doi:10.1016/S0022-2836(03)00590-4 J. Mol. Biol. (2003) 330, 287–296

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

General Method for Sequence-independent Site-directed Chimeragenesis

Kaori Hiraga and Frances H. Arnold*

Division of Chemistry and Chemical Engineering California Institute of Technology, Pasadena, CA 91125, USA

*Corresponding author

We have developed a simple and general method that allows for the facile recombination of distantly related (or unrelated) proteins at multiple discrete sites. To evaluate the sequence-independent site-directed chimeragenesis (SISDC) method, we have recombined β-lactamases TEM-1 and PSE-4 at seven sites, examined the quality of the chimeric genes created, and screened the library of 2^8 (256) chimeras for functional enzymes. Probe hybridization and sequencing analyses revealed that SISDC generated a random library with little sequence bias and in which all targeted fragments were recombined in the desired order. Sequencing the genes from clones having functional lactamases identified 14 unique chimeras. These chimeras are characterized by a lower level of disruption, as calculated by the SCHEMA algorithm, than the library as a whole. These results illustrate the use of SISDC in creating designed chimeric protein libraries and further illustrate the ability of SCHEMA to identify chimeras whose folded structures are likely not to be disrupted by recombination.

Keywords: DNA shuffling; non-homologous recombination; directed evolution; SCHEMA

Introduction

Recombination *via* shuffling of genes is a powerful tool for the laboratory evolution of proteins.[1,2] In this process, sequence fragments from related genes are exchanged to produce a combinatorial library of recombined mutants; screening or selection identifies the sequences that produce proteins with altered or improved functions.[3–8] Most methods for producing such recombined, or chimeric, gene libraries are based on DNA shuffling, in which small DNA segments are assembled or hybridized using local homologous sequences. Thus, to create diverse libraries with multiple crossovers, the parental genes must share high levels of sequence identity (usually >70%).[9] Several methods have been reported for creating chimeric protein libraries independent of homology, including ITCHY,[10] SCRATCHY[11] and SHIPREC.[12] These methods, however, generate large numbers of non-functional sequences, due to mutations, insertions and deletions, and only a few crossovers at best. Their ability to create diverse libraries of functional proteins has not been demonstrated convincingly.

Studies of protein evolution and computational approaches have identified various smaller, structural "building blocks" from which proteins have been (and could be) assembled. Appropriately chosen domains, motifs, modules or exons[13–17] should be useful elements for generating libraries rich in functional proteins for structure–function studies or for discovering new proteins by laboratory evolution.[18–21] Constructing such libraries of chimeric sequences, however, requires a method in which the designated modules from different sources can be shuffled, independent of their sequence identity and maintaining a desired module order. Several different methods could conceivably be used for this purpose. For example, site-directed recombination mediated by PCR using chimeric oligonucleotide primers,[21,22] gene splicing by overlap extension PCR (SOEing)[23] and random multi-recombinant PCR (RM-PCR)[19] permit the shuffling of DNA fragments without homologous sequences. However, creating a library of highly mosaic chimeras from multiple parents would require a large number of oligonucleotide primers, and designing appropriate primers would be complicated. DNA fragments

Abbreviations used: SISDC, sequence-independent site-directed chimeragenesis; SOEing, splicing by overlap extension PCR; RM-PCR, random multi-recombinant PCR; MIC, the minimum inhibitory concentration.

E-mail address of the corresponding author: frances@cheme.caltech.edu

flanked by restriction enzyme sites can be used to construct chimeric protein libraries.[24,25] However, the ligation sites specify the amino acids at every junction between fragments, which strongly biases the protein sequences that can be explored.

Here, we present a simple and general method for site-directed recombination which allows crossovers at multiple sites and is independent of DNA sequence identity shared by the parent genes. In the method described here, the ligation of building blocks will easily regenerate a full-length and highly recombined library through the combinatorial assembly of DNA fragments from different parents. All fragments can be connected in the right order, without any insertions at crossover points. This method can be used for recombination of multiple parent sequences, or elements from multiple parents.

To evaluate this sequence-independent site-directed chimeragenesis (SISDC) method, we have used it to recombine two distantly related β-lactamases, TEM-1 and PSE-4 (40% amino acid identity and 49% nucleotide sequence identity),[26,27] at seven sites to generate a library of 2^8 (256) chimeric sequences. Crossover sites were identified using the computational algorithm SCHEMA, which uses structural information to predict polypeptide elements that can be swapped among related proteins with minimal disruption.[20] In addition, we screened our library for lactamase function (antibiotic resistance) and examined whether the sequences of functional chimeras were those predicted by SCHEMA to have low disruption.

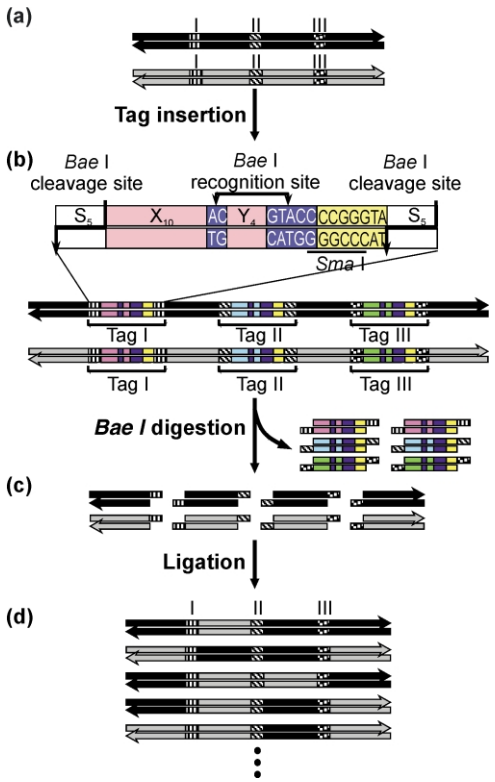

Figure 1. A schematic of sequence-independent site-directed chimeragenesis (SISDC), illustrating recombination between two parents at three sites (four elements). (a) The nucleotide sequences of the parent genes (A and B) are aligned, and consensus sequences at each targeted site (I, II and III) are determined. (b) Marker tags (I–III) are inserted into targeted sites (I–III). The marker tags contain consensus 5 bp sequence (S_5), a type IIb endonuclease (*Bae* I) recognition sequence (blue), two variable regions (X_{10} and Y_4: pink for site I, cyan for site II and green for site III), a constant region (yellow) and a *Sma* I recognition site (underline). (c) Inserted tag regions are removed by treatment with endonuclease and specific sticky ends are produced. (d) Mixed fragments can identify their neighbors, and the fragments from different parents ligate with each other in the right order. After this step, digestion by *Sma* I can eliminate any untreated tag remaining in the final, chimeric library.

Results and Discussion

The SISDC method

A schematic of the SISDC method is shown in Figure 1. The method can be used to generate a chimeric library from multiple parents at multiple crossover sites. For purposes of illustration, Figure 1 demonstrates SISDC for four sequence elements from two parents. Once crossover sites are determined in the aligned parent nucleotide sequences (Figure 1(a)), marker tags containing a type IIb endonuclease recognition sequence (*Bae* I) are inserted into the targeted sites in the parent genes (Figure 1(b)). *Bae* I digests double-stranded DNA specifically on both sides of its recognition sequence (10/15)ACNNNNGTAYC(12/7), leaving two 3′ overhangs.[28] The upstream (10 bp), downstream (7 bp) and middle (4 bp) of the *Bae* I recognition sites are customizable, as are the two cleavage sites (5 bp). In our design of the marker tags (Table 1), the upstream and middle regions of the recognition sites (X_{10} and Y_4) have a different sequence for each targeted position. These regions are used for assembly PCR to construct parental genes (described below). The two 5 bp cleavage sites (S_5) correspond to the consensus sequence at each junction and are reserved for directing the correct ligation of the targeted fragments with no insertions. In addition, the downstream sequence of the *Bae* I recognition sites contains a recognition site for *Sma* I, which is used to eliminate untreated tags remaining in the final chimeric library. After the inserted tag regions are removed by treatment with *Bae* I, a series of specific sticky ends (S_5) are produced (Figure 1(c)). Finally, mixed fragments identify their neighbors and ligate with each other in the right order to yield a combinatorial library of chimeras (Figure 1(d)). In the current study, we constructed a chimeric library in which eight sequence elements from two parents were shuffled. To accomplish this, two half-length libraries were prepared separately according to Figure 1; the two

Table 1. Design of primers for sequence-independent site-directed chimeragenesis (SISDC) of PSE-4 and TEM-1 β-lactamases

Tag	Site	Marker tag sequence (5′ to 3′)[a] S_5-X_{10}-**AC**-Y_4-**GTAC**<u>C</u>-CCGGGTA-S_5	Primer (5′ to 3′)[b]
N	N-term	CAAGCTT-*GGTACC*-C	PNasp:CAAGCTTGGTACCCatgcttttatataaaatgtgtgacaa TNasp:CAAGCTTGGTACCCatgagtattcaacatttccgtgtc
Tag1	41	GCACG-ATGACTTAAG-**AC**-TGAC- **GTAC**<u>C</u>-CCGGGTA-GCACG	P1N:CTTAAGACTGACGTACCCCGGGTAGCAcgtataggtgtttccgttcttg P1C:CGGGGTACGTCAGTCTTAAGTCATCGTgcagaaagagaaacttcaattgc T1N:CTTAAGACTGACGTACCCCGGGTAgcacgagtgggttacatcgaa T1C:CGGGGTACGTCAGTCTTAAGTCATcgtgcacccaactgatcttca
Tag2	70	AGCAC-CTGAATCGAT-**AC**-TCAC- **GTAC**<u>C</u>-CCGGGTA-AGCAC	P2N:ATCGATACTCCAGTACCCCGGGAAAGCacttttaaaacaatagcttgcgc P2C:CGGGGTACTGGAGTATCGATTCAGGTGcttgttaacgggaagcgctga T2N:ATCGATACTCCAGTACCCCGGGAAagcacttttaaagttctgctatgt T2C:CGGGGTACTGGAGTATCGATTCAGgtgctcatcattggaaaacgtt
Tag3	106	TCACC-GTGACGTAGC-**AC**-GCTA- **GTAC**<u>C</u>-CCGGGTA-TCACC	P3N:GCTAGCACGCATGTACCCCGGGTTTCAcctgtaatagaaaagcaagtagg P3C:CGGGGTACATGCGTGCTAGCTCACGGTgaataggtcacaagatctgcttt T3N:GCTAGCACGCATGTACCCCGGGTTtcaccagtcacagaaaagcatc T3C:CGGGGTACATGCGTGCTAGCTCACggtgagtactcaaccaagtcat
Tag4	133	ACTGC-CTAAACGCGT-**AC**-GACT- **GTAC**<u>C</u>-CCGGGTA-ACTGC	P4N:ACGCGTACGACTGTACCCCGGGATactgcggcaaatatcatcctaa P4C:CGGGGTACAGTCGTACGCGTTTAGgcagtattatcacttgtagtcata T4N:ACGCGTACGACTGTACCCCGGGATactgcggccaacttacttctg T4C:CGGGGTACAGTCGTACGCGTTTAGgcagtgttatcactcatggttat
Tag5	166	GAACC-GTAACCGCGG-**AC**-TATT- **GTAC**<u>C</u>-CCGGGTA-GAACC	P5N:CCGCGGACTATTGTACCCCGGGACGAAcctgatttaaatgaaggtaagct P5C:CGGGGTACAATAGTCCGCGGTTACGGTtcaatacggtctagacgagtct T5N:CCGCGGACTATTGTACCCCGGGACgaaccggagctgaatgaagc T5C:CGGGGTACAATAGTCCGCGGTTACggttcccaacgatcaaggcg
Tag6	227	GCTGG-TTAATCGCGA-**AC**-ATCC- **GTAC**<u>C</u>-CCGGGTA-GCTGG	P6N:TCGCGAACATCCGTACCCCGGGAGGCTggatggaacattgcggatcg P6C:CGGGGTACGGATGTTCGCGATTAACCAgccggcaatactgaacgtagta T6N:TCGCGAACATCCGTACCCCGGGAGgctggctggtttattgctgata T6C:CGGGGTACGGATGTTCGCGATTAAccagccggaagggccgag
Tag7	267	CAGGC-ATAATCCGGA-**AC**-ATTA- **GTAC**<u>C</u>-CCGGGTA-CAGGC	P7N:TCCGGAACATTAGTACCCCGGGTGcaggcttcaatggaagagcga P7C:CGGGGTACTAATGTTCCGGATTATgcctgtgtttgagctagataga T7N:TCCGGAACATTAGTACCCCGGGTGcaggcaactatggatgaacgaa T7C:CGGGGTACTAATGTTCCGGATTATgcctg-actccccgtcgtgt
C	C-term	*CCTGCAGG*-TTG	PCsse:CAACCTGCAGGtcagcgcgactgtgatgtataa TCsse:CAACCTGCAGGttaccaatgcttaatcagtgagg

[a] Bold letters, *Bae*I recognition site; underline, *Sma*I recognition site. S_5, consensus sequence; sticky ends by *Bae*I digestion. X_{10} and Y_4, variable sequences for assembly PCR for parental gene construction. Italics, *Asp718* or *Sse*8387I recognition site of N-terminal or C-terminal primer, respectively.
[b] Lower-case letters, complementary sequence to a parent for SCHEMA gene fragment amplification.

libraries were then ligated to yield a full-length library, as described below.

SCHEMA library construction

The computational algorithm SCHEMA was developed in this laboratory to identify structural elements that can be swapped among related proteins with minimal disruption.[20] Using the three-dimensional structure to generate a contact matrix of interactions, SCHEMA calculates the number of interactions, E_s, that are disrupted when specified elements are replaced by their counterparts from another parent protein. To identify favorable crossover locations, one version of SCHEMA scans the protein sequence with a window of defined size to create a disruption profile whose peaks represent contiguous polypeptides with the largest number of internal interactions. Figure 2 shows the SCHEMA profiles calculated from the TEM-1 and PSE-4 β-lactamase structures. The profiles both contain eight major peaks. Crossovers in the regions separating these peaks are expected to yield low levels of structural disruption in many of the resulting recombined proteins. The chimeric library (TP library) was constructed by recombining gene fragments encoding these eight SCHEMA fragments of TEM-1 or PSE-4.

The parental half-length genes, in which a series of the marker tags were inserted at the SCHEMA boundaries, were constructed as shown in Figure 3. Targeted fragments were amplified using synthetic oligonucleotides that contained a marker tag, which includes a recognition sequence of the type IIb restriction enzyme, *Bae*I, and a *Sma*I recognition sequence (Figure 3(a)). Mixtures of these PCR fragments were then assembled into half-length genes with tag insertions by a self-priming overlap PCR (Figure 3(b)–(d)). The X_{10} and Y_4 regions in the marker tag sequences (Table 1) annealed to each other successfully in the assembly PCR, and all of the expected fragments

were amplified in each step. The parental half-length genes were cloned into the plasmid, and their DNA sequences were verified. All of the parental genes were inserted inversely into the cloning site of a TOPO plasmid to inhibit expression of the foreign protein. This helps to maintain full library diversity during library construction.

The parental genes were amplified again by PCR and subjected to digestion by *Bae* I. After the small marker tags (38 bp) were removed by column purification, the DNA fragments were ligated *via* the 5 bp cohesive ends of the consensus sequences, resulting in the half-length products without tag sequences. Figure 4 shows the agarose electrophoresis gel containing the fragments after ligation and PCR. When the ligation product was amplified without treatment with *Sma* I, the bands were slightly smeared, presumably due to incomplete digestion of original parent genes containing inserted tags (see Figure 1(b)). However, after treatment with *Sma* I the main products had the expected sizes. PCR products were cloned back into a TOPO vector and transformed into TOP10 cells. The size of the half-length TP library was estimated to be larger than 300 (see Materials and Methods), well exceeding the 16 possible combinations. Then, the half-length libraries were digested from a TOPO vector and equal amounts were mixed and cloned into an expression vector, pPROtet. Library plasmid DNA was prepared by transformation of ligation reactions into DH5α-pro cells, followed by growth in LB medium containing chloramphenicol. The size of the full-length TP library was estimated to be larger than 5000, well exceeding the 256 possible combinations. In this expression system, the expression of recombinant protein is inhibited tightly in the absence of anhydrotetracycline, because the tetracycline-regulated promoter, P_{Ltet-1}, of pPROtet is repressed tightly by the Tet repressor expressed constitutively in DH5α-pro. In this way, biases in the pre-screened library caused by protein expression are avoided.

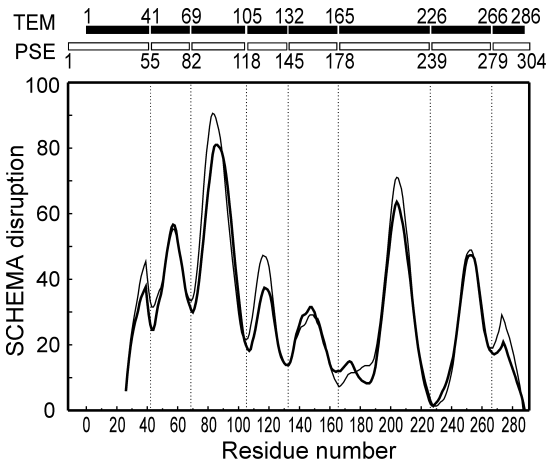

Figure 2. SCHEMA profiles of β-lactamases TEM-1 (bold line) and PSE-4 (thin line). Residues are numbered, based on the sequence of TEM-1, and vertical dotted lines show the crossover positions selected for this study. The resulting SCHEMA fragments are shown at the top, with the corresponding residue numbers (TEM-1, filled bar; PSE-4, open bar).

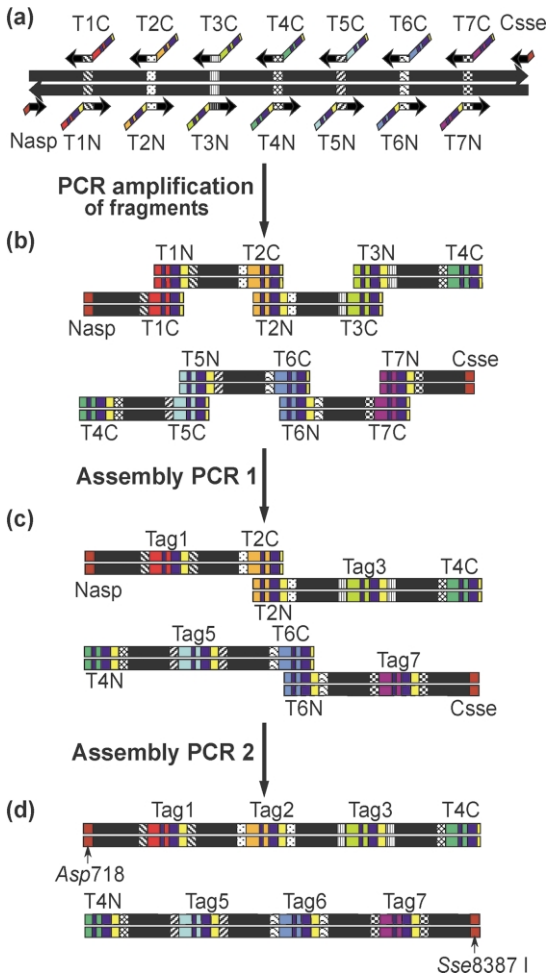

Figure 3. Construction of the parental half-length gene of TEM-1 β-lactamase with tag insertions for SISDC. (a) PCR amplification of fragments with marker tags. The primer names correspond to those in Table 1. (Parental half-length genes of PSE-4 β-lactamase were constructed using a second set of primers, also listed in Table 1.) (b)–(d) Tag-inserted parent genes are constructed by sequential assembly PCR, using complementary sequences contained in the marker tags. The parental half-length genes are then subjected to SISDC.

Reconstruction of full-length TEM-1 and PSE-4 genes

To verify the method, we recombined the fragments of each wild-type β-lactamase (TEM-1 and PSE-4) using the procedure described above and characterized the fragment reconstruction efficiency for each gene. In all, 56 transformants from each assembly were picked at random, and their sequences were analyzed by probe hybridization. All eight SCHEMA fragments were found in 51 of

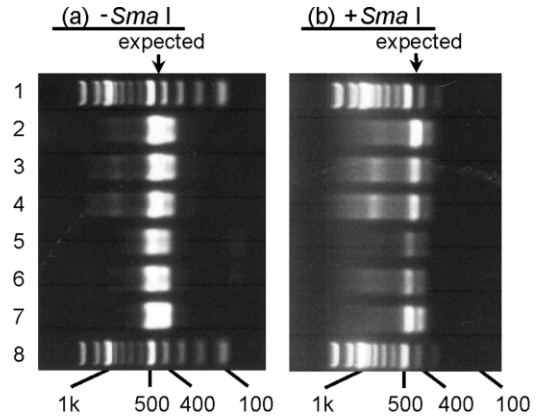

Figure 4. PCR amplification of half-length library TP1-4 after ligation. (a) Treatment with *Sma*I omitted before PCR and (b) treatment with *Sma*I used before PCR. Lanes 1 and 8, 100 bp DNA ladder; lane 2, recombined T1-4 fragment amplified by primer set TNasp and T4C; lanes 3–6, recombined TP1-4 libraries amplified by primer set TNasp and T4C, TNasp and P4C, PNasp and T4C, and PNasp and P4C, respectively; lane 7, recombined P1-4 fragment amplified by primer set PNasp and P4C.

56 clones from TEM-1 reconstructions (91%) and 53 of 56 clones from PSE-4 reconstructions (95%). Furthermore, complete sequences were determined from ten randomly selected transformants from the TEM-1 and PSE-4 reconstructions. In five of the ten clones, full-length wild-type sequence was reconstructed successfully without any mutation or frame-shift. Not a single point mutation was found among 8880 nucleotides sequenced. However, a single base-pair deletion was found at seven positions (0.08%), which caused a frame-shift in 50% of the clones. These single deletions occurred mainly at the ligation site of *Bae*I, and were probably caused by the restriction digestion.

The enzymatic activities of 24 clones picked at random from each reconstructed pool were checked by growth on LB + ampicillin plates. For TEM-1 reconstructions, nine of 24 clones (38%) could be grown on plates containing ampicillin (>3000 μg/ml), and three of 24 clones (13%) could be grown on plates containing ampicillin at 10 μg/ml and 100 μg/ml, but not >3000 μg/ml. For PSE-4 reconstructions, 13 of 24 clones (58%) could be grown on plates containing ampicillin at >3000 μg/ml. Thus, we estimated that the wild-type β-lactamase genes were reconstructed successfully with about 50% efficiency, and 10% have reduced activities, probably caused by point mutations.

Quality of the TP library

The recombined gene library encoding TEM-1 and PSE-4 β-lactamases (TP library) was analyzed by oligonucleotide probe hybridization.[9] From 128 clones picked at random, 88% were found to contain all eight fragments, and 87 unique compositions were identified. Figure 5(a) shows the overall frequency of TEM-1 sequences, indicating that there is little bias in fragment incorporation, except for fragment 2. By counting the number of instances where neighboring probe sites were occupied by sequences from different parents, we measured an average of 3.3 crossovers/gene for the pre-screened TP library. This is very close to the average of 3.5 crossovers we would expect if the fragments were incorporated at random. Furthermore, no bias was observed in the frequencies of crossovers (Figure 5(c); open circles), indicating that ligation reactions between neighboring elements occurred uniformly for all of the fragments from the two parents.

Eight transformants selected at random from the TP library were sequenced. Two point mutations (0.03%) and five single base-pair deletions (0.07%) were found among 7104 nucleotides sequenced, and 50% of the sequences were full-length genes without any mutations or frame-shifts. All of the mutations were silent, occurring in the third nucleotide of the codons. In addition, all of

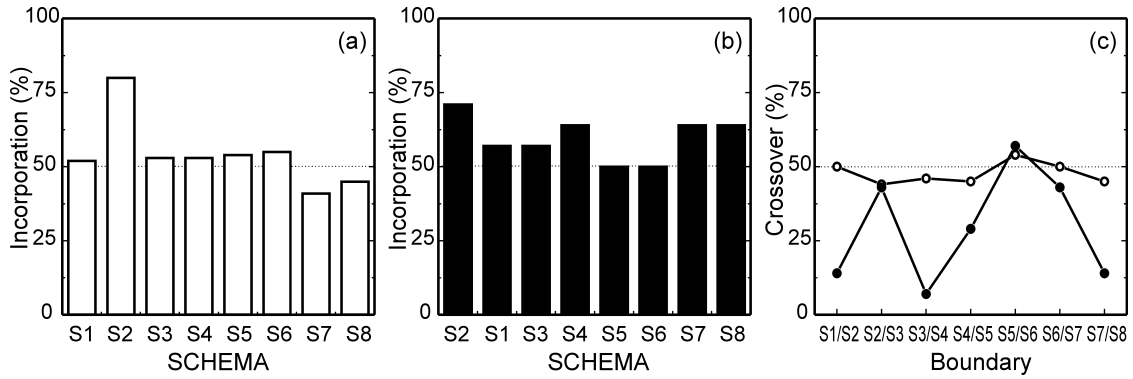

Figure 5. SCHEMA fragment incorporation and crossover positions in the TP library and functional chimeras identified by screening for antibiotic resistance. (a) The fractions of SCHEMA fragments from TEM-1 in the TP library, analyzed by probe hybridization (open bars) and (b) 14 functional chimeras analyzed by DNA sequencing (filled bars). (c) The frequency of crossovers at each SCHEMA boundary in the TP library analyzed by probe hybridization (open circles) and in 14 functional chimeras analyzed by sequencing (filled circles).

mutations and 1 bp deletions occurred at the ligation site of *Bae* I; we presume they arose during enzymatic digestion, as described above.

Screening of active chimeras

To screen for active clones, the TP library was transformed into XL10 cells and plated on LB plates containing chloramphenicol. XL10 does not have the Tet repressor, and protein expression is not repressed. Of 1472 transformants from the TP library 48 (3.3%) were found to survive on an LB plate containing 10 μg/ml or a higher concentration of ampicillin. DNA sequencing revealed that six clones (0.4%) encode wild-type β-lactamases, and 42 (2.9%) encode chimeric proteins. Taking into account that half of the original library contained full-length genes, wild-type clones account for ~0.8% of the library, consistent with expected value of 0.78% (two out of 256 possible sequences). The fraction of active chimeras is 5.8%. In all, 14 unique chimeric sequences were identified (Table 2). Active chimeras exhibited between one and four crossovers per gene. The average, 2.1, is smaller than that of the original TP library (average = 3.3).

Figure 5(b) illustrates the incorporation of fragments from TEM-1 in the active chimeras, showing that fragments from this parent were slightly favored in the active enzymes. The greatest bias appeared in the first fragment, where approximately 75% were from TEM-1. For all predicted crossover locations, crossover events were found in at least one of the functional mutants, indicating these positions are acceptable for recombination (Table 2). However, the frequencies of crossovers were not uniform for each location (Figure 5(c); filled circles). Crossovers were observed frequently at S2/S3, S5/S6 and S6/S7, but were relatively rare at S1/S2 and S3/S4. Crossover positions chosen from the SCHEMA profile are apparently not equally preferred in functional chimeras.

Four pairs of "mirror chimeras" were observed: 345P and 345T, 5P and 5T, 3456P and 3456T, and 6P and 6T (Table 2). Furthermore, most functional chimeras have an even number of crossovers (two or four). Similar results were obtained from another, much larger library of chimeric β-lactamases, in which the crossovers were allowed at both the minima and maxima of the SCHEMA profile.[29] All of the functional chimeras selected from this library had an even number of crossovers, suggesting that interactions between polypeptides far apart in primary sequence strongly affect the folding and function of the chimeric enzymes. In fact, the N and C-terminal helices are packed tightly against one another in the three-dimensional β-lactamase structures.[26,27] These interactions between distant sequence elements are apparently important for retention of lactamase function, although they were not considered in calculating the SCHEMA profile of Figure 2.

An obvious feature of the active chimeras is that they have low SCHEMA disruption (E_s) compared to the average of the library. Figure 6(a) shows the distribution of the effective number of mutations and E_s values for every chimera in the library (small bars): the active chimeras are indicated

Table 2. Sequences, calculated disruption number (E_s), number of mutations, and activities of functional chimeras

Chimera	Sequence[a] 1 2 3 4 5 6 7 8	Number of crossovers	E_s[b]	Number of mutations[c]	MIC[d] (μg/ml ampicillin)
345P		2	13	54	90
345T		2	13	54	100
8P		1	18	12	85
5P		2	20	14	1000
5T		2	20	14	45
3456P		2	27	85	12
3456T		2	27	85	80
2P		2	30	19	90
6P		2	32	31	600
6T		2	32	31	90
178T[e]		2	33	88	5
35P		4	35	39	40
67P		2	40	53	6
56P		2	44	29	7

[a] Sequences are represented by SCHEMA fragment composition: black box, TEM-1; gray box, PSE-4.
[b] E_s, calculated disruption (= number of contacts broken by recombination).
[c] Number of effective mutations is the minimum number of amino acid substitutions required to convert a chimera into one of parental sequence at those residues found in the TEM-1 structure[26] used for SCHEMA calculation (the amino acid residues aligning with residues 26–286 of TEM-1).
[d] MIC, minimal inhibitory concentrations of ampicillin, determined using liquid cultures.
[e] 178T has a point mutation leading to amino acid substitution K146I.

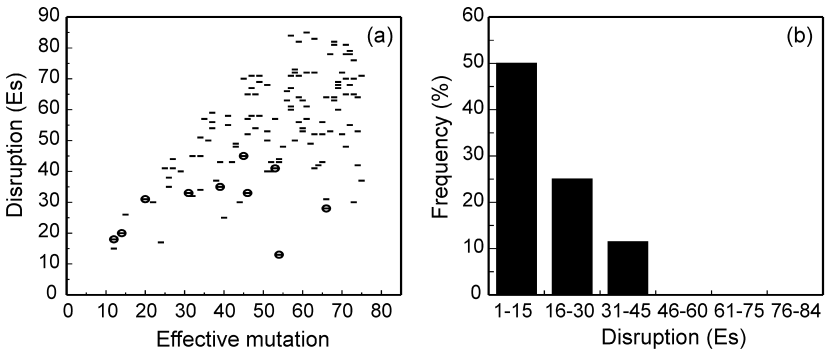

Figure 6. (a) Distribution of calculated E_s values and the effective number of mutations for all possible chimeras in TP library (small bars) and functional chimeras (open circles). E_s was determined for each chimera using the TEM-1 structure.[26] Effective number of mutations is the minimum number of amino acid substitutions required to convert a chimera into one of parental sequence at those residues used for SCHEMA calculation, i.e. the amino acid residues aligned with residues 26–286 of TEM-1. (b) Probability of functional chimeras. The percentage of chimeras found to retain function for various ranges of disruption number (E_s) was calculated by dividing the number of active chimeras by the number of possible chimeras in the library.

(open circles). While the number of effective mutations in functional chimeras is distributed over a broad range, their E_s values are low (maximum 45) compared to the average E_s for all chimeras with the same level of mutation. The significance of the E_s value becomes clearer for the chimeras with higher mutation levels. For functional chimeras with high mutation number (>45), most of the E_s values are near the lowest possible.

We analyzed the relationship between E_s and the proportion of active chimeras. As shown in Figure 6(b), the functional chimeras were highly concentrated at lower E_s values; active chimeras were obtained only for $E_s \leq 45$, and about 30% of possible chimeras were identified as functional for $E_s \leq 30$. It appears that the E_s value calculated by SCHEMA is a good indicator of the level of structural disruption caused by recombination and can be used to predict crossover positions leading to folded and functional chimeras.[29]

Advantages of SISDC

The method described here can be used to generate libraries of highly mosaic chimeras from two or more genes. It consists of four steps of simple genetic techniques: (1) amplification of building blocks with marker tags by PCR, (2) restriction digestion of tag regions, (3) a single ligation between elements, and (4) library amplification by PCR. We included another step, the construction of tag-inserted parental genes, between amplification of building blocks and restriction digestion. In theory, this step can be skipped by cutting the amplified fragments with *Bae*I and subsequent ligation. This shortcut method is simple, but a minor problem with this approach might be the experimental difficulty of controlling the concentrations of amplified fragments. In the approach used here, the molar concentration of every fragment can be easily controlled, thus decreasing bias in parental incorporation.

Because a specific sticky end is designed for each boundary, all of the fragments should be ligated in the right order, duplication of fragments is avoided, and parental-length chimeras should be assembled easily. It is necessary that the parental genes all share similar unique consensus sequences at the desired recombination sites, resulting in one or two amino acids of identity at each crossover site. In this study, a 5 bp consensus sequence was required as a *Bae*I cleavage site. If another type IIb enzyme is used, for example *Bsa*XI, whose recognition sequence is (9/12)AC(NNNNN)GTCC(10/7), the required consensus sequence is only 3 bp (one amino acid residue), although this modification would probably result in reduced efficiency in the subsequent ligation step.

A major advantage of the SISDC method is that the number of primers for inserting the tag sequence into the parental genes scales linearly with the number of parental genes. The number of primers is $2 \times P \times E$ for this method, *versus* $P^2 \times (E-1) + 2P$ for methods that use chimeric oligonucleotides, such as SOEing, where P is the number of parents and E is the number of elements shuffled. Thus, when more than two parent genes are recombined, SISDC uses fewer primers than conventional SOEing. In addition, SISDC is expected to be less mutagenic than PCR-mediated methods, because the random errors introduced during iterative PCR steps are avoided. In this method, two PCR steps were used after sequence certification of the parental plasmids, for amplification of tag-inserted parental genes and amplification of the ligated library. The first PCR step can be avoided by using a restriction digestion from the parental plasmids. Most importantly, we have shown that a library constructed by SISDC has little bias in incorporation of the different elements and crossover frequency. Biases are expected with PCR-based methods, due to the difficulty of estimating the actual concentration of each fragment to be recombined after the first PCR, and the difficulty of finding an annealing temperature for the chimeric primers and the template fragments in the second PCR that does not introduce bias.

In conclusion, this simple and cost-effective method will be especially useful for the construction of chimeric libraries from multiple parents with several crossovers. Sequence identity between parent genes is not necessary, and only one or two amino acids at each crossover position need to be fixed. Combinatorial recombination of multiple elements can be accomplished with a single ligation step, with control over the order of the elements. If desired, various modifications, such as insertion, deletion, and rearrangement, can be incorporated easily. This method can be used for different applications of *in vitro* recombination, including SCHEMA and exon or domain shuffling.

Materials and Methods

Materials

Enzymes were purchased from New England Biolabs, Beverly, MA (*Bae*I, *Sma*I, and Vent DNA polymerase), Amersham Pharmacia Biotech, Piscataway, NJ (*Sse*8387I) and Roche Applied Science, Hague Road, IN (*Asp*718, *Pst*I, and rapid DNA ligation kit). *Escherichia coli* strain TOP10 (Invitrogen, Carlsbad, CA), DH5α-Pro (BD Biosciences Clonetech, Palo Alto, CA), and XL10 (Stratagene, La Jolla, CA) were used for cloning and plasmid maintenance of parental genes, recombined genes, and for screening of active chimeras, respectively. Luria-Bertani (LB) medium (10 g/l of Tryptone, 5 g/l of yeast extract, and 10 g/l of NaCl) containing appropriate antibiotics (ampicillin and/or chloramphenicol, Sigma, St. Louis, MO) was used for culturing *E. coli* cells. All PCR was carried out on a MJ Research PTC-200 thermal cycler (Watertown, MA).

Primer design for construction of marker tag-inserted parental gene

The nucleotide sequences of the TEM-1 and PSE-4 genes were aligned (Figure 1(a)) and the genes coding for the β-lactamases were divided into eight fragments based on the SCHEMA disruption profile (Figure 2).[20] Crossovers were allowed between TEM-1 residues 40–41, 68–69, 104–105, 131–132, 164–165, 225–226 and 265–266, and structurally related residues in PSE-4, 54–55, 81–82, 117–118, 144–145, 177–178, 238–239 and 278–279. At each of these boundaries, the TEM-1 and PSE-4 β-lactamases share the same amino acids. Thus, unique 5 bp consensus sequences for each crossover point (see below) can be chosen without introducing any amino acid substitutions.

The marker tags were designed to encode 5 bp consensus sequences at both ends (S_5: specific to each crossover point, resulting in a synonymous mutation) and a recognition sequence of a restriction enzyme, *Bae*I (Figure 1(b) and Table 1). The *Bae*I recognition sequence consisted of various X_{10} and Y_4 regions for assembly PCR for parental gene construction. Also, a *Sma*I site was included to allow removal of residual parental genes from the final library. A synthetic primer consisted of a complementary sequence to a parent gene flanked by a part of a marker tag (14 bp) to allow amplification of the DNA in the eight segments (Table 1).

Creation of parental plasmids

Parental plasmids, in which a variety of *Bae*I recognition sequences were inserted into seven targeted positions of the TEM-1 and PSE-4 genes, were created by step-wise assembly PCR (Figure 3). First, gene fragments that encode the SCHEMA fragments of β-lactamases from *E. coli* and *Pseudomonas aeruginosa* (T1 ~ 8 for TEM-1, and P1 ~ 8 for PSE-4) were amplified by PCR from plasmids pSTBlue-1 (Novagen, Madison, WI) and pMON711,[30] respectively (Figure 3(a) and (b)), using the set of flanking primers listed in Table 1. A 50 μl reaction mixture contained Vent DNA polymerase buffer, 0.2 mM each dNTP, 10 pmol of each primer, two units of Vent DNA polymerase and 10 ng of template plasmid. Reaction mixtures were heated at 95 °C for five minutes, followed by 25 cycles of incubation at 95 °C for one minute, 52 °C for 30 seconds, 72 °C for 30 seconds, and a final incubation at 72 °C for ten minutes. The products were verified on and purified from Seakem GTG agarose gels (BioWhittaker Molecular Applications, Rockland, ME) with a Zymoclean Gel DNA recovery kit (Zymo Research, Orange, CA), and redissolved in 20 μl of 10 mM Tris–HCl (pH 8.5). Next, T12, T34, T56, T78, P12, P34, P56 and P78 fragments were created by assembly PCR, using complementary sequences designed in marker tags (Figure 3(c)). The assembly reaction of DNA fragments was carried out in Vent DNA polymerase buffer, 0.2 mM each dNTP, 5 μl of each DNA fragment, 20 pmol of each primer, and four units of Vent DNA polymerase in a total volume of 100 μl. Reaction mixtures were heated at 95 °C for five minutes, followed by 25 cycles of incubation at 95 °C for one minute, 52 °C for 30 seconds, 72 °C 30 seconds, and final incubation at 72 °C for 20 minutes. The products were purified as described above. The half-length gene fragments (T1234, T5678, P1234, and P5678) were constructed by assembly PCR as for the first assembly reaction (Figure 3(d)) and purified by agarose gel extraction. The final products were cloned into pCR-Blunt II-TOPO vector (Invitrogen, Carlsbad, CA) for DNA maintenance and transformed into TOP10 cells. The parental plasmids were purified by QIAquick plasmid mini-prep kit (Qiagen, Valencia, CA), and all tag-insertions were certified by DNA sequencing.

Construction of half-length library

The half-length gene fragments were amplified again by PCR from the parental plasmids described above, using a set of primers (TNasp and T4C for T1234, T4N and TCsse for T5678, PNasp and P4C for P1234, P4N and PCsse for P5678). About 4 μg of fragments was subjected to digestion by *Bae*I. After purification using a QIAquick spin column (Qiagen, Valencia, CA), the digested fragments from T1234 and P1234, or T5678 and P5678, were mixed and ligated by T4 DNA ligase in a 10 μl reaction mix. After removal of untreated DNA fragments with *Sma*I, the recombined half-gene libraries were amplified by PCR using a set of primers (TNasp/PNasp and T4C/P4C for library 1–4 or T4N/P4N and TCsse/PCsse for library 5–8). The PCR products purified by gel-extraction were cloned back into pCR-Blunt II TOPO vector and transformed into TOP10. The number of independent clones in the half-length library was estimated by counting the number of colonies from an aliquot of transformants. Half-length library plasmids were collected from overnight culture in LB medium

containing 34 μg/ml of chloramphenicol at 37 °C, and were purified using a QIAquick plasmid mini-prep-kit (Qiagen, Valencia, CA).

Construction of full-length library

The half-length libraries were digested from the TOPO vector with *Asp*718/*Bae*I (for library 1–4) or *Bae*I/*Sse*838I (for library 5–8). After gel-purification and DNA quantification, equal amounts of the half-gene library fragments were mixed and ligated together into the *Asp*718/*Pst*I sites of the expression plasmid pPROtet (Clonetech, Palo Alto, CA). Competent *E. coli* DH5α-Pro cells were transformed with the ligation mixture and grown overnight at 37 °C in LB medium containing 34 μg/ml of chloramphenicol. The number of independent clones in the library was estimated by counting the number of colonies from an aliquot of transformants. SCHEMA fragments from TEM-1 or PSE-4 were ligated separately to reconstruct the wild-type.

Construction of positive control vector

As positive controls, TEM-1 and PSE-4 were amplified by PCR from pSTBlue-1 and pMON711, respectively, using a set of primers (TNasp and TCsse for TEM-1, PNasp and PCsse for PSE-4), as described above. PCR products were digested with *Asp*718/*Sse*8387I and cloned into the *Asp*718/*Pst*I sites of pPROtet by ligation (pPRO-TEM and pPRO-PSE).

Library characterization by probe hybridization analysis

Clones from the pre-screened library were analyzed by probe hybridization in a macroarray format as described.[9] An oligonucleotide probe was designed to bind specifically to each of the parents at the eight SCHEMA fragments at approximately the same temperature. The 16 probes for the β-lactamase libraries were obtained from Qiagen Operon (Alameda, CA).

Screening for active chimeras

The TP library was transformed into XL10 cells and plated on LB plates containing 34 μg/ml of chloramphenicol. Colonies (1472) from the pre-screened library were picked randomly from overnight cultured LB plates and inoculated into the 96 deep-well plates, which were filled with 250 μl of LB medium containing 34 μg/ml of chloramphenicol. The plates were incubated overnight with shaking at 37 °C and replicated onto LB plates containing 34 μg/ml of chloramphenicol and various concentrations of ampicillin (0, 10, 50, 100, 500, 1000, 2000, and 3000 μg/ml). The XL10 cells transformed with pPROtet (negative control) do not grow on plates with ampicillin at concentrations of 4 μg/ml or higher, while XL10 cells transformed with pPRO-TEM and pPRO-PSE (positive controls) do grow on plates with 3000 μg/ml of ampicillin. After overnight incubation, the active clones that grew on plates containing ampicillin were picked and cultured in 2 ml of LB containing 34 μg/ml of chloramphenicol. The plasmids were recovered using a QIAquick plasmid mini-prep kit and subjected to sequencing. The activities of screened chimeras were measured as the minimum inhibitory concentration (MIC) of antibiotic. The candidate plasmids were introduced into XL10 again, and the isolated colonies were incubated in 400 μl of LB medium containing 34 μg/ml of chloramphenicol and a range of concentrations of ampicillin. The MIC values for XL10 transformed with pPROtet, pPRO-TEM, and pPRO-PSE were 4 μg/ml, 3000 μg/ml, and 6000 μg/ml, respectively.

DNA sequence analysis

All sequencing was performed at the Sequence Analysis Facility, Biology Division, California Institute of Technology.

Acknowledgements

We thank Christopher A. Voigt for the SCHEMA calculation, and Jeffrey Endelman and Jonathan Silberg for helpful discussions on data analysis. This research was supported, in part, by a JSPS Postdoctoral Fellowship for Research Abroad (to K.H.).

References

1. Minshull, J. & Stemmer, W. P. C. (1999). Protein evolution by molecular breeding. *Curr. Opin. Chem. Biol.* **3**, 284–290.
2. Stevenson, J. D. & Benkovic, S. J. (2002). Combinatorial approaches to engineering hybrid enzymes. *J. Chem. Soc.* **2**, 1483–1493.
3. Coco, W. M., Levinson, W. E., Crist, M. J., Hektor, H. J., Darzins, A., Pienkos, P. T. *et al.* (2001). DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nature Biotechnol.* **19**, 354–359.
4. Crameri, A., Raillard, S. A., Bermudez, E. & Stemmer, W. P. C. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, **391**, 288–291.
5. Moore, J. C., Jin, H. M., Kuchner, O. & Arnold, F. H. (1997). Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272**, 336–347.
6. Shao, Z., Zhao, H., Giver, L. & Arnold, F. H. (1998). Random-priming *in vitro* recombination: an effective tool for directed evolution. *Nucl. Acids Res.* **26**, 681–683.
7. Stemmer, W. P. C. (1994). Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, **370**, 389–391.
8. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. & Arnold, F. H. (1998). Molecular evolution by staggered extension process (StEP) *in vitro* recombination. *Nature Biotechnol.* **16**, 258–261.
9. Joern, J. M., Meinhold, P. & Arnold, F. H. (2002). Analysis of shuffled gene libraries. *J. Mol. Biol.* **316**, 643–656.
10. Ostermeier, M., Shim, J. H. & Benkovic, S. J. (1999). A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotechnol.* **17**, 1205–1209.
11. Lutz, S., Ostermeier, M., Moore, G. L., Maranas, C. D. & Benkovic, S. J. (2001). Creating multiple-crossover

DNA libraries independent of sequence identity. *Proc. Natl Acad. Sci. USA*, **98**, 11248–11253.
12. Sieber, V., Martinez, C. A. & Arnold, F. H. (2001). Libraries of hybrid proteins from distantly related sequences. *Nature Biotechnol.* **19**, 456–460.
13. Apic, G., Gough, J. & Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325.
14. de Souza, S. J., Long, M. & Gilbert, W. (1996). Introns and gene evolution. *Genes Cells*, **1**, 493–505.
15. Doolittle, R. F. (1995). The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314.
16. Gilbert, W. (1978). Why genes in pieces? *Nature*, **271**, 501.
17. Thornton, J. M., Orengo, C. A., Todd, A. E. & Pearl, F. M. G. (1999). Protein folds, functions and evolution. *J. Mol. Biol.* **293**, 333–342.
18. O'Maille, P. E., Bakhtina, M. & Tsai, M. D. (2002). Structure-based combinatorial protein engineering (SCOPE). *J. Mol. Biol.* **321**, 677–691.
19. Tsuji, T., Onimaru, M. & Yanagawa, H. (2001). Random multi-recombinant PCR for the construction of combinatorial protein libraries. *Nucl. Acids Res.* **29**, E97.
20. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nature Struct. Biol.* **9**, 553–558.
21. Kolkman, J. A. & Stemmer, W. P. C. (2001). Directed evolution of proteins by exon shuffling. *Nature Biotechnol.* **19**, 423–428.
22. Tsuji, T., Yoshida, K., Satoh, A., Kohno, T., Kobayashi, K. & Yanagawa, H. (1999). Foldability of barnase mutants obtained by permutation of modules or secondary structure units. *J. Mol. Biol.* **286**, 1581–1596.
23. Horton, R. M., Hunt, H. D., Ho, S. N., Pullen, J. K. & Pease, L. R. (1989). Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene*, **77**, 61–68.
24. Cho, G., Keefe, A. D., Liu, R., Wilson, D. S. & Szostak, J. W. (2000). Constructing high complexity synthetic libraries of long ORFs using *in vitro* selection. *J. Mol. Biol.* **297**, 309–319.
25. Riechmann, L. & Winter, G. (2000). Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc. Natl Acad. Sci. USA*, **97**, 10068–10073.
26. Jelsch, C., Mourey, L., Masson, J. M. & Samama, J. P. (1993). Crystal structure of *Escherichia coli* TEM1 β-lactamase at 1.8 Å resolution. *Proteins: Struct. Funct. Genet.* **16**, 364–383.
27. Lim, D., Sanschagrin, F., Passmore, L., De Castro, L., Levesque, R. C. & Strynadka, N. C. J. (2001). Insights into the molecular basis for the carbenicillinase activity of PSE-4 β-lactamase from crystallographic and kinetic studies. *Biochemistry*, **40**, 395–402.
28. Sears, L. E., Zhou, B., Aliotta, J. M., Morgan, R. D. & Kong, H. (1996). *Bae* I, another unusual *Bcg* I-like restriction endonuclease. *Nucl. Acids Res.* **24**, 3590–3592.
29. Meyer, M. M., Silberg, J. J., Voigt, C. A., Endelman, J. B., Mayo, S. L., Wang, Z.-G. & Arnold, F. H. (2003). Library analysis of SCHEMA-guided protein recombination. *Protein Sci.* In the press.
30. Sabbagh, Y., Theriault, E., Sanschagrin, F., Voyer, N., Palzkill, T. & Levesque, R. C. (1998). Characterization of a PSE-4 mutant with different properties in relation to penicillanic acid sulfones: importance of residues 216–218 in class A β-lactamases. *Antimicrob. Agents Chemother.* **42**, 2319–2325.

Edited by F. Schmid

(Received 11 February 2003; received in revised form 27 April 2003; accepted 28 April 2003)