METHODS IN ENZYMOLOGY, Vol. 388 (2004).

[4] SCHEMA-Guided Protein Recombination

By JONATHAN J. SILBERG, JEFFREY B. ENDELMAN, and FRANCES H. ARNOLD

Introduction

SCHEMA is a scoring function that predicts which elements in homologous proteins can be swapped without disturbing the integrity of the structure.[1] Using the structural coordinates of the parent proteins, SCHEMA identifies pairs of residues that are interacting and determines the number of interactions, E, that are broken when a chimeric protein inherits portions of its sequence from different parents. E appears to be a good metric for anticipating structural conservation when homologous proteins are recombined. Analysis of well-defined libraries of β-lactamase chimeras revealed that chimeras with low E retained function with higher probability than chimeras with the same effective level of mutation but higher E or chosen at random.[2] Another study also showed that E is a useful measure for anticipating disruption in chimeras of a larger, cofactor-containing protein, cytochrome P450.[3]

Using SCHEMA, libraries of chimeras can be compared *in silico* to determine which one is expected to contain the highest fraction of folded (and potentially interesting) sequences for laboratory evolution studies.[2–4] These libraries can be synthesized *in vitro* using site-directed recombination methods (see Fig. 1), which allow for the simultaneous recombination of two or more parents at specified locations.[2,5] This approach can be used to make chimeric libraries from any parent sequences. In addition, the sequence diversity of folded and functional chimeras encoded in the library can be controlled, i.e., the number of possible unique sequences and the average level of mutation of chimeras predicted to retain structure, can be

[1] C. A. Voigt, C. Martinez, Z. G. Wang, S. L. Mayo, and F. H. Arnold, *Nature Struct. Biol.* **9,** 553 (2002).

[2] M. M. Meyer, J. J. Silberg, C. A. Voigt, J. B. Endelman, S. L. Mayo, Z. G. Wang, and F. H. Arnold, *Protein Sci.* **12,** 1686 (2003).

[3] C. R. Otey, J. J. Silberg, C. A. Voigt, J. B. Endelman, G. Bandara, and F. H. Arnold, *Chem. Biol.* **11,** 309 (2004).

[4] D. A. Drummond, J. J. Silberg, J. B. Endelman, C. A. Wilke, and F. H. Arnold, submitted for publication.

[5] K. Hiraga and F. H. Arnold, *J. Mol. Biol.* **330,** 287 (2003).

0076-6879/04 $35.00

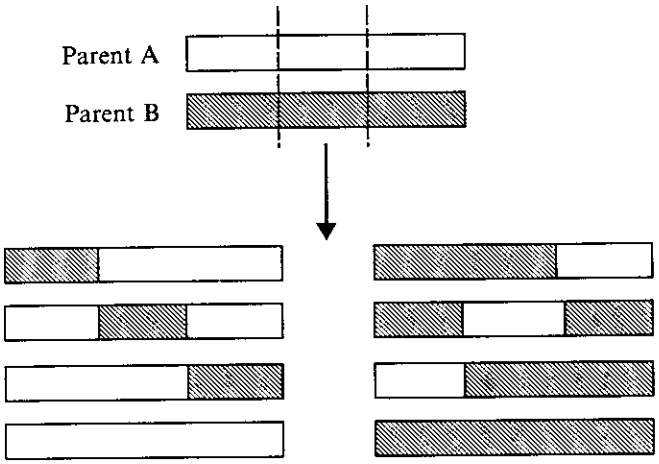

FIG. 1. Library synthesis by site-directed recombination. Sequence elements encoding structurally related polypeptides are swapped at defined locations (dashed lines) in two or more homologous proteins. This yields a library containing $h^y - h$ unique chimeras, where h is the number of parents recombined and y is the number of sequence elements that are exchanged.

used to guide the selection of crossover locations and crossover number. In contrast, annealing-based recombination or "DNA shuffling" techniques, such as Stemmer shuffling,[6,7] StEP,[8] and *in vivo* methods,[9] generate crossovers only in regions of sequence identity and therefore can not generate diverse libraries from more distant parent sequences. The sequence-independent random recombination methods now available (SHIPREC,[10] ITCHY,[11] or SCRATCHY[12]) do not make multiple crossovers efficiently and therefore create libraries of very limited diversity.

This article outlines the procedure used for calculating E for a chimera and discusses ideas for optimizing the design of combinatorial libraries for directed evolution.

[6] W. P. Stemmer, *Proc. Natl. Acad. Sci. USA* **91,** 10747 (1994).
[7] W. P. Stemmer, *Nature* **370,** 389 (1994).
[8] H. Zhao, L. Giver, Z. Shao, J. A. Affholter, and F. H. Arnold, *Nature Biotechnol.* **16,** 258 (1998).
[9] A. A. Volkov, Z. Shao, and F. H. Arnold, *Nucleic Acids Res.* **27,** e18 (1999).
[10] V. Sieber, A. Pluckthun, and F. X. Schmid, *Nature Biotechnol.* **16,** 955 (1998).
[11] M. Ostermeier, A. E. Nixon, and S. J. Benkovic, *Bioorg. Med. Chem.* **7,** 2139 (1999).
[12] S. Lutz, M. Ostermeier, G. L. Moore, C. D. Maranas, and S. J. Benkovic, *Proc. Natl. Acad. Sci. USA* **98,** 11248 (2001).