Protein building blocks preserved by recombination

Christopher A. Voigt[1], Carlos Martinez[2], Zhen-Gang Wang[2], Stephen L. Mayo[3] and Frances H. Arnold[2]

Published online: 3 June 2002, corrected 10 June 2002 (details online) doi:10.1038/nsb805

Borrowing concepts from the schema theory of genetic algorithms, we have developed a computational algorithm to identify the fragments of proteins, or schemas, that can be recombined without disturbing the integrity of the three-dimensional structure. When recombination leaves these schemas undisturbed, the hybrid proteins are more likely to be folded and functional. Crossovers found by screening libraries of several randomly shuffled proteins for functional hybrids strongly correlate with those predicted by this approach. Experimental results from the construction of hybrids of two β-lactamases that share 40% amino acid identity demonstrate a threshold in the amount of schema disruption that the hybrid protein can tolerate. To the extent that introns function to promote recombination within proteins, natural selection would serve to bias their locations to schema boundaries.

In vitro recombination is a powerful tool for the tuning and optimization of proteins. It promotes the combination of traits from multiple parents onto a single offspring, thus exploiting information obtained in previous rounds of selection[1–3]. Recombination plays a key role in the natural evolution of proteins, notably in the generation of diverse antibodies, synthases and proteases[4]. In these examples, crossovers occur at well-defined domain boundaries. The role of recombination in evolution is less well understood when the domain structure of a protein is not obvious. Here, we introduce a computational algorithm to divide a protein structure into pieces that can be swapped by recombination and compare the predictions with data generated by *in vitro* recombination experiments.

Ever since the first protein structures were elucidated, researchers have attempted to divide their otherwise complicated topologies into well-defined domains, defined variously as secondary structure units, structural elements that fold independently or clusters of residues close in geometric space[5–12]. An operationally relevant domain definition is a protein fragment that can be swapped among related structures. The locations of certain types of introns were shown to occur at structural domain boundaries, suggesting that larger proteins are composed of smaller domains discovered earlier in evolution and pieced together by gene duplication and recombination[8,13–15]. Using *in vitro* recombination experiments to observe that a crossover is acceptable, rather than inferring it from the existence of introns, provides a direct approach to understanding how domains can be interchanged to create new functional proteins.

The SCHEMA algorithm

Optimal recombination points have been suggested to allow swapping of structural domains[4,16–18]. Identifying what these smaller building blocks look like has been difficult. Research in computer science has demonstrated that the optimal crossover locations in genetic algorithms correspond to those that retain and combine clusters of bits that interact favorably

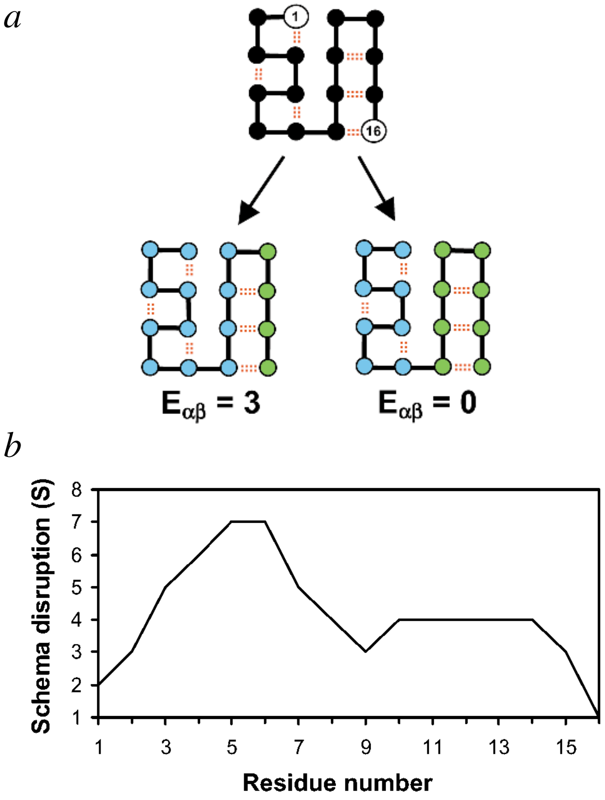

Fig. 1 Illustration of schema disruption. ***a***, Black lines in the structure represent peptide bonds, and the red dotted lines are interactions between amino acid side chains. Two hybrid proteins are shown. When the last four residues come from one parent and the remaining residues come from the other parent, three interactions are disrupted. When the last eight residues come from the same parent, then there is no disruption. According to our schema theory, achieving folded hybrid proteins is more likely when the fewest interactions are disrupted. ***b***, The schema profile of this structure calculated with a window size w = 6.

[1]Biochemistry and Molecular Biophysics, California Institute of Technology, mail code 210-41, Pasadena, California 91125, USA. [2]Division of Chemistry and Chemical Engineering, California Institute of Technology, mail code 210-41, Pasadena, California 91125, USA. [3]Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, mail code 147-75, Pasadena, California 91125, USA.

Correspondence should be addressed to F.H.A. *email: frances@cheme.caltech.edu*, S.L.M. *email: steve@mayo.caltech.edu* or Z.G.W. *email: zgw@cheme.caltech.edu*

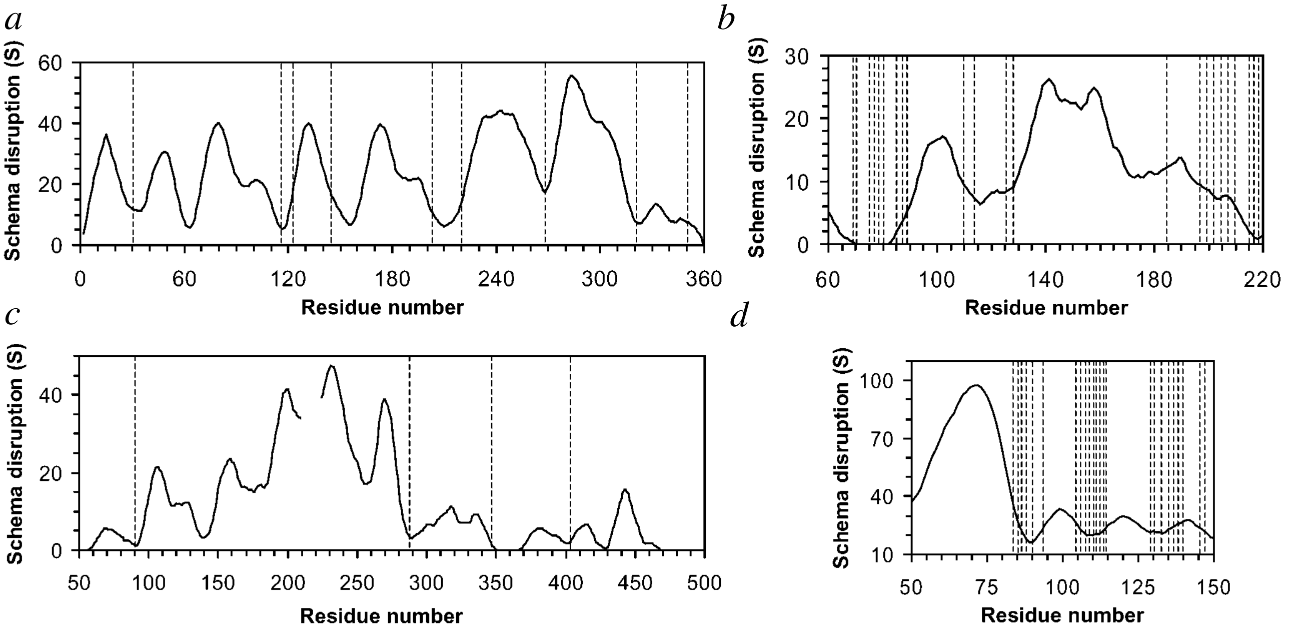

Fig. 2 Schema disruption profiles compared with *in vitro* recombination data. Hatched lines indicate where a recombination event resulted in a functional hybrid protein. All calculations were done using Eq. 2 with a window size of 14 residues and d_c = 4.5 Å. ***a***, Schema profile, as determined from the cephalosporinase structure, compared with the observed crossover points in DNA shuffling[3] and designed hybrid[21] experiments. ***b***, A comparison of the schema profile of Savinase with the crossovers that led to improvement of properties of subtilisin[22]. The crossovers between subtilisins that led to improved thermostability, activity at high or low pH, or stability in organic solvent are indicated. ***c***, A schema disruption calculation of the P450 2C5 structure, based on the sequences of rat and bacterial c17[23]. The dashed lines indicate where single-crossover recombination events led to folded hybrids. Note that residues 212–222 are missing from the structure, represented by a break in the schema profile. ***d***, Schema profile for recombination of PurN and GART glycinamide ribonucleotide transformylase[24,25]. Recombination was allowed to occur only in the 100-amino acid region between residues 50 and 150. The single crossovers that led to functional hybrid proteins are indicated.

(a 'schema')[1,19,20]. Solutions in which recombination divides a schema such that an offspring inherits fractions of it from different parents are generally less fit. To identify the equivalent of schema in proteins, we have developed a computational algorithm, SCHEMA, which can predict fragments that must be inherited from the same parent. Therefore, schemas will be the building blocks from which novel proteins can be assembled by recombination.

SCHEMA calculates the interactions between residues and then determines the number of interactions that are disrupted in the creation of a hybrid protein. A disruption occurs when an interaction is broken when different amino acids are inherited from each parent (Fig. 1). In the simplest implementation, two residues are considered interacting if any of their atoms (excluding hydrogen atoms) are within a cutoff distance d_c = 4.5 Å, which corresponds to 5–8 interactions per residue. Ideally, an algorithm would search all possible crossover combinations and determine the associated disruption for each. Analyzing multiple crossovers by this method leads to combinatorial difficulties both in the calculation and the visualization of the data. SCHEMA overcomes this limitation by scanning the protein structure with a defined window size. Calculating how many interactions are disrupted when a crossover is made generates a schema profile S; if S is large for residue i, then the residue is involved in a more compact schema (see Methods). Crossovers that correspond to minima of the schema profile preserve the maximum number of internal interactions and, therefore, are favored.

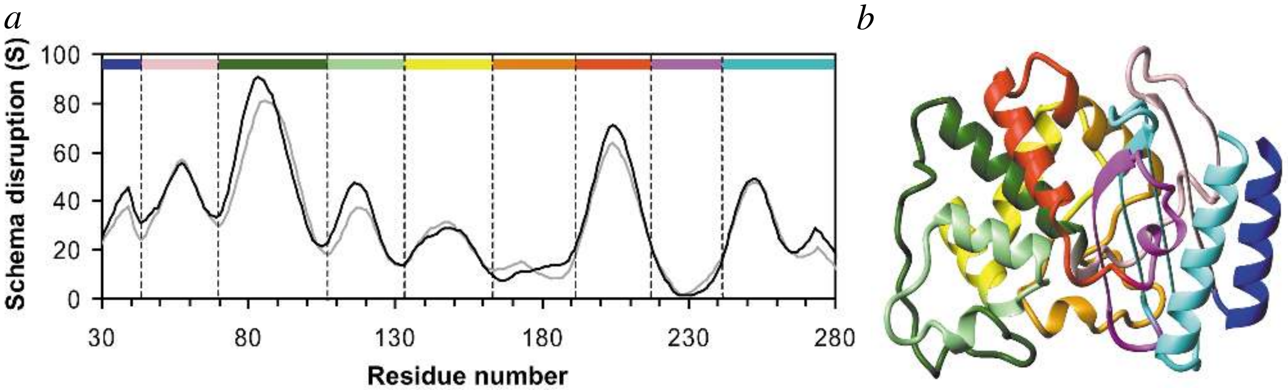

Fig. 3 β-lactamase schema. ***a***, Schema disruption profile for recombination of β-lactamases TEM-1 and PSE-4. Nearly identical results are obtained when the calculation is run on the TEM-1 (gray line) and PSE-4 (black line) structures. The orange and purple regions mark large basins of disruption minima. Crossovers are predicted to be acceptable throughout these basins. ***b***, The predicted schema mapped onto the three dimensional structure of TEM-1 β-lactamase. This figure was generated using MolMol[40].

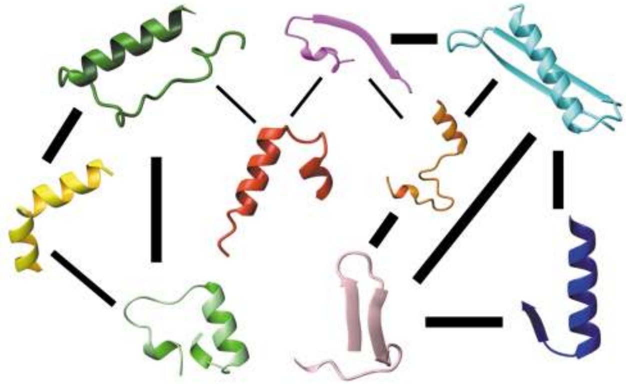

Fig. 4 Interschema interactions. The number of interactions between schema are averaged between the PSE-4 and TEM-1 structures. The thickness of each line is proportional to the number of interactions between two schemas, as calculated using Eq. 1. The thickest lines represent highly interacting schemas (>8 interactions). Medium lines = 5–8 interactions, and thin lines = 2–4 interactions. Note that the purple and orange fragments are not true schemas; rather, they represent extended minima in the schema profile (Fig. 3*a*). This figure was generated using MolMol[40].

Correlation with *in vitro* recombination

The SCHEMA calculation was tested against five experiments in which the genetic information from several parents was recombined to create random libraries of hybrid proteins. In each experiment, a subset of the crossovers survives the screen or selection by retaining (or improving) function. We compared the locations of the functional crossovers with the calculated schema profiles for functional hybrids of cephalosporinases[3,21], subtilisins[22], cytochromes P450 (ref. 23) and glycinamide-ribonucleotide transformylases[24,25] (Fig. 2). Nearly all of the observed crossovers appear in regions corresponding to minima in the schema profiles. The recombination techniques used in these experiments vary significantly, demonstrating the robustness of the predictions.

We find that the window size that best predicts the locations of crossovers in selected libraries is 14, which results in domain sizes of ~20–30 residues. Typically, three types of schema are observed: (i) bundles of α-helices (ii) an α-helix combined with β-strands and (iii) β-strands connected by a hairpin turn. Although the algorithm finds these schemas often, there are numerous interesting exceptions. For example, crossovers are frequently predicted to occur in the center of α-helixes. In addition, there are schema composed of complicated topologies with little discernable secondary structure.

The regions where crossovers are predicted to be deleterious are also noteworthy. For example, crossovers in loops can be highly disruptive if they divide interacting units of secondary structure. A common motif that demonstrates this effect is a single α-helix that is connected by a loop to a β-strand. A single crossover in the loop will disrupt interactions between these secondary structural elements. By the same reasoning, recombining isolated units of secondary structure can be disruptive.

Designing β-lactamase hybrids

Although there is good agreement between the schema profile and the positions of crossovers found during *in vitro* recombination experiments, this agreement does not tell us the degree to which the total amount of schema disruption can be tolerated in a given hybrid. To test this aspect, we recombined two β-lactamases (TEM-1 and PSE-4) that share only 40% amino acid sequence identity but have highly similar structures[21,26,27] and compared the hybrid activities with their calculated disruption. The calculated schema profile of β-lactamase (Fig. 3*a*) was used to divide the structure into schemas (Fig. 3*b*); the degree to which the schemas interact was then calculated (Fig. 4). Based on the calculations, we designed hybrid proteins that have increasing disruption (Fig. 5) but show no correlation with the size of the recombined fragment or with the number of effective mutations corresponding to the recombination event (Table 1). We then constructed this series of hybrid β-lactamases by piecing together DNA fragments of TEM-1 and PSE-4 by PCR[21,28] (see Methods). In addition, we constructed the sequence mirrors of several hybrids. For example, for a two-crossover hybrid (three fragments), we constructed the hybrid in which the first fragment

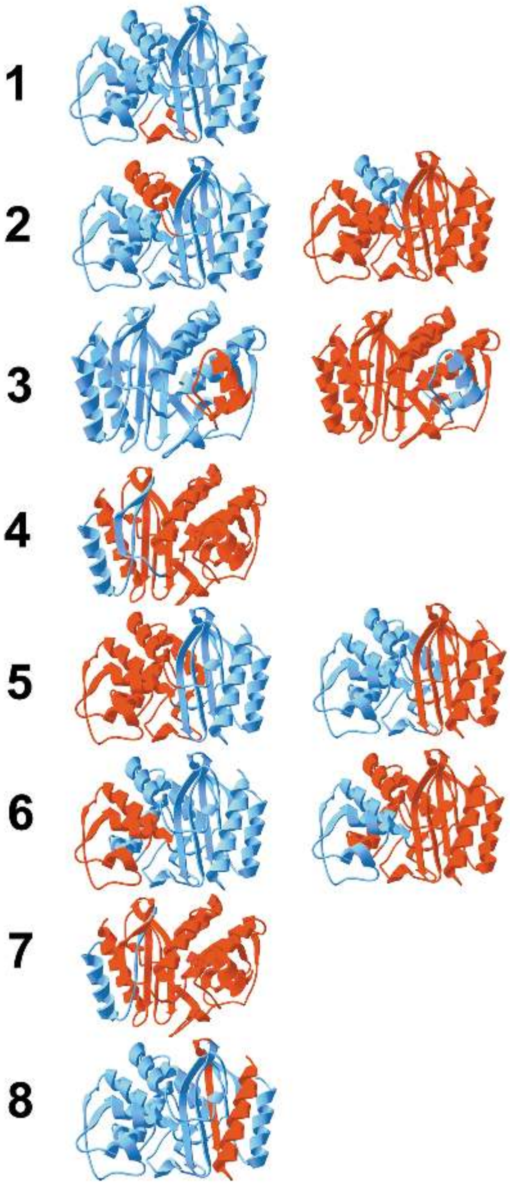

Fig. 5 Designed β-lactamase hybrids. Structures of the designed hybrids of β-lactamase TEM-1 (red) and PSE-4 (blue), shown in order of increasing disruption (Table 1).

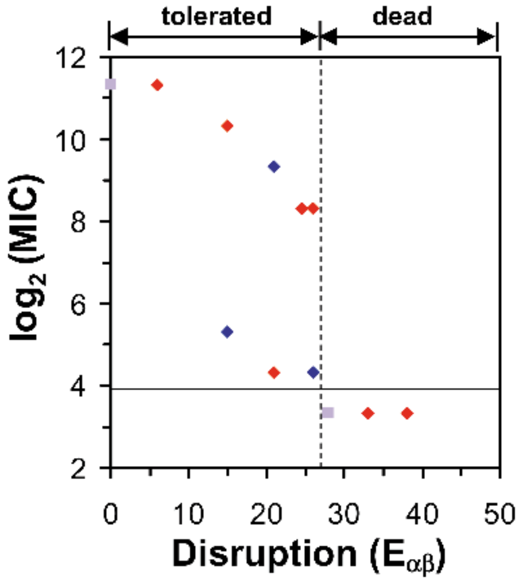

Fig. 6 Activities of hybrid proteins as a function of their disruption. The lower line marks the point at which the MIC represents the background antibiotic resistance of the *Escherichia coli* cells. Activity is lost at $E_{\alpha\beta} \approx 27$. Below the transition, the recombination events are nondisruptive. Above the transition, the hybrids are nonfunctional. The region just before the transition may be optimal for creating hybrid protein libraries: here diversity is maximized while structure disturbance is minimized. The color of the points indicates the parent of the first fragment: red is PSE-4 ('A'), blue is TEM-1 ('B') and purple indicates that the red and blue points overlap.

is from PSE-4 (labeled 'A'), as well as that having the first fragment from TEM-1 (labeled 'B').

We tested each hybrid protein for activity by measuring the minimum concentration of ampicillin required to inhibit cell growth, the minimum inhibitory concentration (MIC) (see Methods). Wild type TEM-1 and PSE-4 are highly active towards ampicillin (MIC ~ 2,560 μg ml^{-1}) and have similar activities towards various β-lactam substrates[21]. The MIC value is a complex combination of effects, including expression, stability and activity[29,30]. Here, we use the antibiotic resistance merely to determine whether a hybrid β-lactamase is folded and functional and not to rank the individual enzyme activities. By measuring the MIC conferred by each hybrid, we found a sharp transition in disruption, beyond which hybrids are nonfunctional (Fig. 6). This transition does not correlate with the number of mutations that effectively occur when the hybrid is constructed (Table 1). The transition divides the graph into two regions: 'tolerated' (nondisruptive) and 'dead' (highly disruptive). The region just before the transition may be the optimal level of disruption to target in creating libraries of hybrids. In this way, diversity is maximized while the fraction of the library that is nonfunctional or unfolded is minimized.

The eight hybrids that show activity (1A–5B) have interesting characteristics. Many have at least one crossover at a buried position. Additionally, a crossover occurs in the middle of a helix for two hybrids (2A and 2B) and at the end of a β-strand in hybrid 1A. Finally, six of the hybrids (2A, 2B, 3A, 3B, 5A and 5B) have crossovers near the active site. Notably, several hybrids that were nonfunctional (7A and 8A) have crossovers that occur in a loop on the surface, with only a few residues recombined at the termini. Crossovers in loops are often considered to be nondisruptive, yet our algorithm correctly identified a crossover as strongly disruptive in this context. Finally, we constructed two hybrids (4A and 7A) that differ only by 12 residues near the N-terminus. Hybrid 4A was found to be functional, whereas hybrid 7A was not. This distinction would be hard to predict based on visualizing the differences mapped onto the three-dimensional structure alone.

The natural selection of intron locations

Go[8,13] discovered a correlation between the location of introns and isolated geometrical domains, a correlation that has held for a wide range of proteins. This correlation has been interpreted as evidence for the 'introns-early' theory of evolution, which states that the first large proteins were constructed from smaller domains through recombination and gene duplication[14,15]. The merging of genes resulted in the separation of the coding DNA by regions of noncoding DNA (introns). Over evolutionary time, the introns disappeared where they were no longer necessary or were disadvantageous — for example, in the restricted genome sizes of prokaryotes. Proponents of this theory have argued that if introns appeared late in evolution, their locations would appear random with respect to structural domains[14,15]. Our results indicate that the correlation between introns and domains could occur as a result of natural selection, even if the introns appeared late.

Of the many proposed functions of introns, one is that they facilitate the swapping of exons[14,15]. If the probability of a crossover is equal across the gene, then a long region of noncoding DNA will bias the crossovers toward a specific region of the fully spliced gene. Cycles of recombination and selection can bias the location of introns if the ability of an intron to promote shuffling contributes to the fitness of an organism. If, in a population of these organisms, introns were randomly distributed throughout the gene, there would be a selective advantage to

Table 1 Designed TEM-1–PSE-4 hybrid β-lactamases

	Crossover 1		Crossover 2[1]				
Hybrid[2]	Number	Context[3]	Number	Context[3]	m[4]	$E_{\alpha\beta}$	MIC
1A[5]	163	loop, surface	179	strand, core	7	6	2,560[6]
2A	189	helix, core	216	loop, surface, as	18	15	1,280
2B	189	helix, core	216	loop, surface, as	18	15	40
3A	130	loop, core, as	163	loop, surface	13	21	20
3B	130	loop, core, as	163	loop, surface	13	21	320
4A	65	loop, surface			42	25	320
5A	70	loop, core, as	216	loop, surface, as	83	26	320
5B	70	loop, core, as	216	loop, surface, as	83	26	20
6A	70	loop, core, as	130	loop, core, as	41	27	10[7]
6B	70	loop, core, as	130	loop, core, as	41	27	10[7]
7A	53	loop, surface			42	33	10[7]
8A	254	loop, surface			23	37	10[7]

[1] This portion is left blank if the hybrid protein only has a single crossover.
[2] The letter in the name indicates the parent that composes the first portion of the gene, where A is PSE-4 and B is TEM-1. For the double crossover mutants, an 'A' indicates a gene structure of A-B-A and 'B' indicates B-A-B.
[3] The context of the side chain of the residue where the cut occurs. The notation 'as' indicates that the crossover occurs near the active site.
[4] Number of mutations that occur when the smaller fragment of one parent is inserted into the larger context of the remaining parent.
[5] This hybrid has been constructed by Levesque and coworkers[21].
[6] Wild type activity of both PSE-4 and TEM-1.
[7] The MIC of XL1-BLUE cells. No β-lactamase activity is observed.

those individuals whose introns appeared in regions that are the most likely to result in successful shuffling events. We have observed this directly in *in vitro* recombination experiments. When crossovers are randomly distributed throughout the gene, the subset that preserve the schema are also the most likely to result in folded, functional hybrids. Therefore, if introns promote recombination, they will most likely reside in low-disruption regions after selection.

Conclusions

Crossovers that lead to folded, functional hybrid proteins occur at positions that minimize the number of disrupted interactions. A simple model of interacting residues can capture this and correctly predict acceptable crossover locations. An important application of this approach will be to accelerate molecular optimization by laboratory evolution methods through the use of computational tools[31–34]. Combinatorial libraries with targeted crossovers can dramatically improve an evolutionary search by significantly reducing the number of mutants that must be screened to obtain specific functional changes. The elucidation and experimental verification of evolutionary dynamics will allow the design of a new generation of evolutionary methods that maximize our ability to discover new biological molecules.

Methods

Calculating the schema profile. The schema disruption of a hybrid protein is the number of interactions that are broken when a certain pattern of fragments is inherited from each of the parents. If a hybrid protein is constructed from two parents where fragment(s) α is/are inherited from parent A and fragment(s) β is/are inherited from parent B, then the disruption, $E_{\alpha\beta}$, of this hybrid can be calculated by

$$E_{\alpha\beta} = \sum_{i\in\alpha}\sum_{j\in\beta} c_{ij} P_{ij} \qquad (1)$$

where c_{ij} = 1 if residues i and j are within distance d_c; otherwise, c_{ij} = 0. Eq. 1 calculates the exact disruption caused by a particular hybrid construction (Table 1; Fig. 6).

The probabilities P_{ij} account for the fact that there is no disruption if the amino acid identities of the residue pair i,j in the set of potential hybrids are the same as in any of the parents. An alignment of the parental amino acid sequences is used to calculate the likelihood that novel combinations of amino acids will be inherited for a given residue pair. To determine the probability, the number of novel combinations is divided by the total number of combinations, p(p – 1), where p is the number of parents.

Eq. 1 can be used to calculate the disruption of any particular hybrid construction. However, when analyzing data from *in vitro* recombination experiments, the number of possible hybrid combinations prohibits the calculation of the disruption of all possible hybrids and the condensation of this information into a useful format. To compare recombination results with the schema disruption theory, we have developed an algorithm that searches for the most likely regions for crossovers to be nondisruptive. The inputs into the SCHEMA program are the coordinates of the three-dimensional structure and an alignment of the parental sequences. The structure of only one parent is required under the assumption that the parents must share similar structures for *in vitro* recombination to be successful. A window of residues w is defined, and the number of internal interactions within this window is counted. In choosing the window size, the assumption is made that the probability that two or more crossovers occurring in the window is small. The window is then slid along the protein structure and a profile is generated where the schema profile of each residue in the window is incremented by the amount of disruption created by a crossover in that region. The numerical value of the schema profile function S at residue i is defined by

$$S_i = \sum_{j=i-w+1}^{i}\ \sum_{k=j}^{j+w-2}\ \sum_{l=k+1}^{j+w-1} c_{kl} P_{kl} \qquad (2)$$

If a residue has a large S_i, then it probably participates in a compact schema. A low S_i indicates that a crossover is probably tolerated at that position. For all of the calculations presented in this manuscript, the parameters are d_c = 4.5 Å and w = 14 residues. The qualitative features of the profiles, such as the location of the minima, are relatively insensitive to the specific values of these parameters (data not shown).

Sequence alignments. Sequence alignments were performed using the BLAST algorithm with the BLOSUM 62 (ref. 35) similarity matrix and open gap / extension gap penalties of 11 / 1. In general, the sequence identity between the parents is >60%, reducing the ambiguity of the alignment. For the β-lactamase TEM-1/PSE-4 system (40% identity), the availability of both structures made a structural alignment possible (using the SwissProt software package: *http://www.expasy.ch/sprot/sprot-top.html*).

Recombination data sets. For cephalosporinase, the schema profile was calculated based on the structure of cephalosporinase[36]. The crossovers that led to improved moxalactam antibiotic resistance were located in regions of low schema disruption[3] (Fig. 2*a*). Further, an independent experiment was performed by Levesque and co-workers[21] in which a fragment was taken from the β-lactamase TEM-1 gene and inserted into the PSE-4 gene. The resulting hybrid protein was found to have wild type activity towards various antibiotics. When this fragment is mapped onto the cephalosporinase structure, it corresponds to a low-disruption region.

For the subtilisin families, Minshull and co-workers[22] recombined a set of 26 subtilisin genes by DNA shuffling and screened the recombinant mutants for improved thermostability, high and low pH activity, and activity in organic solvent. When aligned, the 26 genes fall into four well-defined families. Within each family, the genes have ~99% sequence identity. Crossovers between parents that have this high degree of sequence identity are impossible to analyze by schema disruption. However, the sequence identity between parents from different families ranged from 10 to 24%. It is possible, then, to compare the crossovers between families with the schema profile. In the experiments, crossovers were allowed in the region between residues 60 and 224. The remaining portions of the sequence (1–60 and 224–269) were taken from the Savinase gene. The structure of Savinase was used to calculate the schema profile[37] (Fig. 2*b*). Nearly all of the sequences of the 26 parental genes are unavailable. To overcome this, we ran a BLAST search and selected a *Bacillus halodurus* serine protease (SwissProt entry P41363), which is 65% identical to the Savinase sequence. The probabilities required by Eq. 1 were estimated based on an alignment of these two sequences.

For cytochromes P450, a recombination experiment was performed on two P450c17 genes (rat and human) sharing 68% sequence identity, and a variety of functional hybrid proteins were discovered[23]. The structure of c17 is unknown; however, a structure of a homologous mammalian membrane-bound P450 2C5 has been solved[38]. The equivalent locations for the crossovers were determined by aligning the parental sequences used in the experiment with the 2C5 sequence (Fig. 2*c*).

For glycinamide ribonucleotide transformylase, Benkovic and co-workers[24,25] recombined PurN and GART glycinamide ribonucleotide transformylase and selected functional hybrid proteins. In this experiment, recombination was restricted to occur between amino acid positions 50 and 150. The schema profile was calculated from the structure of PurN[39] (Fig. 2*d*).

Hybrid gene construction. The oligonucleotide fragments corresponding to the peptide schemas were made *via* PCR amplification, where the primers at either end contain a short piece of DNA that overlaps with the preceding gene fragment[28]. This overlap ensures that the fragments will re-anneal to produce a full-length gene. The promoter for PSE-4 in the PMON vector[21] was used for the 'A' fragments, and the promoter for TEM-1 in the PSTBlue-1 vector (Novagen) was used for the 'B' fragments. The PCR protocol is to initially heat the template vectors and primers at 95 °C and then perform 25 cycles of heating at 94 °C for 45 s, cooling at 52 °C for 45 s and extending at 72 °C for 1 min. The fragment is then gel purified and concentrated either through ethanol precipitation (for fragments < 100 bp) or using a Zymoclean-5 gel extraction kit (Zymo Research) (for fragments > 100 bp). Once the oligonucleotide fragments are isolated, they are re-annealed to create a complete gene fragment through a second PCR amplification step. The forward and reverse primers have the sequences for the restriction sites of *Eco*RI and *Hin*dIII, respectively, so that the complete genes can be inserted into the PMON vector modified to contain these restriction sites. The times and temperatures are identical to the previous amplification round. A pre-PCR step can be used to improve the purity of the amplified genes. This PCR protocol is 25 iterations of 95 °C for 30 s, 5 °C for 30 s and 72 °C for 2 min. A final extension of 10 min at 72 °C is done after the cycles are complete. The fragments are purified using the Zymoclean-5 gel extraction kit. Finally, the fragments are ligated into the PMON vector, which has kanamycin resistance. The vectors containing the hybrid genes are transformed into XL1-BLUE supercompetent ($>10^9$) cells (Stratagene) and grown on plates that contain 10 μg ml^{-1} kanamycin. Colonies are isolated and the vector is extracted and sequenced. Some of the recombinant genes contained point mutations after the construction process (~0.06% nucleotide changes per gene).

MIC screening. Each hybrid β-lactamase is tested for its activity towards the degradation of the antibiotic ampicillin. To rapidly screen for this property, agar plates are made with following exponentially increasing concentrations of ampicillin: 10, 20, 40, 80, 160, 320, 640 and 1,280 μg ml^{-1}. Aliquots of transformed cells are spread on the plates and allowed to grow at 37 °C for 24 h. More active hybrids will grow on plates with greater concentrations of ampicillin. The activity is measured as the MIC — that is, the lowest concentration of ampicillin that kills the cells. The XL1-BLUE cells naturally have a MIC of 10; thus, β-lactamase activity cannot be measured below this point. The wild type TEM-1 and PSE-4 enzymes have MICs of 2,560.

Acknowledgments

C.A.V. is supported by a National Science Foundation graduate research fellowship and the California Institute of Technology Initiative in Computational Molecular Biology, a Burroughs Wellcome funded program for science at the interface. Z.G.W. acknowledges the support by the W.M. Keck Foundation. S.L.M. is supported by the Howard Hughes Medical Institute, the Ralph M. Parsons Foundation and an IBM Shared University Research Grant. The PSE-4 gene and the PMON vector were provided by R.C. Levesque (Université Laval, Québec, Canada).

Competing interests statement

The authors declare that they have no competing financial interests.

Received 21 December, 2001; accepted 23 April, 2002.

1. Holland, J. *Adaptation in Natural and Artificial Systems* (The University of Michigan Press, Ann Arbor; 1975).
2. Stemmer, W.P.C. Rapid evolution of a protein *in-vitro* by DNA shuffling. *Nature* **370**, 389–391 (1994).
3. Crameri, A., Raillard, S-A., Bermudez, E. & Stemmer, W.P.C. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288–291 (1998).
4. Ostermeier, M. & Benkovic, S.J. Evolution of protein function by domain swapping. *Adv. Protein Chem.* **55**, 29–77 (2000).
5. Rossman, M.G. & Liljas, A. Recognition of structural domains in globular proteins. *J. Mol. Biol.* **85**, 177–181 (1974).
6. Crippen, G.M. Tree structural organization of proteins. *J. Mol. Biol.* **126**, 315–332 (1978).
7. Rose, G.D. Hierarchic organization of domains in globular-proteins *J. Mol. Biol.* **134**, 447–470 (1979).
8. Go, M. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**, 90–92 (1981).
9. Zehfus, M.H. & Rose, G.D. Compact domains in proteins. *Biochemistry* **25**, 5759–5765 (1986).
10. Holm, L. & Sander, C. Parser for protein folding units. *Proteins* **19**, 256–268 (1994).
11. Panchenko, A.R., Luthey-Schulten, Z. & Wolynes, P.G. Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci. USA* **93**, 2008–2013 (1996).
12. Tsai, C.-J., Maizel, J.V. & Nussinov, R. Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc. Natl. Acad. Sci. USA* **97**, 12038–12043 (2000).
13. Go, M. Modular structural units, exons, and function in chicken lysozyme. *Proc. Natl. Acad. Sci. USA* **80**, 1964–1968 (1983).
14. de Souza, S.J., Long, M., Schoenbach, L., Roy, S.W. & Gilbert, W. Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl. Acad. Sci. USA* **93**, 14632–14636 (1996).
15. Gilbert, W., de Souza, S.J. & Long, M.Y. Origin of genes. *Proc. Natl. Acad. Sci. USA* **94**, 7698–7703 (1997).
16. Ranganathan, A. *et al.* Knowledge-based design of bimodular and trimodular polyketide synthases based on domain and module swaps: a route to simple statin analogues. *Chem. Biol.* **6**, 731–741 (1999).
17. Bogarad, L.D. & Deem, M.W. A hierarchal approach to protein molecular evolution. *Proc. Natl. Acad. Sci. USA* **96**, 2591–2595 (1999).
18. Riechmann, L. & Winter, G. Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc. Natl. Acad. Sci. USA* **97**, 10068–10073 (2000).
19. Forrest, S. & Mitchell, M. *Foundations of Genetic Algorithms 2* (ed. Whitley, L.D.) 109 (Morgan Kaufmann, San Mateo; 1993).
20. Mitchell, M. *An Introduction to Genetic Algorithms* (The MIT Press, Cambridge, Massachusetts; 1996).
21. Sanschagrin, F., Theriault, E., Sabbagh, Y., Voyer, N. & Levesque, R.C. Combinatorial biochemistry and shuffling TEM, SHV and *Streptomyces albus* omega loops in PSE-4 class A β-lactamase. *J. Antimicrob. Chemo.* **45**, 517–519 (2000).
22. Ness, J.E. *et al.* DNA shuffling of subgenomic sequences of subtilisin. *Nature Biotech.* **17**, 893–896 (1999).
23. Brock, B.J. & Waterman, M.R. The use of random chimeragenesis to study structure/function properties of rat and human P450c17. *Arch. Biochem. Biophys.* **373**, 401–408 (2000).
24. Ostermeier, M., Shim, J.H. & Benkovic, S.J. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotech.* **17**, 1205–1209 (1999).
25. Lutz, S., Ostermeier, M. & Benkovic, S.J. Rapid generation of incremental truncation libraries for protein engineering using α-phosphothioate nucleotides. *Nucleic Acids Res.* 29, e16 (2001).
26. Jelsch, C., Mourey, L., Masson, J.M. & Samama, J.P. Crystal-structure of *Escherichia coli* TEM-1 β-lactamase at 1.8-Å resolution. *Proteins* **16**, 364–383 (1993).
27. Lim, D. *et al.* Insights into the molecular basis for carbenicillinase activity of PSE-4 β-lactamase from crystallographic and kinetic studies. *Biochemistry* **40**, 395–402 (2001).
28. Horton, R.M. PCR-mediated recombination and mutagenesis. *Mol. Biotech.* **3**, 93–99 (1995).
29. Palzkill, T. & Botstein, D. Probing β-lactamase structure and function using random replacement mutagenesis. *Proteins* **14**, 19–44 (1992).
30. Huang, W.Z., Petrosino, J., Hirsch, M., Shenkin, P.S. & Palzkill, T. Amino acid sequence determinants of β-lactamase structure and activity. *J. Mol. Biol.* **258**, 688–703 (1996).
31. Voigt, C.A., Mayo, S.L., Arnold, F.H. & Wang, Z.-G. Computational method to reduce the search space for directed protein evolution. *Proc. Natl. Acad. Sci. USA* **98**, 3778–3783 (2001).
32. Voigt, C.A., Kauffman, S. & Wang, Z.-G. Rational evolutionary design: the theory of *in vitro* protein evolution. *Adv. Protein Chem.* **55**, 79–160 (2000).
33. Voigt, C.A., Mayo, S.L., Arnold, F.H., & Wang, Z.-G., Computationally focusing the directed evolution of proteins. *J. Cell. Biochem.* **Suppl. 37**, 58–63 (2001).
34. Bolon, D.N., Voigt, C.A. & Mayo, S.L. *De novo* design of biocatalysts. *Curr. Opin. Chem. Biol.* **6**, 125–129 (2002).
35. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).
36. Lobkovsky, E. *et al.* Evolution of an enzyme-activity — crystallographic structure at 2-Å resolution of cephalosporinase from the AmpC gene of *Enterobacter cloacae*-P99 and comparison with a class-A penicillinase. *Proc. Natl. Acad. USA* **90**, 11257–11261 (1993).
37. Betzel, C. *et al.* Crystal-structure of the alkaline proteinase savinase from *Bacillus lentus* at 1.4-Å resolution. *J. Mol. Biol.* **223**, 427–445 (1992).
38. Williams, P.A., Cosme, J., Sridhar, V., Johnson, E.F. & Mcree, D.E. Mammalian microsomal cytochrome P450 monooxygenase: structural adaptations for membrane binding and functional diversity. *Mol. Cell.* **93**, 121–131 (2000).
39. Almassy, R.J., Janson, C.A., Kan, C.C. & Hostomska, Z. Structures of apo and complexed *Escherichia coli* glycinamide ribonucleotide transformylase. *Proc. Natl. Acad. Sci. USA* **89**, 6114–6118 (1992).
40. Koradi, R., Billeter, M. & Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14**, 51–55 (1996).