the broad spectrum chlorobenzene dioxygenase from *Burkholderia sp*. strain PS12-Dechlorination of 1,2,4,5-tetrachlorobenzene. *Eur. J. Biochem.* **247**,190-199.
23. Mondello,F.J. (1989). Cloning and expression in *Escherichia coli* of *Pseudomonas* strain LB400 genes encoding polychlorinated biphenyl degradation. *J. Bacteriol*. **171,** 1725-1732.

24. Shafikhani, S., Siegel, R.A., Ferrari, E. & Schellenberger, V. (1997). Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-Based plasmid multimerization. *Biotechniques* **23**, 304-310.

25. Cline, J. & Hogrefe, H. GeneMorph™ PCR mutagenesis kit produces a unique mutational spectrum. (http://www.stratagene.com/vol13_4/p157-162.htm).

26. Murakami, H., Hohsaka, T. & Sisido, M. (2002). Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nat. Biotechnol.* **20**, 76-81.

27. Daugherty, P.S., Chen, G., Iverson, B.L. & Georgiou, G. (2000). Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *PNAS* **97**, 2029-2034.

28. Greener, A., Callahan, M. & Jerpseth, B. (1997). An efficient random mutagenesis technique using an *E. coli* mutator strain. *Molecular Biotechnol.* **7**,189-195.

29. Zhang, N., Stewart, B.G., Moore, J.C., Greasham, R.L., Robinson, D.K., Buckland, B.C. & Lee, C. (2000). Directed evolution of toluene dioxygenase from *Pseudomonas putida* for improved selectivity toward *cis*-indandiol during indene bioconversion. *Metabolic Engineering* **2**, 339-348.

30. Sakamoto, T., Joern, J.M., Arisawa, A. & Arnold, F.H. (2001). Laboratory evolution of toluene dioxygenase to accept 4-picoline as a substrate. *Appl. and Environ. Microbiol.* **67**, 3882-3887.

31. Suenaga, H., Goto, M. & Furukawa, K. (2001). Emergence of multifunctional oxygenase activities by random priming recombination. *J. Biol. Chem.* **276**, 22500-22506.

32. Suenaga, H., Mitsuoka, M., Ura, Y., Watanabe, T. & Furukawa, K. (2001). Directed evolution of biphenyl dioxygenase: Emergence of enhanced degradation capacity for benzene, toluene, and alkylbenzenes. *J. Bacteriol.* **183**, 5441-5444.

33. Suenaga, H., Watanabe, T., Sato, M., Ngadiman & Furukawa, K. (2002). Alteration of regiospecificity in biphenyl dioxygenase by active-site engineering. *J. Bacteriol.* **184**, 3682-3688.

34. Barriault, D., Plante, M.M. & Sylvestre, M. (2002). Family shuffling of a targeted bphA region to engineer biphenyl dioxygenase. *J. Bacteriol.* **184**, 3794-3800.

35. Kumamaru, T., Suenaga, H., Mitsuoka, M., Watanabe, T. & Furukawa, K. (1998). Enhanced degradation of polychlorinated biphenyls by directed evolution of biphenyl dioxygenase. *Nat. Biotechnol.* **16**,663-666.

36. Bruhlmann, F. & Chen, W. (1999). Tuning biphenyl dioxygenase for extended substrate specificity. *Biotechnology and Bioengineering* **63**,544-551.

Appendix

Calculation of the actual average number of crossovers from probe hybridization data

(Supplemental material for Chapter 7)

Preface

This appendix describes a method for calculating the actual number of crossovers occurring in a chimeric library given a set of probe hybridization data. Using the probe approach, crossovers can be missed when two or more crossovers occur between probe sites. Thus a correction is required to estimate the actual number of crossovers. This calculation is nontrivial when more than two parents have been shuffled. The following section of this Appendix describes the calculation, which can be performed using the Matlab code provided. These materials are also available online at http://cheme.caltech.edu/groups/fha/probes/crossovers.html

Calculation of N_{abx} from probe data

We first assume that crossovers are uniformly distributed between neighboring probe pairs. Define P_{abx} to be the probability that nucleotide $x+1$ is from parent b given that nucleotide x is from parent a. Let N be the number of base pairs between two neighboring probes, c be an index for the number of crossovers that occur over N and n_p be the number of parents that were shuffled. Because the probe data gives us information for the neighboring probe positions, we can calculate the probability that the first nucleotide is from parent a, P^m_a, ("m" is measured) and the probability that nucleotide 1 comes from parent a and nucleotide N comes from parent b, P^m_{ab}. Vectors S and T of length $c+1$ are used

to represent a particular chimera; S is the series of parents that occur, and T is the series of crossover locations plus the end condition $T_{c+1} = N$. Given these parameters, we can calculate the probability P^{ST} of constructing a chimera with arbitrary representative vectors S and T (eqn. 1).

$$(1)\quad P^{ST} = P^{m}_{S_1 X} \prod_{b \neq S_1} (1 - P_{S_1 b X})^{T_1 - 1} \prod_{j=1}^{c} \left[P_{S_j S_{j+1} X} \prod_{b \neq S_{j+1}} (1 - P_{S_{j+1} b X})^{T_{j+1} - T_j - 1} \right]$$

By summing the probabilities P^{ST} for all the S and T combinations that result in observing parent a at probe position X and parent b at probe position X+1, we can determine the probability P^{s}_{abX} ("s" is simulated) of observing parent a at probe position X and parent b at probe position X+1. This is accomplished in eqn. 2 by summing over the number of crossovers, the set $\{S\}_{abc}$ of all parent combinations consistent with given c,a, and b, and the set $\{T\}_c$ of all crossover location combinations.

$$(2)\quad P^{s}_{abX} = \sum_{c=0}^{N-1} \sum_{\{S\}_{abc}} \sum_{\{T\}_c} P^{m}_{S_1 X} \prod_{b \neq S_1} (1 - P_{S_1 b X})^{T_1 - 1} \prod_{j=1}^{c} \left[P_{S_j S_{j+1} X} \prod_{b \neq S_{j+1}} (1 - P_{S_{j+1} b X})^{T_{j+1} - T_j - 1} \right]$$

To determine the actual number of crossovers, eqn. 2 must first be iterated to determine the set of P_{abX} values that cause P^{s}_{abX} to equal P^{m}_{abX}. A suitable measure of the fit is the sum of squares error (SSE) defined in eqn. 3.

(3) $SSE = \sqrt{\sum_a \sum_b (P^m_{abX} - P^s_{abX})^2}$

Prior to iteration, it is useful to reformulate eqn. 2 to reduce computation time. For instance, with n_p = 3, c = 5, and N = 500, the number of possible sequences is unmanageable, ~10^{11}. Define $\langle P_{abcX} \rangle$ to be the average probability contributed by an arbitrarily chosen chimera with parent *a* at probe position *X*, parent *b* at probe position *X*+1 and *c* crossovers. By multiplying $\langle P_{abcX} \rangle$ by the number of sequences starting with *a*, ending with *b*, and containing *c* crossovers (N_{abcX}), we can estimate the probability contribution of sequences with *c* crossovers. In eqn. 4, these terms are summed over the number of crossovers to obtain an expression for P^s_{abX} that is much faster to implement on a computer than eqn. 2.

(4) $P^s_{abX} = \sum_{c=0}^{N-1} N_{abcX} \bullet \langle P_{abcX} \rangle$

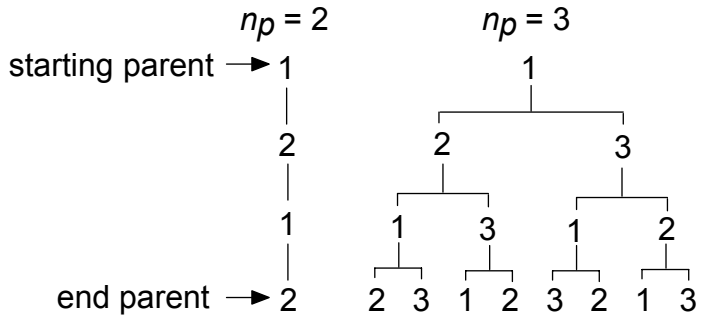
Define $N_{y,c}$ to be the number of ways to position *c* crossovers in an *N*-long sequence. $N^{ab}_{x,c}$ is the number of ways to arrange n_p parents on a sequence with *c* crossovers that starts with parent *a* and ends with parent *b*. Then $N_{c,ab}$ is calculated as follows.

(5) $N_{abcX} = N_{y,c} \bullet N^{ab}_{x,c}$

$N_{y,c}$ is calculated using eqn. 6.

(6) $N_{y,c} = \frac{N!}{c! \bullet (N-c)!}$

$N^{ab}_{x,c}$ is a function of whether $a=b$ or $a\neq b$, but does not depend on the specific identity of *a* and *b*. For example, the diagram below shows all possible parent combinations for $c = 3$.



For $c = 3$ and $n_p = 3$, $N^{11}_{x,3} = 2$ and $N^{12}_{x,3} = 3$. By symmetry, $N^{22}_{x,3} = 2$, $N^{13}_{x,3} = 3$, etc. The chart below shows representative $N^{ab}_{x,c}$ values for other values of *c* and n_p. The data in the chart are consistent with the recursion shown below.

	$n_p = 2$		$n_p = 3$		$n_p = 4$		$n_p = 5$	
c	$N^{11}_{x,c}$	$N^{12}_{x,c}$	$N^{11}_{x,c}$	$N^{12}_{x,c}$	$N^{11}_{x,c}$	$N^{12}_{x,c}$	$N^{11}_{x,c}$	$N^{12}_{x,c}$
0	1	0	1	0	1	0	1	0
1	0	1	0	1	0	1	0	1
2	1	0	2	1	3	2	4	3
3	0	1	2	3	6	7	12	13
4	1	0	6	5	21	20	52	51
5	0	1	10	11	60	61	204	205
6	1	0	22	21	183	182	820	819
7	0	1	42	43	546	547	3276	3277
8	1	0	86	85	1641	1640	13108	13107

The values of $N^{ab}_{x,c}$ for $c = 0$ are dependent only on whether $a = b$.

(7) $N^{a=b}_{x,c=0} = 1$

(8) $N^{a\neq b}_{x,c=0} = 0$

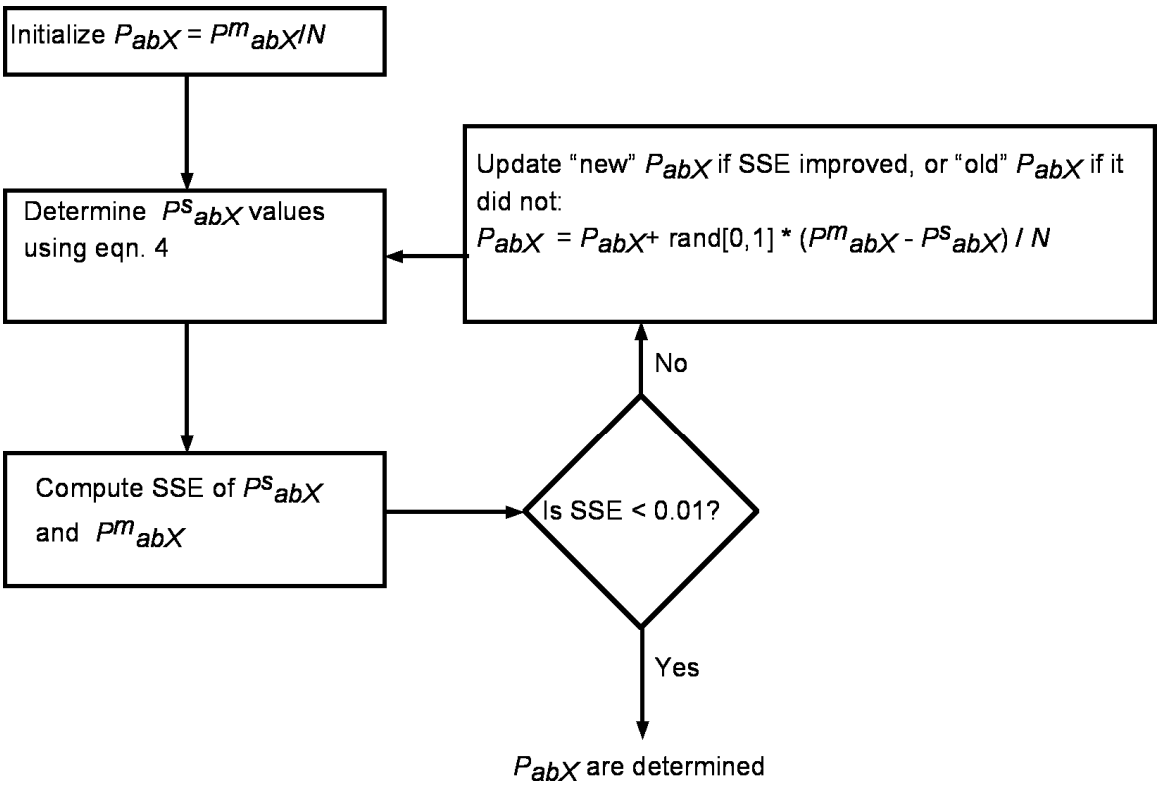
Values of $N^{ab}_{x,c}$ for $c > 0$ can be calculated using the following recursion formulas.

(9) $N^{a=b}_{x,c+1} = (n_p - 1) \bullet N^{a \neq b}_{x,c}$

(10) $N^{a \neq b}_{x,c+1} = (n_p - 2) \bullet N^{a \neq b}_{x,c} + N^{a=b}_{x,c}$

Equation 5 is used finally to calculate N_{abcX}.

Iteration according to the following scheme is used to obtain P_{abX} values for a particular interprobe region.



After iteration to obtain P_{abX} values for a particular interprobe region, the average number N_{abX} of crossovers from parent *a* to parent *b* between probes *X* and *X*+1 can be determined by summing over all possible *S* and *T* the product of the probability P^{ST} of each chimera occurring and the number of each type of crossover N^{ST}_{abX} in the interprobe region as shown in eqn. 11. The simplification using averages done for eqn. 2 can also be performed on this equation.

(11) $$N_{abX} = \sum_{c=1}^{N-1} \sum_{\{S\}_{abc}} \sum_{\{T\}_{abc}} N^{ST}_{abX} \bullet P^{ST}$$

The method described was implemented in Matlab Version 5 (Math Works, Inc., Natick, MA) and used to determine an estimate for the actual number of crossovers occuring between probe positions (N_{abX}) for the two dioxygenase libraries. For each probe pair, the run time to determine P_{abX} values with SSE < 0.01 and then determine N_{abX} was approximately 15 min using a 450 MHz processor. $\langle P_{abcX} \rangle$ was determined for $1 \le c \le 7$ by averaging over 1000 arbitrarily chosen chimeric sequences.

Nomenclature

n_p: number of parents

N: number of base pairs between two neighboring probes

L: number of base pairs in entire gene

x: Nucleotide index (1 to *N*)

X: Probe number index (1 to 6)

a: parent index (1 to n_p)

b: parent index (1 to n_p)

c: index for the number of crossovers that occur over *N*

P^m_{aX}: The "measured" probability that parent a is present at probe position *X*, calculated directly from the probe data.

P^m_{abX}: The "measured" probability that parent *a* is present at probe position *X* and parent *b* is present at probe position *X*+1, calculated directly from the probe data.

P_{abX} : The probability that nucleotide *x*+1 will come from parent *b*, given that nucleotide *x* is from parent *a*, in the region from probe position *X* to *X*+1.

P^s_{abX} : ("*s*" is simulated) The probability of simulating a chimera that has parent *a* at probe position *X* and parent *b* at probe position *X*+1, given a set of P_{abX} and P^m_{aX} values. This is the simulation equivalent of P^m_{abX}.

S: A vector that specifies the series of parents that occur for a possible chimera

$\{S\}_{abc}$: The set of all S vectors that start with *a*, end with *b*, and have length *c*+1.

T: A vector that specifies the positions of *c* crossovers for a possible chimera

$\{T\}_c$: The set of all *T* vectors that specify positions for *c* crossovers distributed over *N*-1 sequence positions.

P^{ST} : The probability of constructing a chimera with arbitrary representative vectors *S* and *T.*

$\langle P_{abcX} \rangle$: The average probability contributed by an arbitrarily chosen chimera with parent *a* at probe position *X*, parent *b* at probe position *X*+1 and *c* crossovers.

N_{abcX} : The number of possible sequences of length N that start with a, end with b, and contain c crossovers.

$N_{y,c}$: The number of ways to position c crossovers in an N-long sequence.

$N^{ab}_{x,c}$: The number of ways to arrange n_p parents on a sequence with c crossovers that starts with parent a and ends with parent b.

N_{abX} : The average number of crossovers from parent a to parent b between probes X and X+1.

N^{ST}_{abX} : The number of crossovers from a to b in the chimera with representative vectors S and T.

Matlab simulation instructions

1. Download all the ".m" files provided at http://cheme.caltech.edu/groups/fha/probes/crossovers.html into the same directory
2. Compile all probe hybridization data into one spreadsheet file with probe sites corresponding to the columns and individual clones as rows. Use "Find and replace" feature to rename genes with integers from 1 to the number of parents. Where data is missing, just insert a zero.
3. Open "data_loader.m" in your text editor of choice and paste your probe data (just numbers, not labels) into the "data = [];" line. When this file is run in matlab, clones with incomplete data will be removed, and the program will figure out how many parents your library was made from.

4. At a matlab prompt, type "nuc_prob_calc". This will call nuc_prob_calc.m, which runs the entire simulation. A bunch of stuff will print out throughout the run which is useful for troubleshooting. Depending on the speed of your computer, each gene segment can take 5-20 minutes to analyze.
5. During each cycle of the iteration, a simulated matrix of probe linkage probabilities is calculated. The sum of this matrix ("sum_Psim") is printed out and should be less than 0.995 and less than or equal to 1.000. If it is too low, you need to consider more crossovers by increasing "num_c" at the top of the "nuc_prob_calc.m" program. The program can simulate up to seven crossovers between neighboring probes. Decreasing "num_c" will make the simulation run faster but compromises accuracy.
6. For each set of neighboring probes, the simulation first does an iteration to determine a matrix of nucleotide-level probabilities. If the fit is improving, the simulation should print out steadily decreasing rms values. By the end of the iteration (nominally cycle 12), the rms ("new_rms") should be < 0.01 for each set of neighboring probes. Try increasing "num_trials" (the number of interation cycles) if the rms is too high.
7. If the iteration is not converging at all (steady rms), try lowering the "scale" variable at the top of the "nuc_prob_calc.m" code. This scales the changes made to the probabilities during each cycle. If the iteration is not converging fast enough for you, try increasing the "scale" variable.

8. After each iteration is completed, the matrices "Psim" and "Pm" are printed out. Pm is the values of probe linkage probabilities from your data, and Psim is the simulated analog. These matrices should be nearly identical.
9. After the simulation, the matrix "C" contains your results. Entering "C(4,1,2)" at a matlab prompt will return the number of crossovers from parent 1 to parent 2 between probes 4 and 5. Entering "sum(sum(sum(C)))" will return the total average number of crossovers for the library. Entering "C(:,1,2)" will return a vector containing the number of crossovers from parent 1 to parent 2 between each set of neighboring probes.

Matlab code

data_loader.m

```
% This file reads in a set of probe data and results in a set of Pm and P_start
% values.  "Pm" is a matrix: (probe number X, parent a, parent b) of probabilities
% of parent a being at probe X and parent b being at probe X+1.
% P_start (probe number X,parent a) is the probability of parent a being at probe
position X

% The data input for the code can be easily done using a spreadsheet.  Columns are
% probe positions, and rows are the individual clones.  A nonzero integer is assigned to
each
% parent.  (e.g., for a three parent library, use 1-3)  If no data is recorded for a
particular probe
% a zero should be used.  These clones will be removed from the analysis automatically.

data = [ %paste your data here with tabs between columns and line breaks between clones
];

dims = size(data);
N = dims(1);
probe_pairs = dims(2) - 1;
np = max(max(data));

% this code removes clones with at least one missing data point (zeroes)
if min(min(data)) == 0
  data = no_spaces(data,probe_pairs+1);
  dims = size(data);
  N = dims(1);
  probe_pairs= dims(2) - 1;
  np = max(max(data));
end

for m=1:np
  for n=1:np
    for j=1:probe_pairs
      Pm_load(j,m,n) = 0;
      for k=1:N
        if data(k,j) == m & data(k,j+1) == n
          Pm_load(j,m,n) = Pm_load(j,m,n) + 1/N;
        end
      end
```

```
    end
  end
end

for m=1:np
  for j=1:probe_pairs+1
    P_start_load(j,m) = 0;
    for k=1:N
      if data(k,j) == m
        P_start_load(j,m) = P_start_load(j,m) + 1/N;
      end
    end
  end
end
```

nuc_prob_calc.m

```
clear;
data_loader       % Reads in probe data and counts pairs
num_c = 5;         % Number of crossovers allowed between probes
N = 150;         % Number of base pairs to consider between probes (Do not raise above
150!)
N_ave = 1000;     % Number of random clones to consider and average over (default ->
1000)
scale = 0.01;     % This will affect the rate of convergence.  Higher values should cause
          % faster convergence, but if too high might prevent convergence (default ->
0.01)
num_trials = 12;   % number of cycles allowed for convergence

% The following code reads in probe data for the probe pair in question
% and stores the relevent numbers in Pm


for pair=1:probe_pairs
  Psim(1:np,1:np) = 0;

  % Now we load in values for Pmab for the probe pair currently running
  for j=1:np
    for k=1:np
      Pm(j,k) = Pm_load(pair,j,k);
    end
  end
  for j=1:np
    P_start(j) = P_start_load(pair,j);
  end


  % Here we initialize values for the probability of crossover/nucleotide.
  for j=1:np
    for k=1:np
      if j!=k
        P(j,k) = Pm(j,k)/N * 1.5;    % The values you put here are somewhat arbitrary
                      % but this seems to give a good first estimate for P
      end
    end
  end
  old_P = P;
  old_rms = .5;

  % This section calculates the number of variants for a given Y,Z,c combination
  % Dependent on N
  for j=1:num_c
    Ncomb(j) = combinations(N,j);
  end
  for j=1:np
    for k=1:np
      if j!=k
        Npc(j,k,1) = 0;
      end
      if j==k
        Npc(j,k,1) = 1;
      end
    end
  end

  for c=2:length(Ncomb)
    for j=1:np
      for k=1:np
        if j!=k
```

```
          Npc(j,k,c) = Npc(1,1,c-1) + (np - 2) * Npc(1,2,c-1);
        end
        if j==k
          Npc(j,k,c) = Npc(1,2,c-1) * (np - 1);
        end
      end
    end
  end
  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  cycles = 0;
  while cycles < num_trials
    cycles = cycles + 1
    rms_sum = 0;
    % The following calculates the rms between Psim and Pm; this is used to
    % decide whether to keep the new values of P or start again with the old values
    for r=1:np
      for s=1:np
        rms_sum = rms_sum + (Psim(r,s) - Pm(r,s))^2;
      end
    end
    new_rms = sqrt(rms_sum);

    if (new_rms > old_rms | min(min(P)) < 0)   % If the P values are getting further from
actual
      P = old_P;  % Revert to old P values
      for j=1:np
        for k=1:np
          if j!=k
            P(j,k) = P(j,k) + rand * (Pm(j,k) - Psim(j,k)) * scale;
          end
        end
      end
    else      % use the new values, but store them in old_P
      old_P = P;
      for j=1:np
        for k=1:np
          if j!=k
            P(j,k) = P(j,k) + rand * (Pm(j,k) - Psim(j,k)) * scale;
          end
        end
      end
      old_rms = new_rms;
    end
    rms  = old_rms  % This will print the old_rms, this should decrease or stay the same
with each
            % cycle.
    % Here we adjust any negative P values to 0
    for A=1:np
      for B=1:np
        if P(A,B) < 0
          P(A,B) = 0;
        end
      end
    end
    create_P_powers;    % This file creates a matrix of P^(1 to N) values so they don't
have
               % to be recomputed each time they are needed
    Psim(1:np,1:np) = 0;
    for Y=1:np
      for Z=1:np
            Psim(Y,Z) = 0;
            low_B = 1;
            high_B = N;

    % Zero actual crossover code
            c = 0;
            if Y==Z
              pcjl = 1;
              for m=1:np
                if m~=Y
                  pcjl = pcjl * Pp((N-1),Y,m);
                end
              end
              Psim(Y,Z) = Psim(Y,Z) + pcjl;
            end

    % One actual crossover code
            c = 1;
            j1 = Y;    % First parent
            j2 = Z;    % Second parent
            Psample = 0;
```

```
        if Y!=Z
          for t=1:N_ave
            c1 = round(rand*N + 0.5);  % Location of first crossover
            if c1 > 2
              pcjl = 1;
              for m=1:np
                if m~=j1
                  pcjl = pcjl * Pp((c1-1-1),j1,m);
                end
              end
              pcjl = pcjl * P(Y,Z);
              for m=1:np
                if m~=j2
                  pcjl = pcjl * Pp((N-c1+1),j2,m);
                end
              end
              Psample = Psample + pcjl;
            end
          end
          Psim(Y,Z) = Psim(Y,Z) + Psample * N/N_ave;
        end
% Two actual crossovers code
        if num_c>1
          c = 2;
          j1 = Y;              % First parent
          j3 = Z;              % Third parent
          Psample = 0;
          t = 0;
          while t<N_ave
            for j2=1:np          % Second parent
              if (j2~=j1 & j2~=j3)
                for cr=1:c
                  cross(cr) = round(rand*N + 0.5);
                end
                cross = sort(cross);
                c1 = cross(1);      % Location of first crossover
                c2 = cross(2);      % Location of second crossover
                if c1>2 & c2-c1 > 1
                  t=t+1;
                  pcjl = 1;
                  for m=1:np
                    if m~=j1
                      pcjl = pcjl * Pp((c1-1-1),j1,m);
                    end
                  end
                  pcjl = pcjl * P(j1,j2);
                  for m=1:np
                    if m~=j2
                      pcjl = pcjl * Pp((c2-1-c1),j2,m);
                    end
                  end
                  pcjl = pcjl * P(j2,j3);
                  for m=1:np
                    if m~=j3
                      pcjl = pcjl * Pp((N-c2+1),j3,m);
                    end
                  end
                  Psample = Psample + pcjl;
                end
              end
            end
          end
          Psim(Y,Z) = Psim(Y,Z) + Psample * Ncomb(c) * Npc(Y,Z,c)/N_ave;
        end
% Three actual crossovers code
        if num_c>2
          c = 3;
          j1 = Y;
          j4 = Z;
          Psample = 0;
          t = 0;
          while t<N_ave
            for j2=1:np
              for j3=1:np
                if (j2~=j1 & j4~=j3) & j2!=j3
                  for cr=1:c
                    cross(cr) = round(rand*N + 0.5);
                  end
                  cross = sort(cross);
                  c1 = cross(1);        % Location of first crossover
                  c2 = cross(2);        % Location of second crossover
```

```
                    c3 = cross(3);        % Location of third crossover
                    if c1>2 & c2-c1 > 1  & c3-c2 > 1
                      t=t+1;
                      pcjl = 1;
                      for m=1:np
                        if m~=j1
                          pcjl = pcjl * Pp((c1-1-1),j1,m);
                        end
                      end
                      pcjl = pcjl * P(j1,j2);
                      for m=1:np
                        if m~=j2
                          pcjl = pcjl * Pp((c2-1-c1),j2,m);
                        end
                      end
                      pcjl = pcjl * P(j2,j3);
                      for m=1:np
                        if m~=j3
                          pcjl = pcjl * Pp((c3-1-c2),j3,m);
                        end
                      end
                      pcjl = pcjl * P(j3,j4);
                      for m=1:np
                        if m~=j4
                          pcjl = pcjl * Pp((N-c3+1),j4,m);
                        end
                      end
                      Psample = Psample + pcjl;
                    end
                  end
                end
              end
            end
            Psim(Y,Z) = Psim(Y,Z) + Psample * Ncomb(c) * Npc(Y,Z,c)/N_ave;
          end
% Four actual crossovers code
          if num_c>3
            c = 4;
            j1 = Y;              % First parent
            j5 = Z;              % Fifth parent
            Psample = 0;
            t = 0;
            while t<N_ave
              for j2=1:np
                for j3=1:np
                  for j4=1:np
                    if (j2~=j1 & j4~=j5) & j2!=j3 & j3!=j4
                      for cr=1:c
                        cross(cr) = round(rand*N + 0.5);
                      end
                      cross = sort(cross);
                      c1 = cross(1);        % Location of first crossover
                      c2 = cross(2);        % Location of second crossover
                      c3 = cross(3);        % Location of third crossover
                      c4 = cross(4);        % Location of fourth crossover
                      if c1>2 & c2-c1 > 1  & c3-c2 > 1 & c4-c3 > 1
                        t=t+1;
                        pcjl = 1;
                        for m=1:np
                          if m~=j1
                            pcjl = pcjl * Pp((c1-1-1),j1,m);
                          end
                        end
                        pcjl = pcjl * P(j1,j2);
                        for m=1:np
                          if m~=j2
                            pcjl = pcjl * Pp((c2-1-c1),j2,m);
                          end
                        end
                        pcjl = pcjl * P(j2,j3);
                        for m=1:np
                          if m~=j3
                            pcjl = pcjl * Pp((c3-1-c2),j3,m);
                          end
                        end
                        pcjl = pcjl * P(j3,j4);
                        for m=1:np
                          if m~=j4
                            pcjl = pcjl * Pp((c4-1-c3),j4,m);
                          end
                        end
```

```
                    pcjl = pcjl * P(j4,j5);
                    for m=1:np
                      if m~=j5
                        pcjl = pcjl * Pp((N-c4+1),j5,m);
                      end
                    end
                    Psample = Psample + pcjl;
                  end
                end
              end
            end
          end
          Psim(Y,Z) = Psim(Y,Z) + Psample * Ncomb(c) * Npc(Y,Z,c)/N_ave;
        end
% Five actual crossovers code
        if num_c>4
          c = 5;
          j1 = Y;
          j6 = Z;
          Psample = 0;
          t = 0;
          while t<N_ave
            for j2=1:np
              for j3=1:np
                for j4=1:np
                  for j5=1:np
                    if (j2~=j1 & j6~=j5) & j2!=j3 & j3!=j4 & j4!=j5
                      for cr=1:c
                        cross(cr) = round(rand*N + 0.5);
                      end
                      cross = sort(cross);
                      c1 = cross(1);        % Location of first crossover
                      c2 = cross(2);        % Location of second crossover
                      c3 = cross(3);        % Location of third crossover
                      c4 = cross(4);        % Location of fourth crossover
                      c5 = cross(5);        % Location of fifth crossover
                      if c1>2 & c2-c1 > 1  & c3-c2 > 1 & c4-c3 > 1 & c5-c4 > 1
                        t=t+1;
                        pcjl = 1;
                        for m=1:np
                          if m~=j1
                            pcjl = pcjl * Pp((c1-1-1),j1,m);
                          end
                        end
                        pcjl = pcjl * P(j1,j2);
                        for m=1:np
                          if m~=j2
                            pcjl = pcjl * Pp((c2-1-c1),j2,m);
                          end
                        end
                        pcjl = pcjl * P(j2,j3);
                        for m=1:np
                          if m~=j3
                            pcjl = pcjl * Pp((c3-1-c2),j3,m);
                          end
                        end
                        pcjl = pcjl * P(j3,j4);
                        for m=1:np
                          if m~=j4
                            pcjl = pcjl * Pp((c4-1-c3),j4,m);
                          end
                        end
                        pcjl = pcjl * P(j4,j5);
                        for m=1:np
                          if m~=j5
                            pcjl = pcjl * Pp((c5-1-c4),j5,m);
                          end
                        end
                        pcjl = pcjl * P(j5,j6);
                        for m=1:np
                          if m~=j6
                            pcjl = pcjl * Pp((N-c5+1),j6,m);
                          end
                        end
                        Psample = Psample + pcjl;
                      end
                    end
                  end
                end
              end
            end
```

```
              end
            end
            Psim(Y,Z) = Psim(Y,Z) + Psample * Ncomb(c) * Npc(Y,Z,c)/N_ave;
          end
    % Six actual crossovers code
          if num_c>5
            c = 6;
            j1 = Y;
            j7 = Z;
            Psample = 0;
            t = 0;
            while t<N_ave
              for j2=1:np
                for j3=1:np
                  for j4=1:np
                    for j5=1:np
                      for j6=1:np
                        if (j2~=j1 & j6~=j7) & j2!=j3 & j3!=j4 & j4!=j5 & j5!=j6
                          for cr=1:c
                            cross(cr) = round(rand*N + 0.5);
                          end
                          cross = sort(cross);
                          c1 = cross(1);
                          c2 = cross(2);
                          c3 = cross(3);
                          c4 = cross(4);
                          c5 = cross(5);
                          c6 = cross(6);
                          if c1>2 & c2-c1 > 1  & c3-c2 > 1 & c4-c3 > 1 & c5-c4 > 1 &
c6-c5 > 1
                            t=t+1;
                            pcjl = 1;
                            for m=1:np
                              if m~=j1
                                pcjl = pcjl * Pp((c1-1-1),j1,m);
                              end
                            end
                            pcjl = pcjl * P(j1,j2);
                            for m=1:np
                              if m~=j2
                                pcjl = pcjl * Pp((c2-1-c1),j2,m);
                              end
                            end
                            pcjl = pcjl * P(j2,j3);
                            for m=1:np
                              if m~=j3
                                pcjl = pcjl * Pp((c3-1-c2),j3,m);
                              end
                            end
                            pcjl = pcjl * P(j3,j4);
                            for m=1:np
                              if m~=j4
                                pcjl = pcjl * Pp((c4-1-c3),j4,m);
                              end
                            end
                            pcjl = pcjl * P(j4,j5);
                            for m=1:np
                              if m~=j5
                                pcjl = pcjl * Pp((c5-1-c4),j5,m);
                              end
                            end
                            pcjl = pcjl * P(j5,j6);
                            for m=1:np
                              if m~=j6
                                pcjl = pcjl * Pp((c6-1-c5),j6,m);
                              end
                            end
                            pcjl = pcjl * P(j6,j7);
                            for m=1:np
                              if m~=j7
                                pcjl = pcjl * Pp((N-c6+1),j7,m);
                              end
                            end
                            Psample = Psample + pcjl;
                          end
                        end
                      end
                    end
                  end
                end
              end
            end
```

```
            end
            Psim(Y,Z) = Psim(Y,Z) + Psample * Ncomb(c) * Npc(Y,Z,c)/N_ave;
          end
    % Seven actual crossovers code
          if num_c>6
            c = 7;
            j1 = Y;
            j8 = Z;
            Psample = 0;
            t = 0;
            while t<N_ave
              for j2=1:np
                for j3=1:np
                  for j4=1:np
                    for j5=1:np
                      for j6=1:np
                        for j7=1:np
                          if (j2~=j1 & j6~=j7) & j2!=j3 & j3!=j4 & j4!=j5 & j5!=j6 &
j6!=j7
                            for cr=1:c
                              cross(cr) = round(rand*N + 0.5);
                            end
                            cross = sort(cross);
                            c1 = cross(1);
                            c2 = cross(2);
                            c3 = cross(3);
                            c4 = cross(4);
                            c5 = cross(5);
                            c6 = cross(6);
                            c7 = cross(7);
                            if c1>2 & c2-c1 > 1  & c3-c2 > 1 & c4-c3 > 1 & c5-c4 > 1 &
c6-c5 > 1 & c7-c6 > 1
                              t=t+1;
                              pcjl = 1;
                              for m=1:np
                                if m~=j1
                                  pcjl = pcjl * Pp((c1-1-1),j1,m);
                                end
                              end
                              pcjl = pcjl * P(j1,j2);
                              for m=1:np
                                if m~=j2
                                  pcjl = pcjl * Pp((c2-1-c1),j2,m);
                                end
                              end
                              pcjl = pcjl * P(j2,j3);
                              for m=1:np
                                if m~=j3
                                  pcjl = pcjl * Pp((c3-1-c2),j3,m);
                                end
                              end
                              pcjl = pcjl * P(j3,j4);
                              for m=1:np
                                if m~=j4
                                  pcjl = pcjl * Pp((c4-1-c3),j4,m);
                                end
                              end
                              pcjl = pcjl * P(j4,j5);
                              for m=1:np
                                if m~=j5
                                  pcjl = pcjl * Pp((c5-1-c4),j5,m);
                                end
                              end
                              pcjl = pcjl * P(j5,j6);
                              for m=1:np
                                if m~=j6
                                  pcjl = pcjl * Pp((c6-1-c5),j6,m);
                                end
                              end
                              pcjl = pcjl * P(j6,j7);
                              for m=1:np
                                if m~=j7
                                  pcjl = pcjl * Pp((c7-1-c6),j7,m);
                                end
                              end
                              pcjl = pcjl * P(j7,j8);
                              for m=1:np
                                if m~=j8
                                  pcjl = pcjl * Pp((N-c7+1),j8,m);
                                end
                              end
```

```
                                Psample = Psample + pcjl;
                              end
                            end
                          end
                        end
                      end
                    end
                  end
                end
                Psim(Y,Z) = Psim(Y,Z) + Psample * Ncomb(c) * Npc(Y,Z,c)/N_ave;
              end
      end
    end
    % Psim values are adjusted to reflect start probabilities according to the probe data
    for r=1:np
      for s=1:np
        Psim(r,s) = Psim(r,s) * P_start(r);
      end
    end
    sum_Psim = sum(sum(Psim))
  end  % end of trials for convergence
  Psim = Psim
  Pm = Pm
  sum(sum(Psim))
  P = old_P
  calc_Nc2
end
```

fact.m

```
function product = fact(num)

product = 1;
while num>0
  product  = product * num;
  num = num - 1;
end
```

combinations.m

```
function answer = combinations(n,t)

% This calculates the number of ways to pick t things from a pool of n things
answer = fact(n)/(fact(t)*fact(n-t));
```

calc_Nc2.m

```
% This code converts the probability of crossover at each base into the number of
% actual crossovers.

num_c = 5;  % Number of crossovers allowed between probes
N = 150;           % Number of base pairs to consider between probes
N_ave = 1000;  % Number of randomly chosen clones with c crossovers to average over

% This section calculates the number of variants for a given Y,Z,c combination
% Dependent on N
for j=1:num_c
  Ncomb(j) = combinations(N,j);
end
for j=1:np
  for k=1:np
    if j!=k
      Npc(j,k,1) = 0;
    end
    if j==k
      Npc(j,k,1) = 1;
    end
  end
end

for c=2:length(Ncomb)
  for j=1:np
    for k=1:np
```

```
      if j!=k
        Npc(j,k,c) = Npc(1,1,c-1) + (np - 2) * Npc(1,2,c-1);
      end
      if j==k
        Npc(j,k,c) = Npc(1,2,c-1) * (np - 1);
      end
    end
  end
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


create_P_powers;      % This file creates a matrix of P^(1 to N) values so they don't
have
         % to be recomputed each time they are needed
N_cr(1:np,1:np) = 0;   % This is the number of crossovers occurring in the segment of
interest between parents a and b
for Y=1:np        % Y and Z represent the parent at probe X and X+1, respectively
  for Z=1:np
% One actual crossover code
        c = 1;
        j1 = Y;    % First parent
        j2 = Z;    % Second parent
        Nsample(1:np,1:np) = 0;
        if Y!=Z
          for t=1:N_ave
            c1 = round(rand*N + 0.5);  % Location of first crossover
            if c1 > 2
              pcjl = 1;
              for m=1:np
                if m~=j1
                  pcjl = pcjl * Pp((c1-1-1),j1,m);
                end
              end
              pcjl = pcjl * P(Y,Z);
              for m=1:np
                if m~=j2
                  pcjl = pcjl * Pp((N-c1+1),j2,m);
                end
              end
              Nsample(j1,j2) = Nsample(j1,j2) + pcjl;
            end
          end
          N_cr = N_cr + Nsample * N * P_start(Y)/N_ave;
        end
% Two actual crossovers code
        if num_c>1
          c = 2;
          j1 = Y;              % First parent
          j3 = Z;              % Third parent
          t = 0;
          Nsample(1:np,1:np) = 0;
          while t<N_ave
            for j2=1:np          % Second parent
              if (j2~=j1 & j2~=j3)
                for cr=1:c
                  cross(cr) = round(rand*N + 0.5);
                end
                cross = sort(cross);
                c1 = cross(1);      % Location of first crossover
                c2 = cross(2);      % Location of second crossover
                if c1>2 & c2-c1 > 1
                  t=t+1;
                  pcjl = 1;
                  for m=1:np
                    if m~=j1
                      pcjl = pcjl * Pp((c1-1-1),j1,m);
                    end
                  end
                  pcjl = pcjl * P(j1,j2);
                  for m=1:np
                    if m~=j2
                      pcjl = pcjl * Pp((c2-1-c1),j2,m);
                    end
                  end
                  pcjl = pcjl * P(j2,j3);
                  for m=1:np
                    if m~=j3
                      pcjl = pcjl * Pp((N-c2+1),j3,m);
                    end
                  end
```

```
                Nsample(j1,j2) = Nsample(j1,j2) + pcjl;
                Nsample(j2,j3) = Nsample(j2,j3) + pcjl;
              end
            end
          end
        end
        N_cr = N_cr + Nsample * Ncomb(c) * Npc(Y,Z,c) * P_start(Y)/N_ave;
      end
% Three actual crossovers code
      if num_c>2
        c = 3;
        j1 = Y;              % First parent
        j4 = Z;              % Fourth parent
        Nsample(1:np,1:np) = 0;
        t = 0;
        while t<N_ave
          for j2=1:np
            for j3=1:np
              if (j2~=j1 & j4~=j3) & j2!=j3
                for cr=1:c
                  cross(cr) = round(rand*N + 0.5);
                end
                cross = sort(cross);
                c1 = cross(1);        % Location of first crossover
                c2 = cross(2);        % Location of second crossover
                c3 = cross(3);        % Location of third crossover
                if c1>2 & c2-c1 > 1  & c3-c2 > 1
                  t=t+1;
                  pcjl = 1;
                  for m=1:np
                    if m~=j1
                      pcjl = pcjl * Pp((c1-1-1),j1,m);
                    end
                  end
                  pcjl = pcjl * P(j1,j2);
                  for m=1:np
                    if m~=j2
                      pcjl = pcjl * Pp((c2-1-c1),j2,m);
                    end
                  end
                  pcjl = pcjl * P(j2,j3);
                  for m=1:np
                    if m~=j3
                      pcjl = pcjl * Pp((c3-1-c2),j3,m);
                    end
                  end
                  pcjl = pcjl * P(j3,j4);
                  for m=1:np
                    if m~=j4
                      pcjl = pcjl * Pp((N-c3+1),j4,m);
                    end
                  end
                  Nsample(j1,j2) = Nsample(j1,j2) + pcjl;
                  Nsample(j2,j3) = Nsample(j2,j3) + pcjl;
                  Nsample(j3,j4) = Nsample(j3,j4) + pcjl;
                end
              end
            end
          end
          N_cr = N_cr + Nsample * Ncomb(c) * Npc(Y,Z,c) * P_start(Y)/N_ave;
        end
% Four actual crossovers code
        if num_c>3
          c = 4;
          j1 = Y;              % First parent
          j5 = Z;              % Fifth parent
          Nsample(1:np,1:np) = 0;
          t = 0;
          while t<N_ave
            for j2=1:np
              for j3=1:np
                for j4=1:np
                  if (j2~=j1 & j4~=j5) & j2!=j3 & j3!=j4
                    for cr=1:c
                      cross(cr) = round(rand*N + 0.5);
                    end
                    cross = sort(cross);
                    c1 = cross(1);        % Location of first crossover
                    c2 = cross(2);        % Location of second crossover
                    c3 = cross(3);        % Location of third crossover
```

```
                    c4 = cross(4);        % Location of fourth crossover
                    if c1>2 & c2-c1 > 1  & c3-c2 > 1 & c4-c3 > 1
                      t=t+1;
                      pcjl = 1;
                      for m=1:np
                        if m~=j1
                          pcjl = pcjl * Pp((c1-1-1),j1,m);
                        end
                      end
                      pcjl = pcjl * P(j1,j2);
                      for m=1:np
                        if m~=j2
                          pcjl = pcjl * Pp((c2-1-c1),j2,m);
                        end
                      end
                      pcjl = pcjl * P(j2,j3);
                      for m=1:np
                        if m~=j3
                          pcjl = pcjl * Pp((c3-1-c2),j3,m);
                        end
                      end
                      pcjl = pcjl * P(j3,j4);
                      for m=1:np
                        if m~=j4
                          pcjl = pcjl * Pp((c4-1-c3),j4,m);
                        end
                      end
                      pcjl = pcjl * P(j4,j5);
                      for m=1:np
                        if m~=j5
                          pcjl = pcjl * Pp((N-c4+1),j5,m);
                        end
                      end
                      Nsample(j1,j2) = Nsample(j1,j2) + pcjl;
                      Nsample(j2,j3) = Nsample(j2,j3) + pcjl;
                      Nsample(j3,j4) = Nsample(j3,j4) + pcjl;
                      Nsample(j4,j5) = Nsample(j4,j5) + pcjl;
                    end
                  end
                end
              end
            end
          end
          N_cr = N_cr + Nsample * Ncomb(c) * Npc(Y,Z,c) * P_start(Y)/N_ave;
        end
% Five actual crossovers code
        if num_c>4
          c = 5;
          j1 = Y;
          j6 = Z;
          Nsample(1:np,1:np) = 0;
          t = 0;
          while t<N_ave
            for j2=1:np
              for j3=1:np
                for j4=1:np
                  for j5=1:np
                    if (j2~=j1 & j6~=j5) & j2!=j3 & j3!=j4 & j4!=j5
                      for cr=1:c
                        cross(cr) = round(rand*N + 0.5);
                      end
                      cross = sort(cross);
                      c1 = cross(1);        % Location of first crossover
                      c2 = cross(2);        % Location of second crossover
                      c3 = cross(3);        % Location of third crossover
                      c4 = cross(4);        % Location of fourth crossover
                      c5 = cross(5);        % Location of fifth crossover
                      if c1>2 & c2-c1 > 1  & c3-c2 > 1 & c4-c3 > 1 & c5-c4 > 1
                        t=t+1;
                        pcjl = 1;
                        for m=1:np
                          if m~=j1
                            pcjl = pcjl * Pp((c1-1-1),j1,m);
                          end
                        end
                        pcjl = pcjl * P(j1,j2);
                        for m=1:np
                          if m~=j2
                            pcjl = pcjl * Pp((c2-1-c1),j2,m);
                          end
                        end
```

```
                        pcj1 = pcj1 * P(j2,j3);
                        for m=1:np
                          if m~=j3
                            pcj1 = pcj1 * Pp((c3-1-c2),j3,m);
                          end
                        end
                        pcj1 = pcj1 * P(j3,j4);
                        for m=1:np
                          if m~=j4
                            pcj1 = pcj1 * Pp((c4-1-c3),j4,m);
                          end
                        end
                        pcj1 = pcj1 * P(j4,j5);
                        for m=1:np
                          if m~=j5
                            pcj1 = pcj1 * Pp((c5-1-c4),j5,m);
                          end
                        end
                        pcj1 = pcj1 * P(j5,j6);
                        for m=1:np
                          if m~=j6
                            pcj1 = pcj1 * Pp((N-c5+1),j6,m);
                          end
                        end
                        Nsample(j1,j2) = Nsample(j1,j2) + pcj1;
                        Nsample(j2,j3) = Nsample(j2,j3) + pcj1;
                        Nsample(j3,j4) = Nsample(j3,j4) + pcj1;
                        Nsample(j4,j5) = Nsample(j4,j5) + pcj1;
                        Nsample(j5,j6) = Nsample(j5,j6) + pcj1;
                      end
                    end
                  end
                end
              end
            end
          end
          N_cr = N_cr + Nsample * Ncomb(c) * Npc(Y,Z,c) * P_start(Y)/N_ave;
        end
% Six actual crossovers code
        if num_c>5
          c = 6;
          j1 = Y;
          j7 = Z;
          Nsample(1:np,1:np) = 0;
          t = 0;
          while t<N_ave
            for j2=1:np
              for j3=1:np
                for j4=1:np
                  for j5=1:np
                    for j6=1:np
                      if (j2~=j1 & j6~=j7) & j2!=j3 & j3!=j4 & j4!=j5 & j5!=j6
                        for cr=1:c
                          cross(cr) = round(rand*N + 0.5);
                        end
                        cross = sort(cross);
                        c1 = cross(1);
                        c2 = cross(2);
                        c3 = cross(3);
                        c4 = cross(4);
                        c5 = cross(5);
                        c6 = cross(6);
                        if c1>2 & c2-c1 > 1  & c3-c2 > 1 & c4-c3 > 1 & c5-c4 > 1 & c6-c5
> 1
                          t=t+1;
                          pcj1 = 1;
                          for m=1:np
                            if m~=j1
                              pcj1 = pcj1 * Pp((c1-1-1),j1,m);
                            end
                          end
                          pcj1 = pcj1 * P(j1,j2);
                          for m=1:np
                            if m~=j2
                              pcj1 = pcj1 * Pp((c2-1-c1),j2,m);
                            end
                          end
                          pcj1 = pcj1 * P(j2,j3);
                          for m=1:np
                            if m~=j3
                              pcj1 = pcj1 * Pp((c3-1-c2),j3,m);
```

```
                        end
                      end
                      pcjl = pcjl * P(j3,j4);
                      for m=1:np
                        if m~=j4
                          pcjl = pcjl * Pp((c4-1-c3),j4,m);
                        end
                      end
                      pcjl = pcjl * P(j4,j5);
                      for m=1:np
                        if m~=j5
                          pcjl = pcjl * Pp((c5-1-c4),j5,m);
                        end
                      end
                      pcjl = pcjl * P(j5,j6);
                      for m=1:np
                        if m~=j6
                          pcjl = pcjl * Pp((c6-1-c5),j6,m);
                        end
                      end
                      pcjl = pcjl * P(j6,j7);
                      for m=1:np
                        if m~=j7
                          pcjl = pcjl * Pp((N-c6+1),j7,m);
                        end
                      end
                      Nsample(j1,j2) = Nsample(j1,j2) + pcjl;
                      Nsample(j2,j3) = Nsample(j2,j3) + pcjl;
                      Nsample(j3,j4) = Nsample(j3,j4) + pcjl;
                      Nsample(j4,j5) = Nsample(j4,j5) + pcjl;
                      Nsample(j5,j6) = Nsample(j5,j6) + pcjl;
                      Nsample(j6,j7) = Nsample(j6,j7) + pcjl;
                    end
                  end
                end
              end
            end
          end
        end
        N_cr = N_cr + Nsample * Ncomb(c) * Npc(Y,Z,c) * P_start(Y)/N_ave;
      end
% Seven actual crossovers code
      if num_c>6
        c = 7;
        j1 = Y;
        j8 = Z;
        Nsample(1:np,1:np) = 0;
        t = 0;
        while t<N_ave
          for j2=1:np
            for j3=1:np
              for j4=1:np
                for j5=1:np
                  for j6=1:np
                    for j7=1:np
                      if (j2~=j1 & j6~=j7) & j2!=j3 & j3!=j4 & j4!=j5 & j5!=j6 & j6!=j7
                        for cr=1:c
                          cross(cr) = round(rand*N + 0.5);
                        end
                        cross = sort(cross);
                        c1 = cross(1);
                        c2 = cross(2);
                        c3 = cross(3);
                        c4 = cross(4);
                        c5 = cross(5);
                        c6 = cross(6);
                        c7 = cross(7);
                        if c1>2 & c2-c1 > 1  & c3-c2 > 1 & c4-c3 > 1 & c5-c4 > 1 & c6-
c5 > 1 & c7-c6 > 1
                          t=t+1;
                          pcjl = 1;
                          for m=1:np
                            if m~=j1
                              pcjl = pcjl * Pp((c1-1-1),j1,m);
                            end
                          end
                          pcjl = pcjl * P(j1,j2);
                          for m=1:np
                            if m~=j2
                              pcjl = pcjl * Pp((c2-1-c1),j2,m);
```

```
                                end
                              end
                              pcjl = pcjl * P(j2,j3);
                              for m=1:np
                                if m~=j3
                                  pcjl = pcjl * Pp((c3-1-c2),j3,m);
                                end
                              end
                              pcjl = pcjl * P(j3,j4);
                              for m=1:np
                                if m~=j4
                                  pcjl = pcjl * Pp((c4-1-c3),j4,m);
                                end
                              end
                              pcjl = pcjl * P(j4,j5);
                              for m=1:np
                                if m~=j5
                                  pcjl = pcjl * Pp((c5-1-c4),j5,m);
                                end
                              end
                              pcjl = pcjl * P(j5,j6);
                              for m=1:np
                                if m~=j6
                                  pcjl = pcjl * Pp((c6-1-c5),j6,m);
                                end
                              end
                              pcjl = pcjl * P(j6,j7);
                              for m=1:np
                                if m~=j7
                                  pcjl = pcjl * Pp((c7-1-c6),j7,m);
                                end
                              end
                              pcjl = pcjl * P(j7,j8);
                              for m=1:np
                                if m~=j8
                                  pcjl = pcjl * Pp((N-c7+1),j8,m);
                                end
                              end
                              Nsample(j1,j2) = Nsample(j1,j2) + pcjl;
                              Nsample(j2,j3) = Nsample(j2,j3) + pcjl;
                              Nsample(j3,j4) = Nsample(j3,j4) + pcjl;
                              Nsample(j4,j5) = Nsample(j4,j5) + pcjl;
                              Nsample(j5,j6) = Nsample(j5,j6) + pcjl;
                              Nsample(j6,j7) = Nsample(j6,j7) + pcjl;
                              Nsample(j7,j8) = Nsample(j7,j8) + pcjl;
                            end
                          end
                        end
                      end
                    end
                  end
                end
              end
            end
            N_cr = N_cr + Nsample * Ncomb(c) * Npc(Y,Z,c) * P_start(Y)/N_ave;
          end
  end
end
% This saves the data for the current probe pair in a matrix called C
for j=1:np
  for k=1:np
    C(pair,j,k) = N_cr(j,k);
  end
end
```

no_spaces.m

```
function array = no_spaces(data,prob_col)

dims = size(data);
len = dims(1);
wid = dims(2);
m = 1;

for j=1:len
  mark = 0;
  for k=1:prob_col
    if data(j,k) == 0
      mark = 1;
```

```
    end
  end
  if mark == 0
    array(m,:) = data(j,:);
    m=m+1;
  end
end
```

create_p_powers.m

```
for j=1:N
  for m=1:np
    for n=1:np
      Pp(j,m,n) = (1-P(m,n))^j;
    end
  end
end
```