Creating multiple-crossover DNA libraries independent of sequence identity

Stefan Lutz*, Marc Ostermeier†, Gregory L. Moore‡, Costas D. Maranas‡, and Stephen J. Benkovic*§

†Department of Chemical Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2694; ‡Department of Chemical Engineering, Pennsylvania State University, 112A Fenske Laboratory, University Park, PA 16802; and *Department of Chemistry, Pennsylvania State University, 414 Wartik Laboratory, University Park, PA 16802

Contributed by Stephen J. Benkovic, August 6, 2001

We have developed, experimentally implemented, and modeled *in silico* a methodology named SCRATCHY that enables the combinatorial engineering of target proteins, independent of sequence identity. The approach combines two methods for recombining genes: incremental truncation for the creation of hybrid enzymes and DNA shuffling. First, incremental truncation for the creation of hybrid enzymes is used to create a comprehensive set of fusions between fragments of genes in a DNA homology-independent fashion. This artificial family is then subjected to a DNA-shuffling step to augment the number of crossovers. SCRATCHY libraries were created from the glycinamide-ribonucleotide formyltransferase (GART) genes from *Escherichia coli* (*purN*) and human (*hGART*). The developed modeling framework *e*SCRATCHY provides insight into the effect of sequence identity and fragmentation length on crossover statistics and draws contrast with DNA shuffling. Sequence analysis of the naive shuffled library identified members with up to three crossovers, and modeling predictions are in good agreement with the experimental findings. Subsequent *in vivo* selection in an auxotrophic *E. coli* host yielded functional hybrid enzymes containing multiple crossovers.

Sequence homology-dependent methods for recombining genes (e.g., DNA shuffling and molecular breeding) have been successful at evolving proteins with improved function (1–8). An inherent limitation of these methods is their dependence on DNA sequence identity for generating diversity. This dependence precludes the creation of crossovers between genes at loci of low homology, biasing crossover positions toward regions of highest homology. In general, a severe bias toward parental recombination is observed when sequences with less than 70% sequence identity are DNA-shuffled. Given the fact that protein structure is more frequently conserved than DNA homology, homology-dependent methods for recombining genes may potentially exclude solutions to protein engineering problems.

The need for a recombination protocol capable of freely exchanging genetic diversity without sequence identity limitations has motivated the creation of *i*ncremental *t*runcation for the *c*reation of *h*ybrid enz*y*mes (ITCHY). ITCHY allows one to create comprehensive fusion libraries between fragments of genes without any sequence dependency (9–11). However, the main drawback of the method, as well as similar techniques (12), is that members of these libraries contain only one crossover per gene. As suggested (13), the DNA shuffling of ITCHY libraries could potentially introduce multiple crossovers between the genes of interest by preserving ITCHY crossovers (prepositioned crossovers) in the starting material and by recombining regions of homology between genes (Fig. 1). This combination of ITCHY and DNA shuffling has been named SCRATCHY.

In this work, using the glycinamide-ribonucleotide formyltransferase (GART) from *Escherichia coli* (PurN) and human (hGART), we experimentally demonstrate that SCRATCHY can generate libraries with multiple crossovers between two genes of interest, independent of sequence homology. In parallel, a computational model of SCRATCHY based on the reported *e*Shuffle algorithm for DNA shuffling (14) is described to predict crossover probability and distribution in the naive SCRATCHY library. The *in silico* case study provides insights on the effects of fragmentation length and sequence identity. The present comparison between experimental and modeling results lays a systematic foundation for exploring solutions to protein engineering problems in a more diverse sequence space. Con-

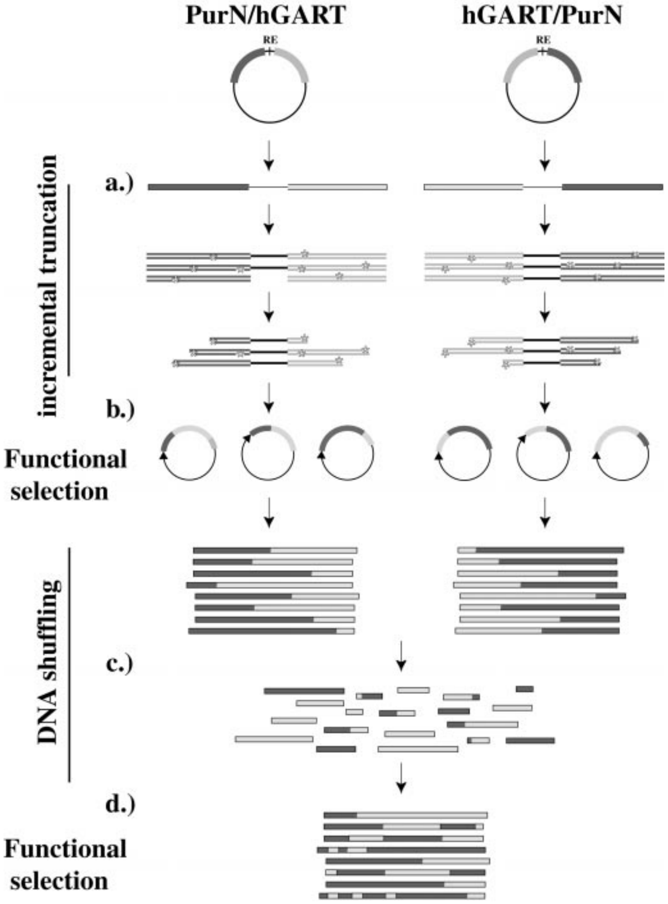


Fig. 1. Schematic overview of SCRATCHY. Initially, individual incremental truncation libraries of the two complementary constructs (pDIM-PGX and pDIM-GPX) were created (*a*). After functional selection (*b*) to recover hybrids of parental size and in-frame, the libraries were mixed and submitted to DNA shuffling (*c*). A final selection (*d*) identifies functional constructs.

Abbreviations: ITCHY, *i*ncremental *t*runcation for the creation of *h*ybrid enzymes; hGART, human glycinamide-ribonucleotide formyltransferase; wt, wild type; PurN, *E. coli* GART; PG, PurN/hGART; GP, hGART/PurN; GG, hGART/hGART.

[§] To whom reprint requests should be addressed. E-mail: sjb1@psu.edu.

www.pnas.org/cgi/doi/10.1073/pnas.201413698

firmation of the presence of function in this expanded sequence space and accessibility by SCRATCHY is provided by the identification of functional hybrid enzymes with more than one crossover.

Materials and Methods

Plasmid Construction. The construction of pDIM-PGX has been described (10). pDIM-GPX was constructed by two substitutions: an initial replacement of the N-terminal *purN* fragment in pDIM-PGX with its *hGART* analog, consisting of DNA coding for amino acids 1–144, followed by the exchange of the C-terminal *hGART* fragment with the corresponding portion of *purN*, consisting of DNA coding for amino acids 54–212.

Incremental Truncation. ITCHY libraries of pDIM-PGX and pDIM-GPX were generated by PCR amplification according to the published protocol for incremental truncation with nucleotide analogs (10). After the nuclease treatment and ligation, the individual libraries were transformed into *E. coli* DH5α-E (Fig. 1*a*).

Selection of the ITCHY Libraries. The incremental truncation libraries were transferred from DH5α-E into the auxotrophic *E. coli* strain TX680F′, and selection of active hybrids (Fig. 1*b*) was carried out on M9 minimum plates [with isopropyl β-D-thiogalactoside (IPTG)] at room temperature as described (15). The plasmid DNA from the resulting colonies was recovered by QIAprep (Qiagen, Valencia, CA).

DNA Shuffling. The two functional hybrid gene libraries were individually amplified by PCR with *Pfu* DNA polymerase (Stratagene), using the following primer pair: T3, 5′-ATTAACCCTCACTAAAGGGA-3′; and PGX$_{rev}$: 5′-ATAAGGGCGACACGGAAATG-3′. After QIAquick PCR purification, the products were quantified by OD$_{260}$ measurement and mixed to equal amounts. Three micrograms of DNA were shuffled essentially as described (16, 17). Briefly, DNaseI treatment-generated fragments (40–100 bp in length), after gel-purification, were reassembled by self-primed PCR, using *Pfu* DNA polymerase (Fig. 1*c*). A second PCR reaction in the presence of specific primer pairs (see Table 1, which is published as supporting information on the PNAS web site, www.pnas.org) yielded product bands of the original gene size, which then were cloned into the pDIM-N2(ΔF′) expression vector, using the *Nde*I and *Spe*I restriction sites (Fig. 1*d*). The DNA was directly transformed into the auxotrophic *E. coli* TX680F′.

Analysis and Selection of the SCRATCHY Library. Random members of the naive libraries were selected and analyzed by PCR and DNA sequencing. Selection of active hybrids was carried out on M9 minimum plates [with and without isopropyl β-D-thiogalactoside (IPTG)] at room temperature and 37°C, respectively, as described (15). The plasmid DNA from a number of the resulting colonies was analyzed by DNA sequencing, and the functionality of the shown hybrids was confirmed by retransformation and selection on M9 plates and ampicillin resistance.

Results and Discussion

Experimental SCRATCHY. Two ITCHY libraries encoding either the PurN/hGART (PGX) or the hGART/PurN (GPX) hybrid pairs were constructed (Fig. 1*a*). After transformation of these libraries, naive libraries of 1×10^6 (pDIM-PGX) and 1.6×10^6 (pDIM-GPX) members, respectively, were obtained, providing extensive coverage of the theoretical library size of 7.3×10^4 possible combinations [(270 bp)^2]. The diversity of both naive libraries was assessed by sequence analysis of several randomly picked members. In either case, members are distributed over the entire sample space, comparable to data from previous libraries (refs. 9 and 10 and Fig. 9, which is published as supporting information on the PNAS web site).

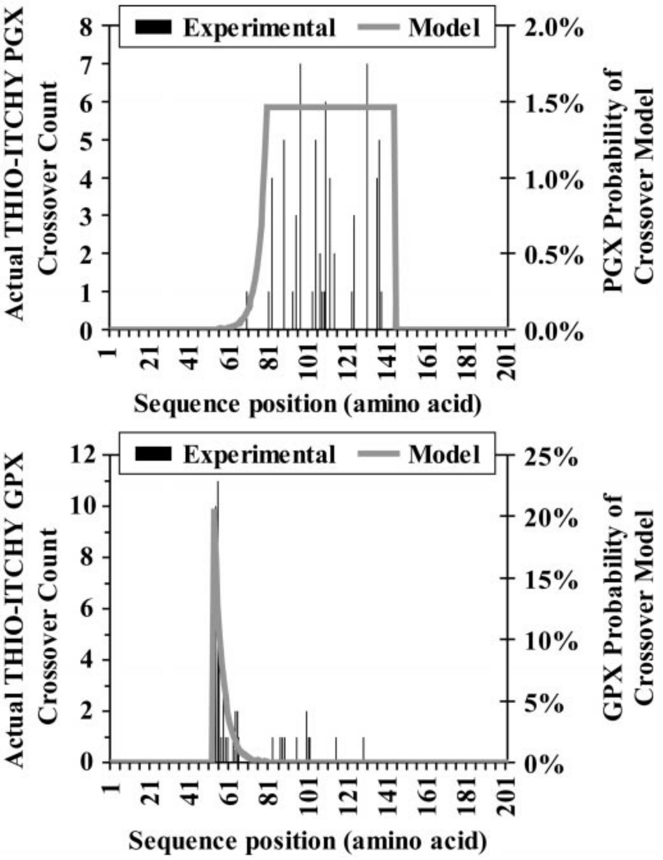


Fig. 2. Profiles of crossover positions for the PGX and GPX libraries, including experimental counts (bars) and smooth-fitted functions of crossover probability (lines).

For the subsequent DNA-shuffling step, hybrid constructs that are approximately of parental size and maintain the correct reading frame beyond the crossover point have the highest potential to be useful in producing functional multicrossover hybrid enzymes. Although a preliminary selection for constructs of parental size can easily be performed by excision of the desired size fragments from an agarose gel, in-frame selection proved difficult. In selecting in-frame constructs by fusing the hybrid gene onto the neomycin-resistance gene, similar to reported chloramphenicol acetyltransferase (CAT) and green fluorescent protein (GFP)-fusion systems (18, 19), we observed a significant amount of false positives for the PGX ITCHY library (data not shown). Presumably, these false positives are the result of internal ribosomal-binding sites in the hybrid gene(s), resulting in expression of kanamycin resistance in a significant percentage of colonies, independent of the reading frame of the hybrid gene. Similar problems have also been also observed in the CAT system (12).

Functional selection provides an alternative approach toward the selection for parental size and in-frame constructs for DNA shuffling. Although the profile of representative sequences in such a library is biased, as shown in Fig. 2, the distribution of the two directional libraries allows for multiple crossovers to occur in the overlapping region. Selection of functional hybrid enzymes from the two ITCHY libraries by complementation of an auxotrophic *E. coli* strain at 22°C yielded ≈150 members per library. Sequence analysis of 22 members of both libraries found 16 unique functional hybrid constructs, distributed over the

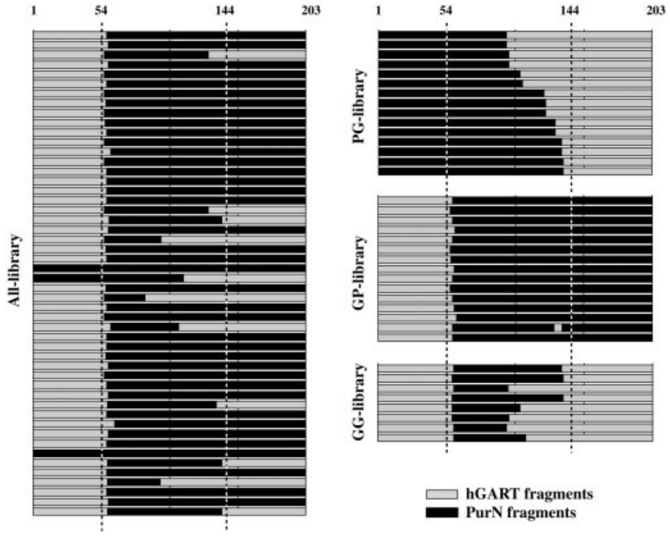


Fig. 3. Naive library sequence data for the All, PG, GP, and GG libraries. The dotted lines indicate the borders of the overlapping region between amino acid positions 54 and 144.

entire range of previously identified functional sequence space (Fig. 9).

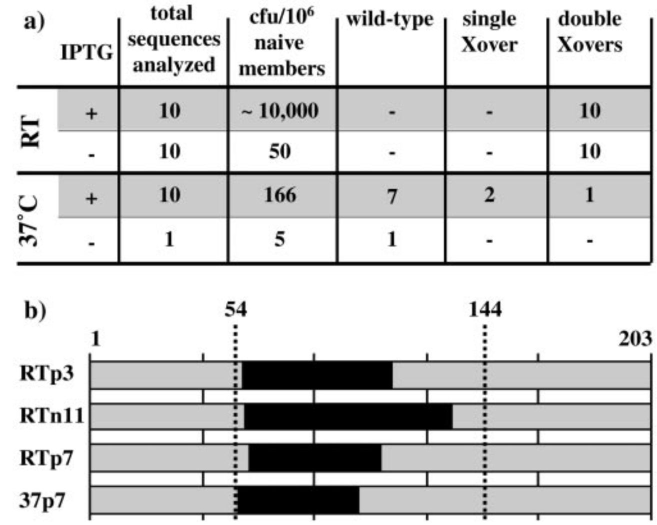
Equal amounts of both selected libraries were DNA-shuffled, and the resulting reassembled sequences were amplified with four individual primer pairs (Table 1). The first pair anneals to outside portions on either side of the gene, yielding a comprehensive library (All library) of possible combinations including wild-type (wt) constructs. The other three primer pairs reach into the N- and C-terminal regions of the hybrid genes, selectively amplifying specific subsets of the shuffled hybrid library. The PG library was amplified by using a *purN*-specific forward primer and an *hGART*-specific reverse primer. The GP library was amplified by using an *hGART*-specific forward primer and a *purN*-specific reverse primer. Both the PG and the GP library selectively filter out members with even numbers of crossovers, including wt, increasing the chances to identify members with higher order crossovers. Finally, the GG library was generated by using *hGART*-specific forward and reverse primers. As predicted from the biased crossover profiles and confirmed by initial sequencing data, the reassembly of wt-hGART is highly unfavorable. As a consequence, the GG library is highly enriched in double-crossover members, making it particularly suitable for the identification of functional hybrids with multiple crossovers. After ligation into the pDIM-N2(ΔF′) vector and transformation into TX680F′, four libraries of 1 to 2 × 10^5 members each were obtained. From these naive libraries, the hybrid genes of over 100 individual colonies were analyzed by DNA sequencing (the results are summarized in Fig. 3).

Naive SCRATCHY Libraries. Analysis of the naive libraries revealed several interesting characteristics. Most importantly, a significant portion of the sample sequences had multiple crossovers. Approximately 20% of the All-library sequences contained two crossover points, and one member of the GP library consisted of four alternating hGART and PurN fragments (three crossovers). When considering the location and number of the crossover points in the sequences, an important experimental bias emerges. The majority of sequences (70%) in the All library are reassembled duplicates of GPX library members, as if the library was present at a higher concentration than the PGX library during DNA shuffling. A series of tests of the DNA-shuffling protocol indicates that the bias arises during fragment reassembly and is intrinsic to the parental sequences. Similar quantitative drifts during coamplification of multiple templates in a single reaction mixture have been reported for competitive PCR experiments (20). Presumably, these effects originate at least partly from differences in thermodynamic parameters of the involved nucleic acids, as well as from competitive binding of duplex DNA to the polymerase, and are augmented further by the low homology between the parental sequences. Experimental optimization of the ratio between the two parental libraries in the starting pools for the DNase treatment led to a more balanced shuffled library (data not shown).

Further examination of the sequencing data reveals a number of additional interesting features. The reassembly of parental wt sequences in SCRATCHY, in contrast to DNA shuffling of low homology sequences, is not dominant. Although few wt-PurN sequences are identified in the naive libraries, wt-hGART is absent. The deficiency of wt-hGART in the recombination mixture is explained by the paucity of a contiguous bridge of hGART fragments traversing the entire gene length as a result of the uneven distribution of fusion points in the two ITCHY libraries (Fig. 2). The same bias, amplified by the higher effective concentration of the GPX library, is also responsible for the predisposition of hGART/PurN/hGART double-crossover sequences over PurN/hGART/PurN hybrids. Reassembly of a PurN/hGART/PurN hybrid requires both a PurN to hGART crossover at the beginning of the overlapping region and an hGART to PurN crossover near the end of the overlapping region. However, both of these crossovers occur infrequently in the starting material, thus explaining their absence. In summary, the data show that the characteristics of the ITCHY libraries are inherited by the SCRATCHY library.

Selected SCRATCHY Libraries. Because of the limitations of the used selection system, the identification of such constructs from the All, PG, or GP libraries would require extensive sampling. The GG library offered an excellent alternative, representing a SCRATCHY sublibrary that is "enriched" in double-crossover constructs. Functional selection for multicrossover hybrid enzymes was performed by plating the shuffled GG library on M9 plates in the presence and absence of isopropyl β-D-thiogalactoside (IPTG). Both systems were grown at two different conditions: at room temperature (RT) or, for more stringent selection, at 37°C. As shown in Fig. 4*a*, selection at ambient temperature yielded a significant percentage of functional hybrids. Sequence analysis of plasmid DNA from 20 randomly picked colonies revealed that the constructs consisted of two crossovers and maintained the correct reading frame. Of the 20 analyzed samples, 14 sequences were unique on the DNA level. At higher temperature, a reduction of the total number of colonies coincided with the identification of a majority of wt-hGART. We hypothesize that the trend is an indicator for the propensity of the multicrossover hybrids to protein misfolding at elevated temperatures, rather than a direct link to catalytic performance. Four representative members of functional double-crossover hybrids are shown in Fig. 4*b*. The identification of functional hybrid enzymes with multiple crossovers demonstrates the potential benefits of SCRATCHY for the exploration of the expanded sequence space.

Modeling SCRATCHY. In conjunction with the experimental work on SCRATCHY, an *in silico* modeling framework for crossover statistics prediction named ***e***SCRATCHY was developed. The modeling framework builds on a recently introduced program (*e*Shuffle) for assessing the generation of crossovers in the context of DNA shuffling (14). The approach utilizes thermodynamic calculations and complete DNA sequence information to model the selectivity of different fragment hybridization events. Then the annealing step statistics are linked with a



	IPTG	total sequences analyzed	cfu/10^6 naive members	wild-type	single Xover	double Xovers
RT	+	10	~ 10,000	-	-	10
	-	10	50	-	-	10
37°C	+	10	166	7	2	1
	-	1	5	1	-	-

Fig. 4. Functional selection of multicrossover hybrids from the GG library. (*a*) Tabulated summary of frequency and characteristics of functional hybrids under the tested selection conditions. (*b*) Four representative examples of functional double-crossover hybrids. The dotted lines indicate the borders of the overlapping region between amino acid positions 54 and 144.

reassembly algorithm to infer crossover allocation statistics in the resulting sequence library. Our algorithm approximates the reassembly process as a series of fragment–fragment annealing events proceeding in the 5′ to 3′ direction. Every possible in-sequence fragment–fragment combination is considered to determine whether a crossover is generated after the extension step. Predictions obtained with ***e***Shuffle were in good agreement with published DNA-shuffling experiments, confirming the aggregation of crossovers in regions of near perfect sequence identity and the presence of synergistic reassembly in family DNA shuffling.

SCRATCHY can be abstracted as the family DNA shuffling of an artificially created superfamily containing all single crossover hybrids between the two genes of interest. The presence of fragments during reassembly that contain prepositioned crossovers extends the sequence space accessed by SCRATCHY compared with the one available to traditional DNA shuffling. Therefore, when fragment–fragment hybridization is considered in the reassembly algorithm of ***e***SCRATCHY, it is necessary to keep track of not only the overlapping region but also whether one or both fragments contain a prepositioned crossover and whether this crossover is located within or outside the overlapping region (Fig. 5). These considerations give rise to three hypothetical yet distinct mechanisms for generating crossovers in contrast to the single mechanism (i.e., the extension of a heteroduplex) encountered in ***e***Shuffle (14). Namely, (*i*) the extension of a heteroduplex as in ***e***Shuffle, (*ii*) the incorporation of a prepositioned crossover, or (*iii*) the extension of a hybrid duplex that occurs when a fragment already containing a prepositioned crossover anneals with another fragment with the crossover positioned in the duplex. Hybrid duplexes are part stabilizing homoduplex and part crossover-generating heteroduplex, presumably enabling the SCRATCHY protocol to generate crossovers within narrower sequence identity stretches than DNA shuffling. It is important to note that these three hypothesized mechanisms reflect, and thus depend on, the abstraction of the proposed reassembly algorithm as a recursive sequence of annealing events. Clearly, the sequence of actual hybridization events occurring in the reacting mixture over multiple cycles defines a process much more complex than the level of detail captured within ***e***SCRATCHY. Specifically, hybrid duplexes may also occur in DNA shuffling, but only after the first reassembly cycle and only between fragments arising from heteroduplex extension in regions of near perfect sequence identity that are largely absent in low sequence identity systems. Annealing choices from all three mechanisms are handled in a straightforward manner within the free energy-based scoring system (14). In addition, the reassembly algorithm is modified to check for each of the three crossover types for every fragment-annealing event.

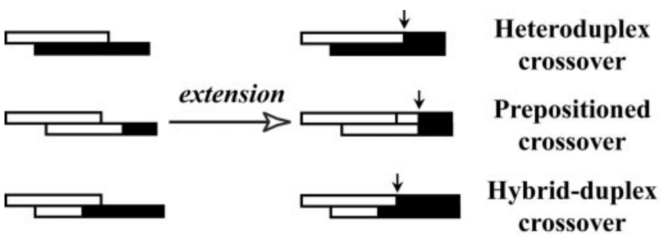


Fig. 5. The three mechanisms for generating crossovers that are tracked *in silico*.

Additional modifications were performed to improve computational performance. The family of single crossover sequences generated in the ITCHY step is much larger than that typically used for molecular breeding, thus the original ***e***Shuffle program (which scales as the square of the number of parental sequences) was customized. Specifically, fragments with identical sequences from different ITCHY parents were pooled, because they do not change the outcome of fragment–fragment extensions considered by the reassembly algorithm. By aggregating their concentrations instead of considering them separately, computing times were reduced to scale linearly with the number of parental sequences. In addition, we found that for fragmentation lengths longer than 40-nt, approximating individual duplex melting curves as step functions at the melting temperature of the duplex provides a tractable and accurate approximation of the annealing thermodynamics, because melting temperatures for larger fragments are significantly above the applied annealing temperature. A 40-nt fragment reassembly confirmed that predictions vary by less than 5% when this approximation is used.

eSCRATCHY was next used to address questions concerning the preservation of prepositioned crossovers in reassembled sequences, as well as its contribution toward multiple-crossover sequences in comparison with those that also occur in homology-based reassembly. In particular, the effect of fragmentation length and pairwise sequence identity on the number and positioning of crossovers produced and the relative contribution of each of the three postulated crossover mechanisms were examined.

In Silico* Case Study.** The *purN*/*hGART* system is first examined in detail. In this case study, both in-frame and parental size selection are "idealized" so that the crossovers present in the ITCHY library are not biased in any manner. Predictions from ***eSCRATCHY indicate that 52% of the reassembled sequences have multiple crossovers for a fragmentation length of 60 nucleotides even though the nucleotide sequence identity is only 49% in the overlapping region. Note that even for fragments as short as 20 nucleotides, predictions by ***e***Shuffle indicate that almost 99.9% of sequences reassembled by DNA shuffling alone will be wt. Interestingly, in contrast to DNA shuffling, ***e***SCRATCHY predicts that fragmentation length has little, if any, effect on the average number of crossovers produced per sequence (Fig. 6). Smaller fragments imply that more annealing choices are available during reassembly and thus more opportunities to generate crossovers, but at the same time, a smaller

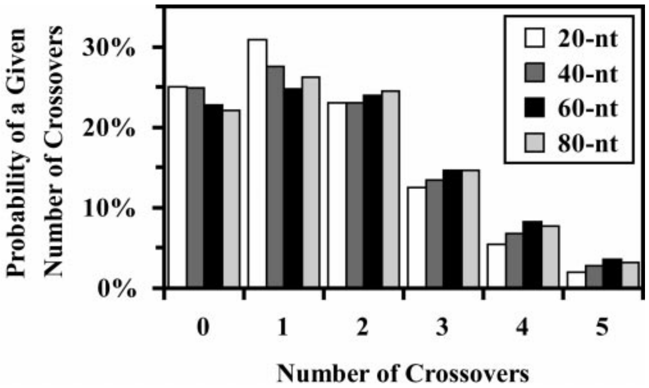


Fig. 6. Probability that a hybrid sequence contains a given number of crossovers after the idealized SCRATCHY of PurN and hGART for fragmentation sizes of 20, 40, 60, and 80 nucleotides (54°C annealing temperature). Note that the distributions are similar for each of the sizes.

proportion of fragments contain prepositioned crossovers. These two effects seem to cancel each other for systems with low sequence identity. Thus, relatively large fragments can be used in SCRATCHY without reducing the number of crossovers, allowing for easier purification, isolation, and reassembly.

In addition, predictions suggest that neglecting hybrid-duplex crossovers in *e*SCRATCHY would produce drastically different results, as these crossovers contribute 47% of the total number of crossovers. This "emergent" mechanism, not present in *e*Shuffle, is almost as frequent as the prepositioned crossover mechanism. Heteroduplex crossovers are negligible as expected for a system with 49% sequence identity. The distribution of crossovers along the sequence is shown in Fig. 10, which is published as supporting information on the PNAS web site. Prepositioned crossovers are present almost uniformly along the entire sequence, showing that the unbiased nature of the ITCHY library is retained. In contrast, hybrid duplex-based crossovers track regions of high sequence identity and involve a less even distribution. Contrary to the homology-based methods, the sum of all types of crossovers fills the entire sequence length with an average frequency of 0.65% per position. The "signature" of DNA shuffling can still be detected in the form of peaks tracking regions of high sequence identity.

Next, we examined the effect of pairwise sequence identity on crossover frequencies for the recombination of the following six sequences with *purN* by using *e*SCRATCHY and *e*Shuffle (sequence identity with *purN* in the overlapping region in parentheses): GAR transformylases from human (49%), *Pseudomonas aeruginosa* (54%), *Pasteurella multocida* (60%), *Vibrio cholerae* (64%), *Salmonella typhimurium* (79%), and methionyl-tRNA formyltransferase from *E. coli* (33%). As seen in Fig. 7, predictions suggest that SCRATCHY is capable of generating crossovers for all sequence pairs, regardless of sequence identity. On the other hand, DNA shuffling requires an approximate "threshold" sequence identity of 60% before any appreciable crossover generation occurs. Even for high sequence identities, we predict that SCRATCHY outperforms DNA shuffling by an average of 1.5 crossovers per sequence. Both prepositioned and hybrid-duplex crossover mechanisms remain prevalent for the entire range of sequence identities, and the heteroduplex mechanism begins to contribute at identities greater than 60% (Fig. 7). After using parameters reflecting the specifics of the actual experimental library described earlier, the predictions of *e*SCRATCHY of the naive *purN*/*hGART* SCRATCHY All library were reexamined and compared with the experimental data.

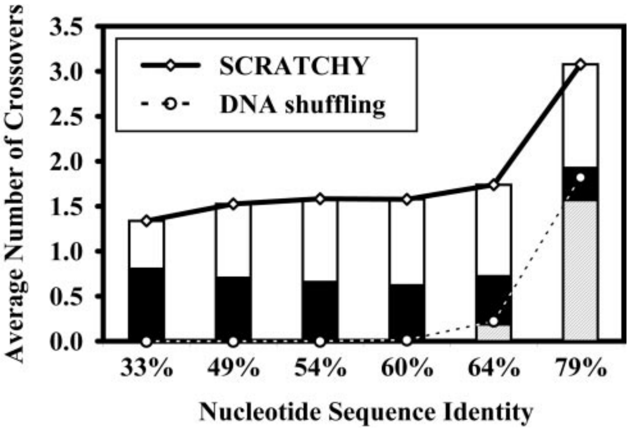


Fig. 7. A comparison of the numbers of crossovers predicted for idealized SCRATCHY and DNA shuffling for sequence pairs of various sequence identities (20-nt fragments, 54°C annealing temperature). White bars, contributions to SCRATCHY from prepositioned crossovers; black bars, hybrid-duplex crossovers; and crosshatched bars, heteroduplex crossovers.

Comparison of Experiment and Model. The main objective of this project, in addition to the experimental demonstration of multiple crossover generation by SCRATCHY, was the development and testing of an *in silico* protocol that will guide and support experimental efforts in future studies. Accurate *in silico* analysis required the integration of two experimental presets: the crossover distribution of the used ITCHY libraries and the fragment reassembly based bias toward hGART/PurN library members. First, the uneven distribution of crossovers caused by the functional selection of the ITCHY library was accounted for in the *e*SCRATCHY program by fitting the observed crossover data with a smooth function (Fig. 2), thus customizing the relative concentration of each of the ITCHY library members. Second, as seen in the naive All library, hGART/PurN library members dominate the reassembly process. This effect was accounted for by adjusting the concentration ratio of the two libraries to 86% GPX:14% PGX. This ratio was calculated by examining the 5′ and 3′ termini of the All-library members. The relative effective concentration of the GPX library was estimated by counting the number of sequences beginning with hGART (47) and ending with PurN (39). Similarly, the PGX library estimate totaled 14 (3 + 11), resulting in the 86:14 ratio. Together, these two modifications result in crossover predictions that are in good agreement with the experimental sequence data for the naive All library. The distribution matches well with what is found experimentally (Fig. 8*a*). The discrepancy between the numbers of multiple crossovers predicted in the idealized case (Fig. 6) and found in the experiment can be attributed to the bias in the starting material. In addition, predictions for crossover position statistics (Fig. 8*b*) capture the uneven nature of crossovers found in the reassembled sequences as a result of the same bias, which also leads to an increased 3.6:1 ratio of prepositioned/hybrid-duplex crossovers compared with the idealized case.

Another interesting aspect is the contribution of crossovers originating from incremental truncation or homology-based recombination. Experimentally, all fusion points observed in the SCRATCHY libraries have counterparts at locations corresponding to prepositioned crossovers, originating from the ITCHY libraries. However, the origin of the crossovers in the homologous region between amino acids 104–114 cannot conclusively been attributed to ITCHY or DNA shuffling. In the *e*SCRATCHY model, heteroduplex crossovers are rare across the entire sequence.

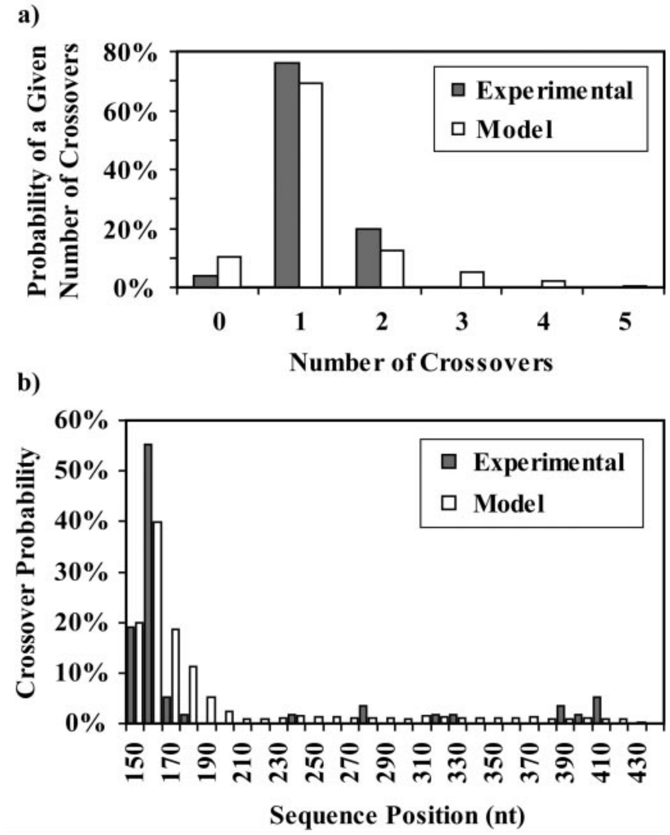


Fig. 8. Comparing *e*SCRATCHY predictions (fragmentation length, 70 nucleotides; annealing temperature, 54°C) for (*a*) the number of crossovers per naive library member and (*b*) naive library crossover positions against experimental data. In *b*, data are grouped in histogram form with each bar representing a range of 10 nucleotides.

Conclusions

SCRATCHY has been implemented experimentally and successfully modeled *in silico*. In the laboratory, the method created functional hybrid enzymes derived from multiple parental fragments. In comparison, the moderate sequence homology of only 49% between PurN and hGART makes construction of similar hybrids by established DNA-shuffling protocols impossible. Although preliminary data indicated less than wt activity for functional SCRATCHY library members with multiple crossovers, our results show that the extended sequence space, accessible exclusively by SCRATCHY, contains function and may provide proteins with changed and improved properties.

In the current work, functional selection of the incremental truncation library was used to sort out sequences suitable for the subsequent DNA-shuffling step. Although acceptable to demonstrate the fundamentals of SCRATCHY on the GART model system, the introduced bias, as reflected in the distorted crossover profiles, is not generally desirable. Likely, many single crossovers that do not produce function on an individual base but may be successful in combination with other crossovers are lost as a result of the functional selection. We have previously created such unbiased SCRATCHY libraries between two genes with 33% identity and found them to contain members with multiple crossovers (M.O., unpublished data). However, the analysis of this library was complicated by problems with in-frame selection, presumably caused by internal ribosome-binding sites. Although we appreciate that these sites may be removed by silent mutagenesis, we find such an approach to be cumbersome and suggest that new in-frame selection methods need to be developed.

In parallel, we have developed a modeling framework named *e*SCRATCHY to address *in silico* questions concerning the application of the SCRATCHY protocol. Crossovers prepositioned in the ITCHY step were shown to be preserved in the reassembled sequences, and the formation of multiple-crossover hybrids was correctly predicted. In contrast to DNA shuffling alone, fragmentation length has little effect on the predicted number of crossovers. The presence of highly biased concentrations at the two ends of the reassembled genes implies that a more complex reassembly procedure may be at hand, warranting further investigation. We demonstrated *in silico* that SCRATCHY outperforms DNA shuffling by an average of 1.5 crossovers per sequence, even at high homologies. Higher levels of recombination have been shown to facilitate the evolution of proteins having desired characteristics (21). Thus, in addition to accessing sequence space that is unattainable by DNA shuffling alone, SCRATCHY may be preferable for the recombination of highly homologous genes as well.

We thank Dr. Shankar Vaidyaraman for help with the model implementation and Judith E. Mank for technical assistance with the DNA sequencing. S.L. was a recipient of a postdoctoral fellowship from the Swiss National Science Foundation. C.D.M. acknowledges support of the National Science Foundation Career Award CTS9701771.

1. Moore, J. C. & Arnold, F. H. (1996) *Nat. Biotechnol.* **14,** 458–467.
2. Crameri, A., Whitehorn, E. A., Tate, E. & Stemmer, W. P. (1996) *Nat. Biotechnol.* **14,** 315–319.
3. Crameri, A., Dawes, G., Rodriguez, E., Silver, S. & Stemmer, W. P. (1997) *Nat. Biotechnol.* **15,** 436–438.
4. Ness, J. E., Welch, M., Giver, L., Bueno, M., Cherry, J. R., Borchert, T. V., Stemmer, W. P. & Minshull, J. (1999) *Nat. Biotechnol.* **17,** 893–896.
5. Joo, H., Lin, Z. & Arnold, F. H. (1999) *Nature (London)* **399,** 670–673.
6. Chang, C. C., Chen, T. T., Cox, B. W., Dawes, G. N., Stemmer, W. P., Punnonen, J. & Patten, P. A. (1999) *Nat. Biotechnol.* **17,** 793–797.
7. Powell, S. K., Kaloss, M. A., Pinkstaff, A., McKee, R., Burimski, I., Pensiero, M., Otto, E., Stemmer, W. P. & Soong, N. W. (2000) *Nat. Biotechnol.* **18,** 1279–1282.
8. Schmidt-Dannert, C., Umeno, D. & Arnold, F. H. (2000) *Nat. Biotechnol.* **18,** 750–753.
9. Ostermeier, M., Shim, J. H. & Benkovic, S. J. (1999) *Nat. Biotechnol.* **17,** 1205–1209.
10. Lutz, S., Ostermeier, M. & Benkovic, S. J. (2001) *Nucleic Acids Res.* **29,** E16.
11. Ostermeier, M. & Benkovic, S. J. (2001) *Biotech. Lett.* **23,** 303–310.
12. Sieber, V., Martinez, C. A. & Arnold, F. H. (2001) *Nat. Biotechnol.* **19,** 456–460.
13. Ostermeier, M., Nixon, A. E. & Benkovic, S. J. (1999) *Bioorg. Med. Chem.* **7,** 2139–2144.
14. Moore, G. L., Maranas, C. D., Lutz, S. & Benkovic, S. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 3226–3231.
15. Ostermeier, M., Nixon, A. E., Shim, J. H. & Benkovic, S. J. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 3562–3567.
16. Stemmer, W. P. (1994) *Nature (London)* **370,** 389–391.
17. Zhao, H. & Arnold, F. H. (1997) *Nucleic Acids Res.* **25,** 1307–1308.
18. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. (1999) *Nat. Biotechnol.* **17,** 691–695.
19. Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D. & Davidson, A. R. (1999) *Protein Sci.* **8,** 1908–1911.
20. Sugimoto, T., Fujita, M., Taguchi, T. & Morita, T. (1993) *Anal. Biochem.* **211,** 170–172.
21. Coco, W. M., Levison, W. E., Crist, M. J., Hektor, H. J., Darzins, A., Pienkos, P. T., Squires, C. H. & Monticello, D. J. (2001) *Nat. Biotechnol.* **19,** 354–359.