ReX: A suite of computational tools for the design, visualization, and analysis of chimeric protein libraries

Weiliang Huang[1,2*], Wayne A. Johnston[2], Mikael Boden[2], and Elizabeth M.J. Gillam[2*]

[1]*School of Pharmacy, The University of Maryland, Baltimore, MD and* [2]*School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, Australia*

BioTechniques 60:91-94 (February 2016) doi 10.2144/000114381

Keywords: crossover; directed evolution; DNA shuffling; recombination; restriction enzyme; chimera; mosaic; mutant library; SCHEMA; Staggered Extension Process; PCR; artificial evolution

Supplementary material for this article is available at www.BioTechniques.com/article/114381.

Directed evolution has greatly facilitated protein engineering and provided new insights into protein structure–function relationships. DNA shuffling using restriction enzymes is a particularly simple and cost-effective means of recombinatorial evolution that is well within the capability of most molecular biologists, but tools for the design and analysis of such experiments are limited. Here we introduce a suite of freely available online tools to make the construction and analysis of chimeric libraries readily accessible to the novice. REcut (http://qpmf.rx.umaryland.edu/REcut.html) facilitates the choice of DNA fragmentation strategy, while Xover (http://qpmf.rx.umaryland.edu/Xover.html) analyzes chimeric mutants to reveal recombination patterns and extract quantitative data.

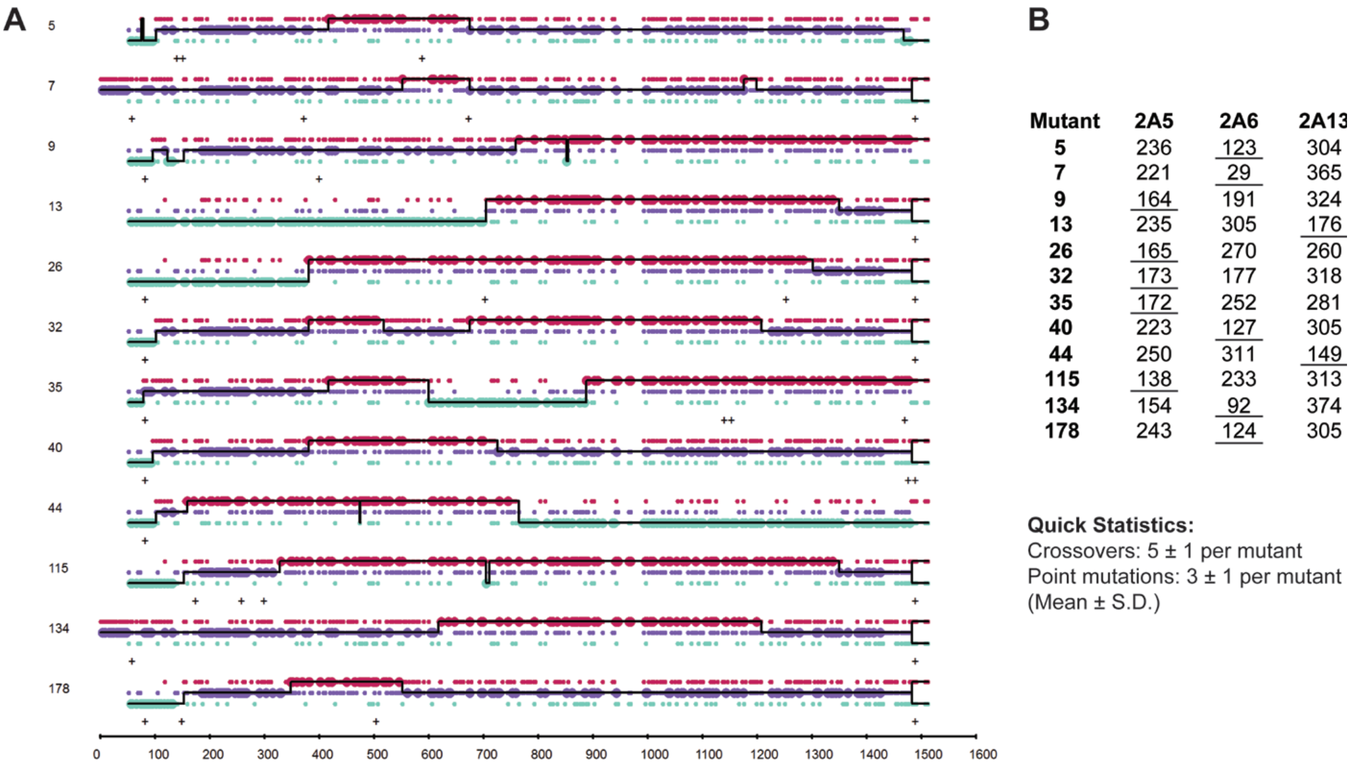
Over the last two decades, recombinatorial-directed evolution techniques such as DNA family shuffling (1) have revolutionized protein engineering. Chimeric mutants generated by recombination of two sequences can also provide important starting points for structure–function analyses that can then be pursued further by site-directed mutagenesis (2). DNA family shuffling involves the production of fragments of DNA from a set of parental genes followed by homology-dependent recombination of fragments in a primerless PCR and subcloning of the resultant chimeric (or mosaic) mutants into a suitable expression vector for screening or selection. While simple in theory, the original DNA shuffling procedure can be difficult to accomplish due to the need to optimize the DNase-mediated fragmentation of parental sequences prior to reassembly of chimeric progeny sequences. Too little fragmentation potentiates reassembly of parental sequences, whereas too much fragmentation leads to non-specific hybridization during the reassembly. A straightforward and efficient alternative method that avoids this problem is restriction-enzyme (RE)–mediated DNA shuffling (3,4), where REs are used for the fragmentation step. This modification allows enhanced control over the size of DNA fragments generated, which is important for library quality (5) but difficult to achieve with DNase. In particular, DNA fragments of ~100–300 bp are favored for DNA shuffling since they minimize regeneration of parental sequences on the one hand and mis-hybridization during the reassembly PCR on the other.

RE-mediated DNA family shuffling is well within the technical capabilities of most beginner molecular biologists, requiring only the ability to perform RE digests, PCR amplifications, and gel extractions. However, computational tools are lacking for the choice of enzymes for the fragmentation step, which is a non-trivial computation given the importance of finding appropriate combinations of REs that can cleave parental genes into DNA fragments with a suitable size and cleavage pattern. Moreover, software is not widely accessible for the analysis of recombination patterns in the chimeric progeny, a challenge common to all DNA chimeragenesis experiments (6,7). Here we report ReX (REcut-Xover), a suite of online tools developed to facilitate the computer-aided design of restriction fragmentation and automated analysis of DNA/protein chimeric patterns. Both REcut and Xover run with all modern web browsers and are made freely available to the research community via the following sites: http://qpmf.rx.umaryland.edu/REcut.html and http://qpmf.rx.umaryland.edu/Xover.html. Web browsers supporting the HTML5 standard, such as Google Chrome and Mozilla Firefox, are recommended but not essential. Tasks can be submitted from Windows, Linux, Macintosh, iOS, and Android devices.

REcut enables the determination of optimal RE sets for fragmentation of genes for DNA shuffling. It computes all possible cleavage patterns that can be generated by different enzyme

METHOD SUMMARY

Here we present ReX, a suite of programs for the design, analysis and depiction of chimeric mutants created by recombinatorial mutagenesis.



Mutant	2A5	2A6	2A13
5	236	$\underline{123}$	304
7	221	$\underline{29}$	365
9	$\underline{164}$	191	324
13	235	305	$\underline{176}$
26	$\underline{165}$	270	260
32	$\underline{173}$	177	318
35	$\underline{172}$	252	281
40	223	$\underline{127}$	305
44	250	311	$\underline{149}$
115	$\underline{138}$	233	313
134	154	$\underline{92}$	374
178	243	$\underline{124}$	305

Quick Statistics:
Crossovers: 5 ± 1 per mutant
Point mutations: 3 ± 1 per mutant
(Mean ± S.D.)

Figure 1. Crossover plot and statistics returned by Xover for selected chimeras from a cytochrome P450 2A subfamily DNA family shuffling library created by RE-mediated DNA shuffling of P450s 2A5, 2A6, and 2A13. (A) Crossover plot: Matches to different parental sequences are shown in different colors: red—P450 2A5; purple—P450 2A6; green—P450 2A13. Large dots represent 100% parental match (i.e., the position in question matches only one parent) and small dots represent more than one parental match (i.e., the residue present in the mutant matches more than one parent). Chimeras are identified by the numbers shown to the left of each set of dots. The solid line for each chimera represents the parent from which it is most likely the sequence is derived. A set of horizontal parallel lines between crossovers indicates a match to multiple parents at an equal probability (e.g., at the 3´ ends of mutants 7, 13, 26, 32, 40, 115, 134, and 178). A mutation is recorded as a plus sign (+) when the residue in a chimera does not match the corresponding position in any parental sequence. A vertical spike (e.g., at position 474 in mutant 44), indicates a single position switch between parents. (B) Statistics for the chimeras shown in (A): Mean crossovers and point mutations per mutant and table of Levenshtein distances (i.e., the minimum number of point mutations required to convert each parent into the mutant in question). The effective mutation is underlined (i.e., the minimum point mutations required to back-mutate a mutant to its closest parent).

combinations on a set of parental sequences and selects the best-suited enzyme combinations to yield fragments within a user-specified size range. Using a database of RE recognition sites compiled based on the New England Biolabs REBASE database (http://rebase.neb.com) (8), the program finds suitable RE sets to fragment the input parental DNA sequences according to the user-defined minimum and maximum fragment lengths. REcut also reports on REs useful for individual parents, which allows the user to individually cut parents with low sequence identity.

For RE-mediated DNA shuffling, two or more combinations of REs are needed to generate sets of DNA fragments with overlapping ends to allow ORF reassembly by PCR. Thus, fully or partially overlapping cleavage sites should be avoided. The minimum number of overlapping bases required for annealing can be set by the user (e.g., for parents that share 80% sequence identity and 50% GC content, self-priming of 2 overlapping fragments would require at least 19 bases of overlap to ensure proper annealing at a T_{ann} of ~ 45°C, as estimated using the formula of Marmur) (9). From the preview page, users can input the minimum overlap and query REcut to select the most appropriate combination of RE sets. A table of minimum bases required for annealing based on the degree of parental sequence identity is provided for the convenience of users in the online help page. An example of the output provided is shown in Supplementary Figures S1 and S2. Alternatively, users can also select RE sets manually (e.g., according to the enzymes at hand in the laboratory) and let REcut verify their complementarity using the manual option (Supplementary Figure S3).

The second part of the suite, Xover, analyzes DNA and protein crossover patterns generated by DNA shuffling or any other DNA recombination technique. Chimeric mutant DNA or protein sequences are submitted in FASTA format along with the corresponding parental sequences. Xover aligns input sequences automatically using Clustal Omega (version 1.2) (10). If pre-aligned sequences are provided, the gaps in the sequences are taken into consideration during the alignment. At each position in the alignment, the residue present in the chimera is compared with that present in each parent. If a match is found at the corresponding position in *n* parents, then the probability that the residue in the chimera is inherited from one of those parents is scored as 1/n.

In the recombination plot generated by Xover, each chimeric sequence is represented by a set of lines of dots, corresponding to each of the library parents depicted in different colors. The size of the dot at each position is proportional to the probability that the residue in the mutant comes from the designated parent; full-size dots are shown when, at the position in question, the residue present in the mutant matches the sequence of only one parent. No symbol is assigned at positions where all parents are identical. Fragments having the longest uninterrupted match with one parent are chosen as the most conservative representation of the recombination pattern of a chimera. The solid line for each chimera represents the recombination pathway between parents. Where multiple pathways are equally probable between recombination points, they are shown as parallel lines over the relevant section of the alignment (e.g., at the 3´ ends of mutants 7, 13, 26, 32, 40, 115, 134, and 178 at around 1480–1520 bp in Figure 1).

At positions where the chimera does not match the corresponding position in any parental sequence, a point mutation is recorded as a plus sign (+). A vertical spike indicates a single position switch between parents (e.g., the spike downwards at ~474 bp in mutant 44 in Figure 1) and may reflect point mutations that coincidentally match another parent or extremely frequent template switching during PCR. A horizontal black bar indicates a fragment deletion/insertion (not shown in Figure 1). A compressed crossover graph is shown on the results page for fast display; the full resolution graph can be downloaded in BMP format for publication use.

The displayed output from Xover includes the mean and standard deviation (SD) of the number of crossovers and point mutations across the set of progeny sequences. A full data set spreadsheet can be downloaded in CSV format, containing all the analyzed data that were used to plot the crossover graph, namely sequence alignments, per-position parent calling, point mutation detection, and detailed crossover information. Xover also provides two unique indicator values for evaluating library quality: Levenshtein distance and effective mutation score. In comparison to simple estimations of the percentages of sequence derived from different parents, Levenshtein distance indicates the sequence difference between mutants and parents (i.e. the minimum number of point mutations required to convert a parent into a mutant in terms of either nucleotides or amino acids, depending on the sequences being analyzed). The parent showing the shortest Levenshtein distance is the closest in sequence to the mutant, and this distance is defined as the effective mutation score. Levenshtein distance and effective mutation score are not affected by identical sequences shared between parents, and so they give an objective evaluation of relationships between parents and mutants in a quantitative and unambiguous way.

Xover provides users with an intuitive and information-rich way to rapidly analyze and depict a large volume of chimeric sequence data automatically and can be used to analyze essentially

any DNA or protein library sequences, such as those data generated by StEP (11), SCHEMA (12), and CLERY (13). In comparison to traditional representations of recombined sequences as multi-colored bars, the customized crossover graph produced by Xover allows a more objective, information-rich presentation of complex chimeric sequence patterns. The CSV file containing the numerical data output can be used for further analysis with spreadsheet programs such as Microsoft Excel and OpenOffice Calc or can be viewed using any plain text editor.

Xover is designed to be fully automatic and requires no data pre-processing or knowledge of scripting, in contrast to the only other programs reported to date for analysis of recombination patterns: Shuffled (6) and Salanto (7). Shuffled can analyze pre-aligned sequences and output a traditional mosaic bar graph. However, the user is required to have knowledge in setting up scripting environments and using system shell commands in order to use it independently. (Over multiple attempts we were unable to successfully query the web interface.) Salanto requires extra setup and data pre-processing by the user, and can only output results in numbers. By contrast, Xover provides an intuitive and information-rich way to rapidly analyze and depict a large volume of chimeric sequence data automatically, and its user-friendly web interface makes it amenable to widespread use.

Together, the two computational tools in the ReX suite should make DNA shuffling accessible to more molecular biologists and support a wider application of recombinatorial-directed evolution to analysis of structure—function relationships as well as problems in protein engineering.

Author contributions

W.H. developed the software, initiated the REcut idea, and wrote the paper. W.A.J. initiated the Xover idea and contributed to the paper. M.B. provided advice on the software and contributed to the paper. E.M.J.G. provided advice on the user interface and wrote the paper.

Acknowledgments

The original versions of this software were developed as part of the PhD studies of W.H., supported by an Australian International Postgraduate Research Scholarship. The authors gratefully acknowledge the helpful feedback provided by Dominic J. B. Hunter and James B. Y. H. Behrendorff on early versions of this software.

Competing interests

The authors declare no competing interests.

References

1. **Crameri, A., S.A. Raillard, E. Bermudez, and W.P.C. Stemmer.** 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature *391*:288-291.
2. **Behrendorff, J.B, C.D. Moore, K.H. Kim, D.H. Kim, C.A. Smith, W.A. Johnston, C.H. Yun, G.S. Yost, and E.M.J. Gillam.** 2012. Directed evolution reveals requisite sequence elements in the functional expression of P450 2F1 in Escherichia coli. Chem. Res. Toxicol. *25*:1964-1974.
3. **Kikuchi, M., K. Ohnishi, and S. Harayama.** 1999. Novel family shuffling methods for the in vitro evolution of enzymes. Gene *236*:159-167.
4. **Huang, W., W.A. Johnston, M.A. Hayes, J.J. De Voss, and E.M.J. Gillam.** 2007. A shuffled CYP2C library with a high degree of structural integrity and functional versatility. Arch. Biochem. Biophys. *467*:193-205.
5. **Hunter, D.J.B., J.B.Y.H. Behrendorff, W.A. Johnston, P.Y. Hayes, W. Huang, B. Bonn, M.A. Hayes, J.J. De Voss, and E.M.J. Gillam.** 2011. Facile production of minor metabolites for drug development using a CYP3A shuffled library. Metab. Eng. *13*:682-693.
6. **Morett, E. and A.G. Garciarrubio.** 2004. Shuffled: a software suite that assists the analysis of recombinant products resulting from DNA shuffling. Biotechniques *37*:354-358.
7. **Schürmann, N., L.G. Trabuco, C. Bender, R.B. Russell, and D. Grimm.** 2013. Molecular dissection of human Argonaute proteins by DNA shuffling. Nat. Struct. Mol. Biol. *20*:818-826.
8. **Roberts, R.J., T. Vincze, J. Posfai, and D. Macelis.** 2015. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. *43*:D298-D299.
9. **Marmur, J. and P. Doty.** 1962. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. J. Mol. Biol. *5*:109-118.
10. **Sievers, F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, et al.** 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*:539.
11. **Zhao, H.** 2004. Staggered extension process in vitro DNA recombination. Methods Enzymol. *388*:42-49.
12. **Voigt, C.A., S.L. Mayo, F.H. Arnold, and Z.G. Wang.** 2001. Computational method to reduce the search space for directed protein evolution. Proc. Natl. Acad. Sci. USA *98*:3778-3783.
13. **Abécassis, V., D. Pompon, and G. Truan.** 2000. High efficiency family shuffling based on multi-step PCR and in vivo DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. Nucleic Acids Res. *28*:e88.

Received 09 August 2015; accepted 21 September 2015.

Address correspondence to Elizabeth M.J. Gillam, School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, 4072 Australia or Weiliang Huang, School of Pharmacy, The University of Maryland, Baltimore, MD 21201. E-mail: e.gillam@uq.edu.au or weiliang.huang@uqconnect.edu.au

To purchase reprints of this article, contact: biotechniques@fosterprinting.com