

[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).

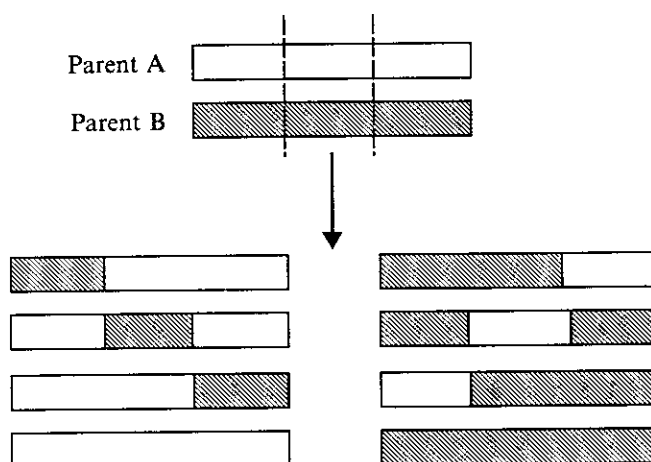


FIG. 1. Library synthesis by site-directed recombination. Sequence elements encoding structurally related polypeptides are swapped at defined locations (dashed lines) in two or more homologous proteins. This yields a library containing h^y-h unique chimeras, where h is the number of parents recombined 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 SCRATCHY¹²) 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).

Methods

Calculating SCHEMA Disruption

Based on the structure of the parent proteins, SCHEMA determines which residues are interacting, defined as those residues within a cutoff distance, and generates a contact matrix.¹ When recombining two parents, the contacts are scaled by the sequence identity of the parents being recombined, i.e., all contacts that cannot be broken by recombination are removed from the matrix. E is determined by counting the number of contacts broken when a chimeric protein inherits portions of its sequence from different parents.

The SCHEMA disruption E of a chimeric sequence s , made by recombining sequence elements from h homologous proteins, is given by

$$E = \sum_{i=1}^N \sum_{j=i+1}^N C_{ij} P(i, j, s_i, s_j), \quad (1)$$

where N is the number of residues that have defined coordinates in the parental structure, $C_{ij} = 1$ if residues i and j are within the cutoff distance d_c (otherwise $C_{ij} = 0$), and s_i designates the parent incorporated at position i in the chimera (e.g., $s_i = 1$ if the sequence is derived from parent #1, $s_i = 2$ if derived from parent #2, etc.). $A(s_i, k)$ is the identity of the residue in parent s_i at position k within the parental amino acid sequence, and $P(i, j, s_i, s_j) = 0$ if any parent has residue $A(s_i, i)$ at position i and residue $A(s_j, j)$ at position j [otherwise $P(i, j, s_i, s_j) = 1$]. It is essential that structurally related residues in each parent are numbered identically, e.g., $A(1, k)$ and $A(2, k)$ should represent structurally related residues in each parent, to ensure that sequence identities among the parents are properly accounted for when calculating E .

Typically, we use $d_c = 4.5 \text{ \AA}$, and hydrogen, backbone nitrogen, and backbone oxygen atoms are excluded from the calculation of E . Small deviations from this value of d_c or the use of all atoms to calculate C_{ij} does not significantly affect the relative E of chimeras being compared, although the magnitude of E changes. When cofactor-containing proteins are recombined, contacts between the cofactor and residues in the proteins are also excluded from calculation of E .³ In this simple model, contacts between the cofactor and the protein cannot be broken by the recombination of related proteins.

Ideally, PDB coordinates for all the parent sequences are available, and a structure-based alignment is performed. For parents whose sequences differ in length, this ensures that structurally related residues in each parent are numbered identically. This can be done using free software packages

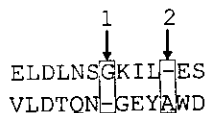


FIG. 2. Treatment of gaps in sequence alignments. A hypothetical sequence alignment used in SCHEMA calculations is shown. PDB coordinates corresponding to the top sequence are being used by SCHEMA to calculate C_{ij} , and the bottom sequence represents a homologous protein for which no structural information is available. At position 1, the atomic coordinates of glycine are defined in the PDB file, so the gap in the second parent is treated as a mutation relative to G when computing E and m . Because there are no coordinates for position 2, it is ignored when computing E and m .

such as SwissProt or the combinatorial-extension algorithm.^{13,14} If structural coordinates are available for different conformational states of the parents being recombined, it is best to assess E using the coordinates for each conformation to ensure that both states of the chimeras are likely to exhibit similar low disruption. When the structure of only one parent is available, sequence alignments can be performed using the BLAST algorithm.¹⁵

Often alignment of the parents requires the insertion of gaps within the primary amino acid sequence of one or more of the parents (see Fig. 2). When gaps are introduced into the parent whose structural coordinates are being used to generate the contact matrix C_{ij} , the residues found in the other parents are ignored when calculating E because there is no corresponding structural information. In contrast, when gaps occur in any parent other than the one used for structural information, they are treated like real residues that differ in identity from the residues in the other parents.

From Disruption to Probabilities

The fraction of chimeras retaining function has been found to decrease exponentially with E .² If we posit that any disrupted contact has a probability f_d of yielding a nonfunctional chimera and each contact acts statistically independently of the others, the fraction of chimeras at each E predicted to retain function is given by $P_f = (1 - f_d E/N)^N$. In this case, N equals the total number of contacts that could be broken by recombination. When N becomes large, as with proteins, this equation can be approximated by a simple exponential, $P_f = e^{-f_d E}$. This relationship between E and P_f is likely to hold for the recombination of any homologous proteins. However, functional data from different chimeric libraries may yield a range of f_d values. We have found that the sensitivity of the

¹³ N. Guex and M. C. Peitsch, *Electrophoresis* **18**, 2714 (1997).

¹⁴ I. N. Shindyalov and P. E. Bourne, *Protein Eng.* **11**, 739 (1998).

¹⁵ S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. USA* **89**, 10915 (1992).