doi:10.1006/jmbi.2001.5349 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 316, 643–656

JMB

Analysis of Shuffled Gene Libraries

John M. Joern, Peter Meinhold and Frances H. Arnold*

Division of Chemistry and Chemical Engineering 210-41 California Institute of Technology, Pasadena CA 91125, USA

**Corresponding author*

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 to identify key sequence-function relationships.

Keywords: directed evolution; recombination; family shuffling; StEP; random chimeragenesis

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 acid residues (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 those 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 need to

Abbreviations used: BPO, biphenyl dioxygenase; TCDO, tetrachlorobenzene dioxygenase; TDO, toluene dioxygenase.

E-mail address of the corresponding author: frances@cheme.caltech.edu

0022-2836/02/030643–14 $35.00/0

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 co-workers described a multiple macroarray system based on annealing of radioactive oligonucleotide probes to preselected gene positions that 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(±0.8) crossovers on average and a range of 0 to 7, while active clones contained 3.8(±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(±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 mutations operations.

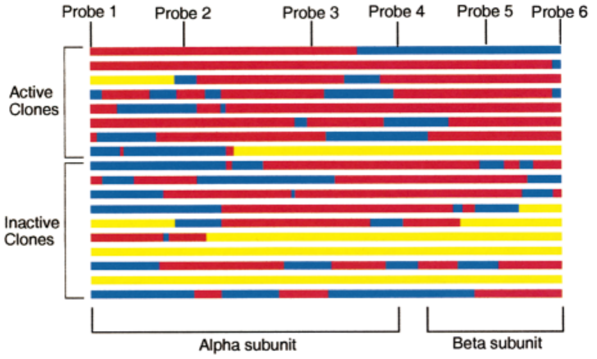


Figure 1. Sequencing results of 18 clones from the library made by shuffling genes encoding the α and β subunits of three dioxygenases. Horizontal colored bars represent the sequences of individual clones from the toluene-active or toluene-inactive subset of the library. Sequence elements from *todC1C2*, *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 a high level of 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 2(a) for an example of how these regions are defined). Figure 2(b) shows that while small regions of contiguous identity <6 bp 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.

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 spaced roughly equally 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

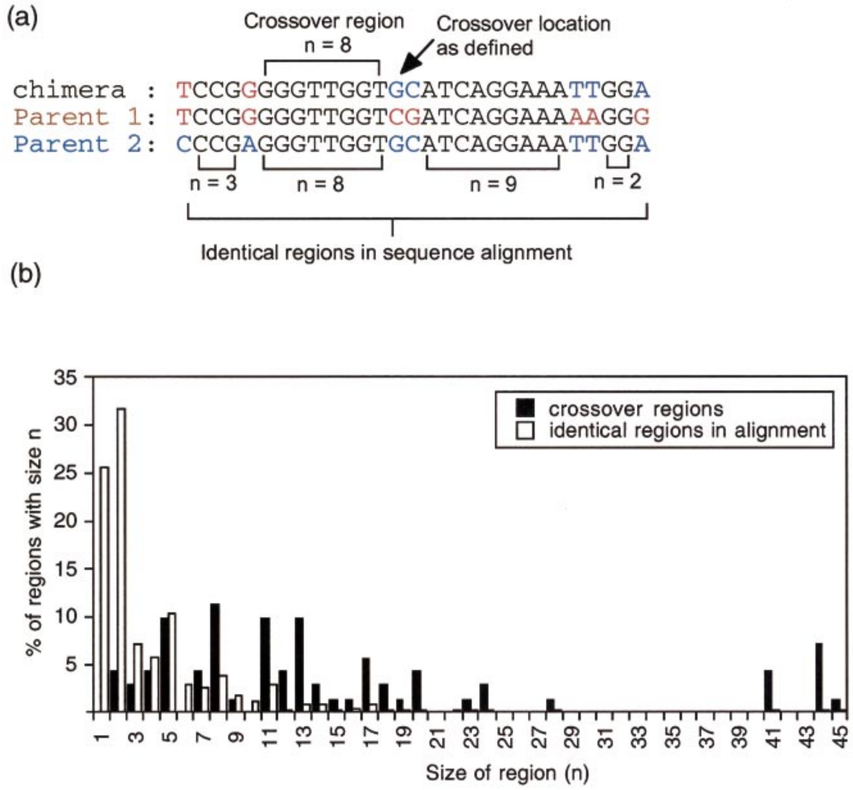


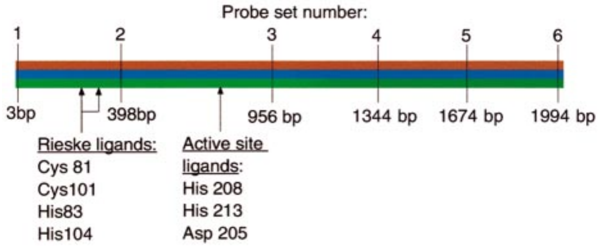
Figure 2. The size distribution of regions where crossovers occurred in 18 sequenced genes of the three-parent library, compared to the size distribution of identical regions in the sequence alignments. (a) Sample section of a sequence alignment containing a crossover. A crossover has occurred between the second and third alleles shown, in a region of 8 bp. Because the exact crossover location cannot be determined even by sequencing, it is defined as the first non-identical 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.

equal. An antibody-alkaline phosphatase complex is used to detect bound label by display of chemiluminescence after free probe is washed away.

For each shuffled dioxygenase library, two 384-well plates containing chimeric clones and the parents (six 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(±0.07) crossovers/gene for the unselected three-parent library and 2.11(±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_{abX}^{m} 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 the Supplementary Material for explanation and calculations).



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 indicated, based on the *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.

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(±0.25) crossovers/gene agrees with the sequencing results (4.20(±0.79) for unselected clones) and is considerably higher than the 1.77(±0.07) observed crossovers. For the two-parent library, the estimated number of crossovers is 5.04(±0.18), compared to only 2.11(±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

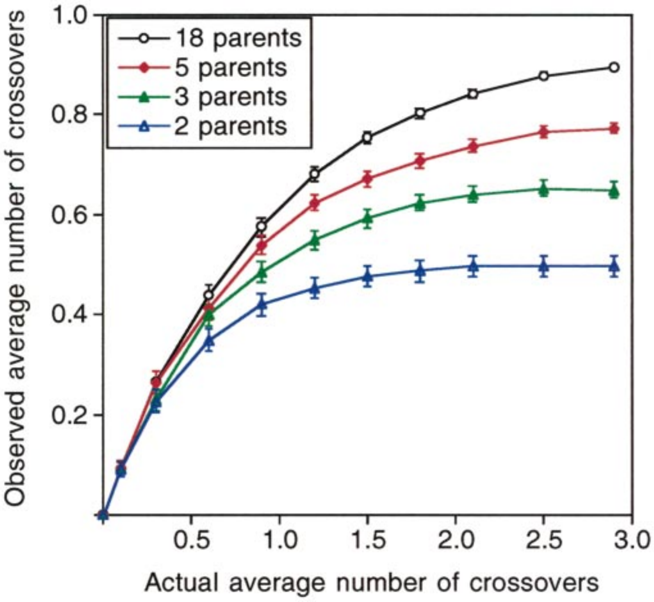


Figure 4. The observed average number of crossovers (the number directly apparent from probe hybridization data) between two probes plotted against the actual average number of crossovers for two probe sites separated by an arbitrary sequence length for different numbers of parents. For the purposes 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.

give rise to larger errors in the actual number of crossovers.

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 percentage 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 (five probes total) are required (4 = maximum number of 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 a similar percentage identity, as in this hypothetical case, additional probe positions should be used.

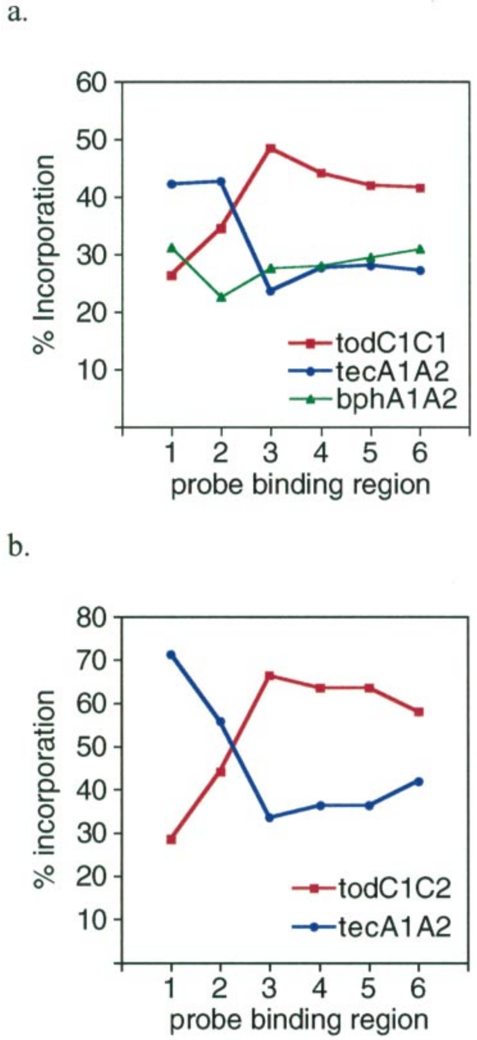


Figure 5. Incorporation of parent sequences at different probe positions in the (a) three-parent dioxygenase library (306 clones) and (b) two-parent library (317 clones) is biased towards *tecA1A2* at the 5′-end of the gene (probe positions 1 and 2) and towards *todC1C2* at the 3′ half (positions 3 to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

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 DNase I-fragmented parental DNA fragments in the shuffling reassembly reaction. Interestingly, however, the parental incorporation also varied from region to region (Figure 5(a)). 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 5(b)), 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.

Unequal amplification[18] or cloning efficiency as well as sequence-dependent variations in DNase I 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 G + C-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 G + C-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. *Escherichia coli* BL21(DE3) was 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 containing only ampicillin. No pJMJ6 transformants were found on the IPTG plates, while the pJMJ2 transformation yielded approximately 1000 transformants. The same transformation mixtures plated without IPTG yielded approximately 1000 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 tetrachlorobenzene dioxygenase (TCDO) could bias parental incorporation, and this could be position-depen-

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

	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

dent. 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 DNase I digestion could 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 a relatively low level of 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*.

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 distributed approximately evenly over all regions probed.

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}) (equation (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 described by 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

† P_{abx} is a function of sequence position x, whereas the P_{abX} variable discussed previously is constant over the region from probe position X to $X + 1$. When P_{abx} is averaged over x, the value is similar to P_{abX}.

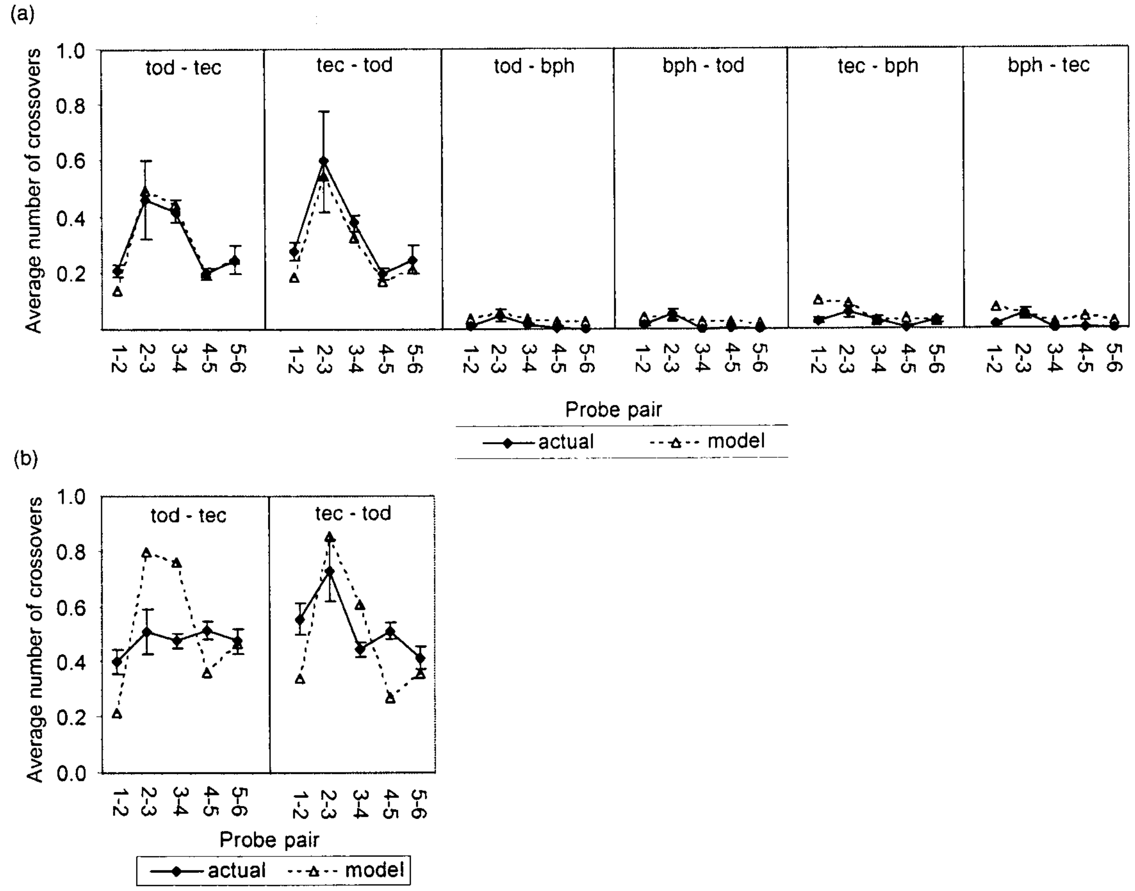


Figure 6. The number of crossovers N_{abX} after correction of the probe data (actual) is compared to the N_{abX} predicted by the model using equation (1) (model) for all types of crossovers. (a) Three-parent dioxygenase library. The average number of crossovers from *todC1C2* to *tecA1A2* within the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

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, $\beta = 1$; when they are not, β must be increased above 1. The number of parents (n_p) is also included:

$$P_{abx} = \frac{(\Delta G_{abx})^{\alpha}}{\sum\limits_{a\neq b}\sum\limits_{b\neq a}\sum\limits_{k=1}^{L}(\Delta G_{abk})^{\alpha}}\, n_p n_c \beta \qquad (1)$$

Crossover is not allowed ($\Delta G_{abx} = 0$) at positions where the region of overlap is 1-2 bp or the free energy change upon duplex formation is positive. To validate and 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 4000 chimeras *in silico* as described above. For n_c, we used values of 3.65 and 5.04 for the three-parent 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 equation (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 6(a)). For the two-parent library, however, the correlation is much weaker (Figure 6(b)). We do not know the reason for this.

Figure 7 compares the actual crossover points determined from the 18 sequenced clones (Figure 7(a)) to the relative probabilities of crossover according to the model (Figure 7(b)). 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 G + C 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]

Recombination and protein function

Of the recombined dioxygenases, only toluene dioxygenase (TDO) shows high activity on toluene; TCDO’s activity is approximately 10 % that of TDO, and biphenyl dioxygenase (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.

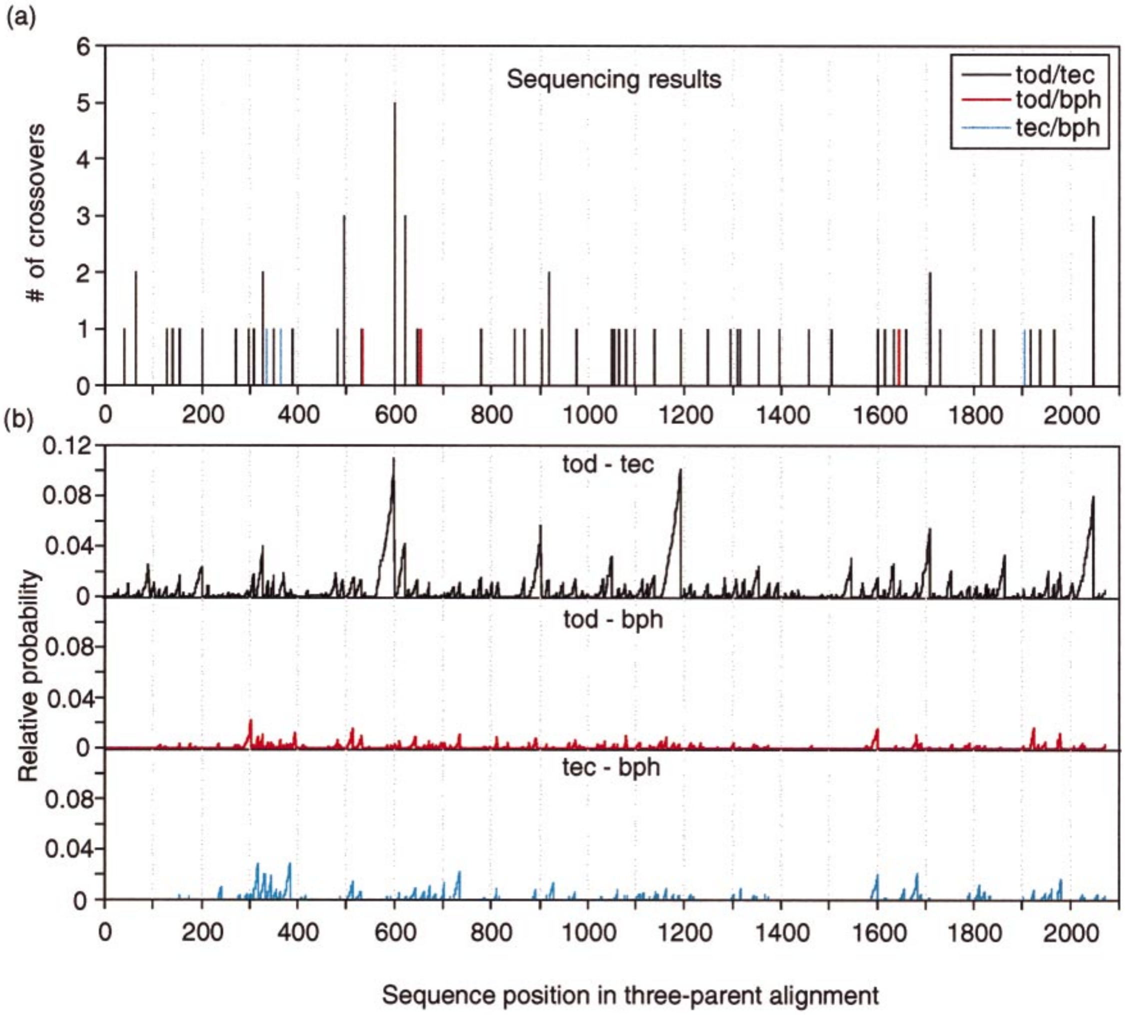


Figure 7. Model evaluation for predicting preferred crossover points. (a) Number of crossovers observed at each sequence position by DNA sequencing of genes from 18 clones. (b) Relative probabilities of crossover calculated according to equation (1), with $a = 1.6$, plotted against sequence position.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

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

	Relative activity toward	
Parent	Indole	Toluene
TDO	*****	*****
TCDO	*****	*
BPO	Not active	Not active

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.

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-parent and the 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.

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.

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 three-dimensional structure by counting the number of interactions jeopardized by a particular arrangement of crossovers (unpublished results). 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 seven of 266 active clones

Table 3. Comparison of the average number of crossovers for unselected clones and clones selected for activity toward indole

	Three-parent library		Two-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 corssovers	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

The measured average number of crossovers is determined directly from the probe hybridization data and corrected for multiple crossovers between probes as described (see the 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.

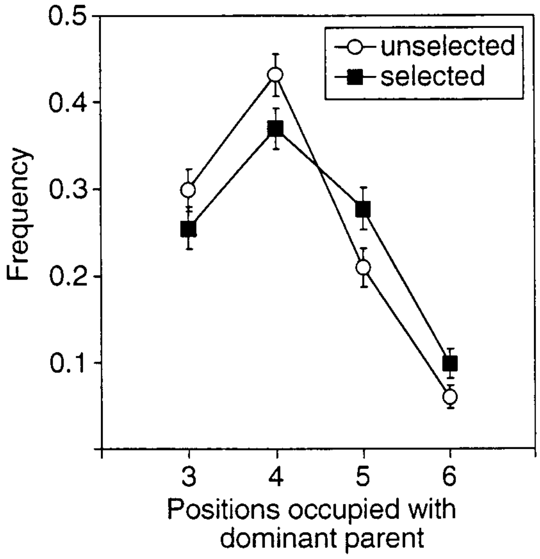


Figure 8. Distribution of the number of probe positions occupied by the dominant parent (the parent present at the most probe positions out of six total) for the two-parent unselected and selected libraries. Unselected and selected libraries comprise 317 and 318 clones, respectively. Error bars represent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

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.

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,[26] 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.[14,15,27] In these studies, chi-

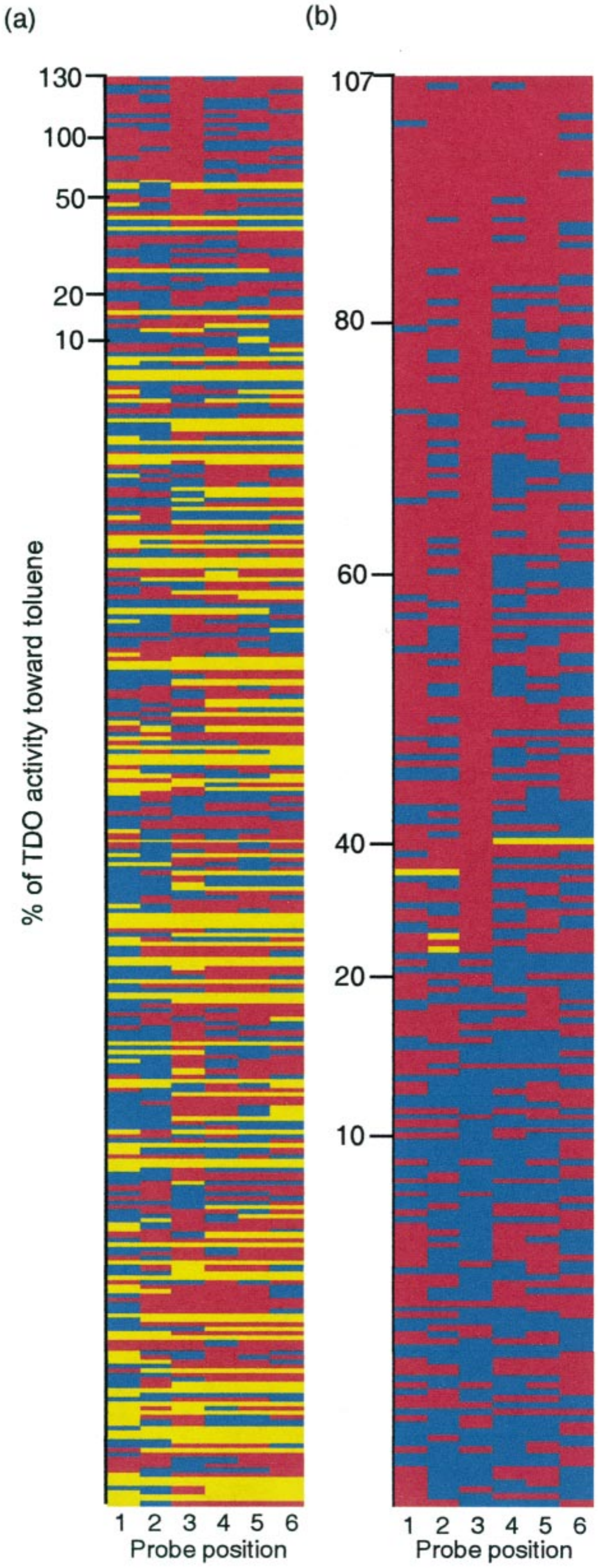


Figure 9. Probe hybridization patterns sorted by relative activity towards toluene and plotted as a heat map using the program Spotfire® (Spotfire, Inc., Cambridge, MA), Wild-type *todC1C2* (toluene dioxygenase) has an activity of 100 %. *todC1C2* is colored red, *tecA1A2* blue and *bphA1A2* yellow. (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.

meric 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 non-natural substrates.[28] At this point, we have not 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, 5000 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 >7 bp) taken n_c (the average number of crossovers) at a time.

If the parent pool contains parents with a low level of 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 wild-type genes, which further reduces diversity. One useful strategy for avoiding reassembly of wild-type 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 sequenced completely. 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 (see the Supplementary Material for descriptions of the plasmids used) was constructed by inserting the todBAD gene fragment from pDTG602[29] between the *Bam*HI and *Xba*I sites of ptrc99A (Amersham Pharmacia Biotech, Piscataway, NJ). Plasmids pJMJ2, pJMJ6, and pJMJ7 were constructed by cloning *todC1C2* from pDTG602,[29] *tecA1A2* from pSTE7,[30] and *bphA1A2* from

LB400,[32] respectively, into the *Kpn*I/*Bam*HI 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 100 μl reaction mixture contained: 10 μl of 10 × Pfu buffer, 2 μl of PCR nucleotide mix (10 mM each), 40 pmol of each primer, five units of Pfu polymerase, 3 μl of dimethylsulfoxide (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 three minutes, followed by 20 cycles of (94 °C for 30 seconds; 52 °C for 30 seconds; 72 °C for five minutes), 72 °C for ten minutes, 4 °C thereafter.

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

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

To amplify full-length (2.1kb) genes, this reassembly reaction was diluted 500-fold 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 medium[32] containing 100 mg/l of ampicillin and 0.4 % (w/v) glucose. Independent colonies were picked randomly using a QpixII colony picker (Genetix, New Milton, UK) and inoculated into the filled wells. The remaining ten wells were then filled with 70 μl of M9-minimal medium. Four wells were left uninoculated and six 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,[32] indole crystals (Sigma, St. Louis, MO) were spread out onto the lid of the plate. The plate was incubated at 30 °C for three 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 that 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 cm × 11.5 cm membranes (Amersham, Piscataway, NJ) placed on M9-minimal medium[32] plates containing 1.5 % (w/v) 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 (three parents × six 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).

Labeled 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 two to three hours at 61 °C in a model 400 Hybridization oven (Robbins Scientific, Sunnyvale, CA). Stringency washes were carried out twice in SSC (15 mM trisodium citrate, 150 mM NaCl, pH 7) for 15 minutes 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 50 mm f/1.4D AF lens (Nikon, Denver, CO). The 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) and 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 of Joern *et al.*[33] 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[32] plates containing 100 mg/l of ampicillin and 0.4 % (w/v) D-glucose. Colonies grew in this gridded format for 14 hours at 30 °C and were transferred to M9 medium[32] containing 4 % (w/v) Bacto-agar, 0.5 mM IPTG, 100 mg/l of ampicillin, 1.6 % (w/v) D-glucose, and 80 mg/l of $FeSO_4 \cdot 7H_2O$ on a 132 mm diameter nitrocellulose membrane (Protran, 0.45 μm, Schleicher & Schuell). The colonies were incubated for 12 minutes 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 % (w/v) Gibbs reagent (added as a 2 % solution in ethanol). After eight to nine minutes, 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 AF20-40 mm lens (Tamron Co., Ltd., Tokyo, Japan) and a 590(±20) nm bandpass filter (Omega Optical, Brattleboro, VT).

The image analysis tool Quantity One (Biorad, Hercules, CA) was used to quantify the relative activities of the clones. A 24 × 16 array with a 2.5 mm × 2.5 mm 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 TDO 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 TDO and the baseline peak intensity for wild-type TDO. The baseline peak intensity varied slightly across the image, and was estimated for each clone by using the minimum peak intensity of the eight 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.

Acknowledgments

The authors thank Jonathan Silberg, Christopher Voigt f and Lillian Pierce for their comments on the manuscript. This work was supported by the U. S. National Science Foundation (BES#9981770) and by Maxygen, Inc.

References

1. Gilfillan, S. (1935). *Inventing the Ship*, Follett Publishing Co., Chicago, IL.
2. Nelson, R. S. & Winter, S. (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. *Nature Biotechnol.* **17**, 259-264.
5. Schmidt-Dannert, C., Umeno, D. & Arnold, F. H. (2000). Molecular breeding of carotenoid biosynthetic pathways. *Nature Biotechnol.* **18**, 750-753.
6. Ness, J. E., Welch, M., Giver, L., Bueno, M., Cherry, J. R., Borchert, T. V. *et al.* (1999). DNA shuffling of subgenomic sequences of subtilisin. *Nature 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. *Nature Biotechnol.* **16**, 258-261.
10. Volkov, A. A., Shao, Z. & Arnold, F. H. (2000). Random chimeragenesis by heteroduplex recombination. *Methods Enzymol.* **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. *Nucl. Acids Res.* **26**, 681-683.
12. Coco, W. M., Levinson, W. E., Crist, M. J., Hektor, H. J., Darzins, A., Peinkos, P. T. *et al.* (2001). DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nature 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. *Nucl. Acids Res.* **28**, e88.
17. Zhao, H. & Arnold, F. H. (1997). Optimization of DNA shuffling for high fidelity recombination. *Nucl. Acids Res.* **25**, 1307-1308.
18. Polz, M. F. & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Appl. 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. *Nucl. Acids Res.* **24**, 4501-4505.
21. Moore, G. L., Maranas, C. D., Lutz, S. & Benkovic, S. J. (2001). Predicting crossover generation in DNA shuffling. *Proc. Natl Acad. Sci. USA,* **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. *Mol. Divers.* **2**, 111-118.
25. Voigt, C. A., Kauffman, S. & Wang, Z. G. (2001). Rational evolutionary design: the theory of *in vitro* protein evolution. *Advan. Protein Chem.* **55**, 79-160.
26. 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.
27. Hansson, L. O. & Mannervik, B. (2000). Use of chimeras generated by DNA shuffling: probing structure-function relationships among glutathione transferases. *Methods Enzymol.* **328**, 463-477.
28. Petrounia, I. P. & Arnold, F. H. (2000). Designed evolution of enzymatic properties. *Curr. Opin. Biotechnol.* **11**, 325-330.
29. 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.
30. 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.
31. Mondello, F. J. (1989). Cloning and expression in *Escherichia coli* of *Pseudomonas* strain LB400 genes encoding polychlorinated biphenyl degradation. *J. Bacteriol.* **171**, 1725-1732.
32. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
33. 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**, 219-223.

Edited by P. E. Wright

(Received 5 October 2001; received in revised form 14 December 2001; accepted 14 December 2001)

http://www.academicpress.com/jmb

Supplementary Material is available on IDEAL