Protein Engineering vol. 16 no. 12 pp. 1025–1034, 2003
DOI: 10.1093/protein/gzg129

Using a residue clash map to functionally characterize protein recombination hybrids

Manish C.Saraf and Costas D.Maranas[1]

Department of Chemical Engineering, The Pennsylvania State University, 112 Fenske Laboratory, University Park, PA 16802, USA

[1]To whom correspondence should be addressed.
E-mail: costas@psu.edu

In this article, we introduce a rapid, protein sequence database-driven approach to characterize all contacting residue pairs present in protein hybrids for inconsistency with protein family structural features. This approach is based on examining contacting residue pairs with different parental origins for different types of potentially unfavorable interactions (i.e. electrostatic repulsion, steric hindrance, cavity formation and hydrogen bond disruption). The identified clashing residue pairs between members of a protein family are then contrasted against functionally characterized hybrid libraries. Comparisons for five different protein recombination studies available in the literature: (i) glycinamide ribonucleotide transformylase (GART) from *Escherichia coli* (purN) and human (hGART), (ii) human Mu class glutathione *S*-transferase (GST) M1-1 and M2-2, (iii) β-lactamase TEM-1 and PSE-4, (iv) catechol-2,3-oxygenase *xylE* and *nahH*, and (v) dioxygenases (toluene dioxygenase, tetrachlorobenzene dioxygenase and biphenyl dioxygenase) reveal that the patterns of identified clashing residue pairs are remarkably consistent with experimentally found patterns of functional crossover profiles. Specifically, we show that the proposed residue clash maps are on average 5.0 times more effective than randomly generated clashes and 1.6 times more effective than residue contact maps at explaining the observed crossover distributions among functional members of hybrid libraries. This suggests that residue clash maps can provide quantitative guidelines for the placement of crossovers in the design of protein recombination experiments.

Keywords: bioinformatics/directed evolution/protein engineering/residue–residue clash

Introduction

Directed evolution is a strategy for improving a specific biological function (thermostability, stereoselectivity, catalytic activity, expanded substrate specificity) through genetic diversification and selection, emulating natural evolution in an accelerated and guided fashion (Moore *et al*., 1997). The diversity generating mechanism commonly entails the exchange of parental DNA fragments in the reassembled sequences through recombination and/or involves altered residue sites through random mutagenesis. One of the key challenges in the use of such directed evolution techniques for protein engineering is that in some cases, particularly when the parental sequences share low sequence identity, the reassembled sequences do not even fold properly and thus are non-functional. Moreover, it has been observed experimentally that the lower the sequence identity between the recombined parental sequences, the larger the proportion of the library that is not functional (Wang, 2000). The majority of the DNA-shuffling methods can only recombine closely related sequences and generate crossovers only within regions of high (i.e. >60%) sequence identity. However, with the advent of more versatile techniques such as ITCHY (Ostermeier *et al*., 1999), SCRATCHY (Lutz *et al*., 2001), SHIPREC (Sieber *et al*., 2001) and sequence-independent site-directed chimeragenesis (SISDC) (Hiraga and Arnold, 2003) greater diversity can be created by recombining distant homologs. This unfortunately often leads to an increasingly large proportion of the combinatorial library being non-functional. In an earlier paper (Saraf *et al*., 2003), we showed that functionally important protein regions are not necessarily conserved and instead found that they are more likely to exhibit strongly correlated substitution patterns with other regions. Moore and Maranas (Moore and Maranas, 2003) utilized the energetics of molecular interactions to identify residue–residue clashes using second-order mean field calculations and found that most of these clashes could be attributed to steric, charge or hydrogen bond disruptions in the hybrids. In this paper, we directly look for these types of clashes based on protein sequence data and compare these predictions against sequence data obtained for functional recombinant libraries.

A number of hypotheses have been proposed (Bogarad and Deem, 1999; Voigt *et al*., 2001) to explain why functional crossovers are not randomly distributed along the sequence but rather form distinct patterns. One of the most recent methods, the SCHEMA algorithm (Voigt *et al*., 2002), postulates that crossover patterns resulting in hybrids with a large number of contacting residue pairs originating from the same parental sequences are more likely to retain their functionality. The key idea here is that each contact is a representation of favorable interaction between the two residues. Thus, by retaining these contacting residues in the hybrids, one retains the favorable interactions that exist in the parental sequences. This interesting approach has led to a number of successful predictions (Hiraga and Arnold, 2003; Meyer *et al*., 2003). One potential shortcoming, however, is that it cannot differentiate between hybrids with different directionality (i.e. an A–B versus a B–A crossover), which often have substantially different functionalities (Lutz *et al*., 2001; Moore and Maranas, 2003). Here, we rethink the effect of having contacting residue pairs with different parental origins. Instead of always counting them as unfavorable, we view such pairs as places where clashes may or may not occur between the contacting residues. This view allows us to re-establish 'context' in the interaction between the residue pair and thus capture the effect of crossover directionality (e.g. an A–B versus a B–A crossover) on function. Specifically, motivated by the results of Moore and

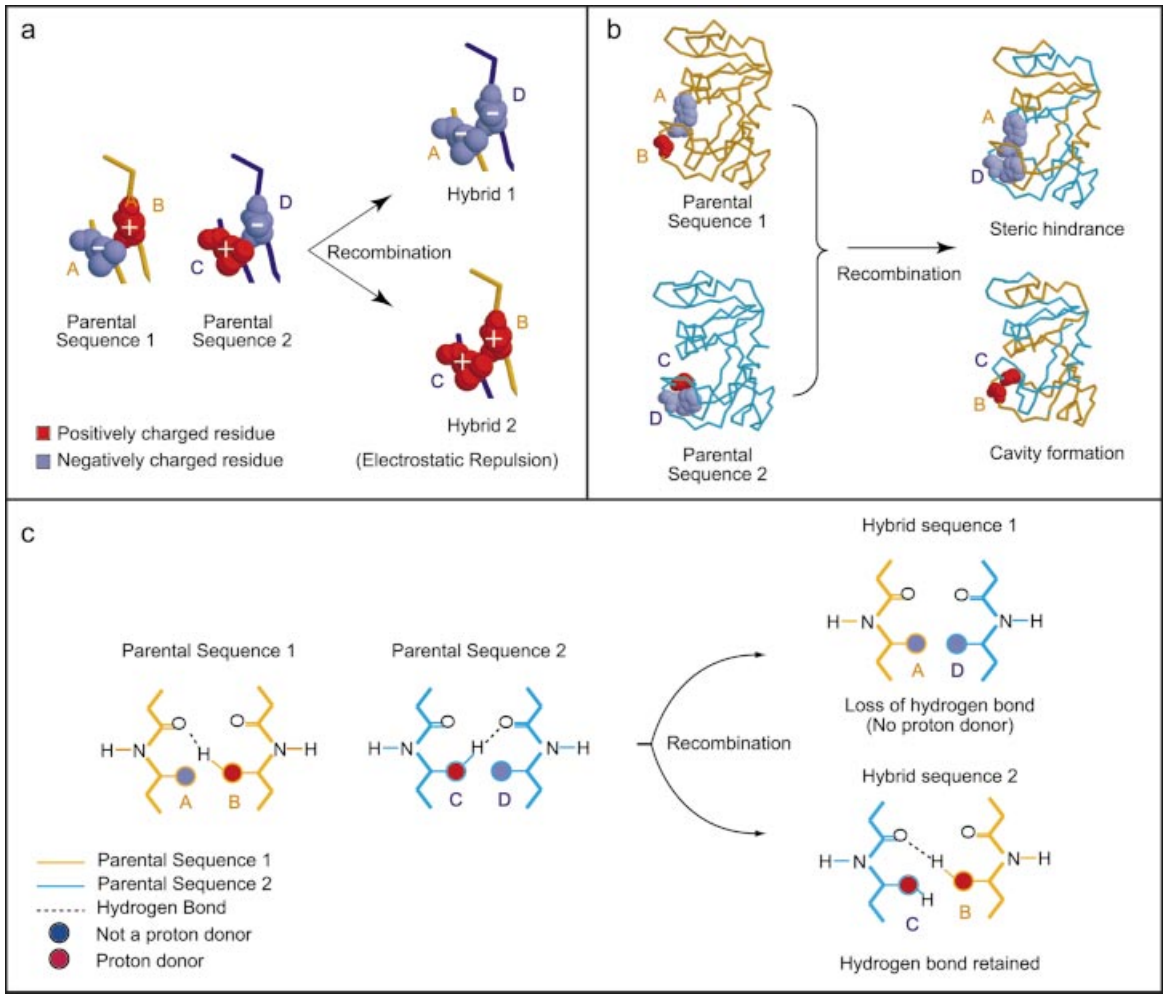

Fig. 1. (**a**) The contacting residues A–B (parental sequence 1) and C–D (parental sequence 2) have opposite charges and different relative positions in the two parental sequences. Recombination results in electrostatic repulsion between residues A–D (–/–) in the first hybrid and B–C (+/+) in the second hybrid. (**b**) The first hybrid retains residues with large side chains from both parental sequences 1 and 2 (A–D) causing steric hindrance. Pairing of the residues with small side chains (C–B) in the second hybrid leads to a cavity formation. (**c**) Hybrid 2 retains proton donors (C, B) from both parental sequences and thus the hydrogen bond between the side chain donor and the backbone acceptor is retained. Alternatively, hybrid 1 retains residues with side chains that have no proton donors (A, D) resulting in the loss of the hydrogen bond between the two residues.

Maranas (2003), we explore three out of the many different mechanisms that may render a contacting residue pair detrimental to the ability of the hybrid to fold properly (i.e. stability) and thus retain its functionality: (i) introduction of repulsive residue pairs such as +/+ or –/–, (ii) disruption of hydrogen bonds due to the formation of donor/donor or acceptor/acceptor pairs and (iii) generation of steric clashes or cavities. It is quite straightforward to show that upon recombination residue clashes such as the repulsive residue pairs, disrupted hydrogen bonds and steric clashes can be introduced due to reversed orientation of charged, acceptor/donor or bulky residue pairs (Figure 1). Other forms of clashes, not considered here, include the disruption of important protein-specific interactions (Oldfield, 2002) such as metal binding motifs (Glusker, 1991), the catalytic triad (Fischer *et al*., 1994; Wallace *et al*., 1997) and a number of ligand binding sites (Chakrabarti, 1993; Copley and Barton, 1994).

The proposed procedure extends the concept of a residue contact map (Voigt *et al*., 2002) by relying on the construction of a residue clash map (i.e. a plot representing all possible clashing residue pairs in the reassembled sequences) based on the properties of the pair of residues that are in contact and have different parental origins. Notably, we find that the pattern of clashing residue pairs is greatly dependent on the crossover directionality. By superimposing these predicted clashing residue pairs against functional crossover statistics available in the literature we find that these clashes are preferentially avoided in the hybrids with %ACC (percent of avoided calculated clashes) ranging from 61 to 100%. Note that here we define %ACC as the percentage of predicted clashes that are avoided by all the functional hybrids available in the data set. In contrast, results obtained based on the residue contact map (i.e. a plot representing all non-conserved contacting residue pairs that have different parental origins) yielded %ACC ranging from 30 to 71% while the results from randomly generated clashes yielded %ACC ranging from 9 to 54%.

Methods

Parental sequences participating in directed evolution, though sometimes highly divergent at the sequence level, share very similar structural traits. This implies that the basic structural

characteristics have to be largely preserved at least among the functional protein hybrids. These structural constraints enable us to construct the contact maps of the hybrids by simply querying the inter-residue distances calculated from the coordinates of the parental sequences obtained from the Protein Data Bank (PDB) (Westbrook *et al*., 2002). Note that the contact map of a parental sequence is the list of all residue pairs whose β-carbons (Cβ), or α-carbons in the absence of Cβ, are within a cut-off distance of 8 Å (Gobel *et al*., 1994). These contacting residue positions are adjusted according to the structural alignment between the two parental sequences using the combinatorial extension (CE) method (Shindyalov and Bourne, 1998). Next, the contact map of the hybrid is generated by retaining only those contacting residue positions that are common to the contact maps of both parental sequences. Pairs of contacting residue positions with at least one residue conserved in both parental sequences are excluded since the corresponding residue pair in the hybrid will always be present at these positions in at least one of the parental sequences. In cases where there are no structural data for a particular parental sequence, a predicted structure is used for identifying contacting residue pairs. The predicted structure is inferred using Swiss-Model (Schwede *et al*., 2003) and a homologous structure as the template. This homologous structure is obtained either from the ExPDB database (http://www.expasy.org/swissmod/SM_Check_ExPDB.html) or using a BLAST search on the PDB (Berman *et al*., 2000) to find the nearest match. In all cases described in this study, the template and the parental sequence whose structure is modeled share a relatively high sequence identity (>60%). It has been reported that predicted structures modeled using templates with such a high sequence identity are fairly reliable (Schwede *et al*., 2003). The Swiss-Model protein modeling server uses the template as an initial structure and replaces the template structure side chains with side chain conformations selected from a backbone-dependent rotamer library. These selections are made using a scoring function trading off favorable interactions such as hydrogen bonds, disulfide bridges and unfavorable close contacts. Side chain placement in the protein structure is fine-tuned through a steepest descent energy minimization algorithm using the GROMOS96 force field (van Gunsteren *et al*., 1996). Next, the contact maps of the hybrids generated as described above are investigated for clashes based on the three mechanisms (i.e. electrostatic repulsion, steric clashes and hydrogen bond disruptions).

Repulsive residue pairs

Residue pairs found in the contact map of the hybrids are screened for +/+ or –/– charge contacts that may be brought about by recombination (Figure 1a). A contacting pair that has a repulsive residue pair (+/+ or –/–) at these positions in either of the parental sequences is not counted since they evidently do not seem to disrupt functionality. Note that the crossover directionality is automatically accounted for since charge repulsion may be generated between residue pairs in one hybrid but not necessarily in the hybrid that has the reverse directionality (Figure 1a). For example, parental contacting residue pairs with a single charged residue (n/+ and +/n) may form upon recombination either a neutral pair (n/n) or a repulsive residue pair (+/+) depending on the directionality of the crossover. Also, lysine and arginine are considered to be positively charged and glutamate and aspartate as negatively charged.

Steric hindrance or cavity formation in the hybrids

A significant reduction in the total volume of a contacting residue pair is likely to give rise to a cavity formation, whereas a corresponding increase may cause steric hindrance. Figure 1b illustrates the effect of such volume changes as a consequence of the reversed orientation of large (residues A, D) and small (residues B, C) side chains in the parental sequences. Cavity formation or steric hindrance is detected by observing whether the combined volume of the contacting residue pair in the resultant hybrid is much lower or higher than the mean combined volume (M) of the same contacting residue pairs in the parental sequences (A+B, C+D):

$$M = \frac{1}{2}\left[(V_A + V_B) + (V_C + V_D)\right] \quad (1)$$

Here V_k is the side chain volume of residue k (k = A, B, C, D) in Å^3. Specifically, the scores S_{AD} and S_{CB} (for hybrids 1 and 2 shown in Figure 1b) are defined separately for hybrids with different crossover directionality as a measure of the deviation from M:

$$S_{AD}\begin{cases}|(V_A + V_D) - (M + \Delta)|, & \text{if } V_A + V_D \geq M + \Delta \text{ (steric hindrance)}\\ |(V_A + V_D) - (M - \Delta)|, & \text{if } V_A + V_D < M - \Delta \text{ (cavity formation)}\end{cases} \quad (2)$$

A parameter [$\Delta = |(V_A + V_B) - (V_C + V_D)|$], which quantifies the extent of difference between the combined volumes of the two parental contacting residue pairs, is introduced into these scores to account for the tolerance of such volume changes. If the contacting residue pairs in both parental sequences are of similar size, they could lead to a small (even zero) value of Δ, thus resulting in artificially inflated scores particularly in cases where the large and small residues have reversed orientation. Therefore, a lower bound is set on Δ equal to 10% of the mean (M):

$$\Delta = \begin{cases} |(V_A + V_B) - (V_C + V_D)| & \text{if } |(V_A + V_B) - (V_C + V_D)| \geq \dfrac{M}{10} \\ \dfrac{M}{10} & \text{if } |(V_A + V_B) - (V_C + V_D)| < \dfrac{M}{10} \end{cases} \quad (3)$$

In general, the core of most proteins has a higher packing fraction as compared with the surface (Munson *et al*., 1996). This suggests that steric clashes are less likely to be tolerated in the protein core (Dupraz *et al*., 1990) as they often lead to packing defects (Song *et al*., 1999; Ratnaparkhi and Varadarajan, 2000). To account for the difference in the tolerance level for steric clashes at the protein surface and in the core, we set different cut-off scores S_c for contacting pairs. Cavity formation and steric hindrance in the core of the protein (i.e. accessible surface area of side chain <8 Å^2) are considered to be significant if they score above a cut-off value, S_c = 15 Å^3, whereas only steric hindrance is considered with a cut-off value of 30 Å^3 at the surface. The accessible surface area of a side chain is obtained by rolling a water probe of radius 1.4 Å over the exposed surface. These calculations are performed using the WHATIF software package (Vriend, 1990).

Hydrogen bond disruption

Protein family members share many common hydrogen bonds, particularly those that are essential for functionality (Agarwal *et al*., 2002; Loll *et al*., 2003). Swapping the positions of the donor and acceptor groups of a hydrogen bond within a sequence preserves the hydrogen bond. However, similarly to

volume and charge clashes, orientation reversals of the donor and acceptor groups in parental sequences lead to hybrids with donor–donor or acceptor–acceptor contacting pairs, thus disrupting the hydrogen bond between the two residues (Figure 1c). Note that hydrogen bonds between two backbone atoms are not of interest here since both the acceptor (CO) and donor (NH) groups are retained upon recombination. Here, we consider all possible cases (i.e. side chain/backbone and side chain/side chain) to identify potentially disrupted hydrogen bonds. The WHATIF software package (Vriend, 1990) is used to detect common hydrogen bonds and identify the donor and acceptor groups of the parental sequences.

Contacting residue pairs identified for hybrids that violate at least one of the above three criteria (i.e. charge repulsion, steric hindrance and hydrogen bond disruption) are denoted as arcs (Figure 2) linking the two residue positions. A crossover occurring between these two positions results in differing parental origins for the two contacting residues, connected by the arc, in the resulting hybrid. This representation of clashes is generalized for hybrids with multiple crossovers by using bicolored arcs to encode the specific directionality of the parental combination leading to a clash. We next examine the effectiveness of the proposed residue clash maps at explaining known functional crossover combinations for a number of protein systems.

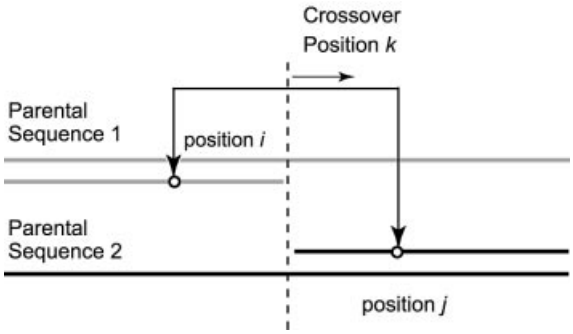

Fig. 2. An unfavorable interaction between the two residues at positions *i* and *j* in the hybrid is represented by an arc between the two positions. The residue at position *i* is retained from parental sequence 1 and *j* from parental sequence 2. Arcs depict any one of the three forms of clashes: (i) electrostatic repulsion, (ii) steric clashes and (iii) hydrogen bond disruption. A crossover at position *k* ($i < k < j$) brings the two contacting residues with different parental origins together, thus forming a clash.

Results

Residue clash maps are generated for the following five systems: (i) glycinamide ribonucleotide transformylase (GART) hybrids from *Escherichia coli* (purN) and human (hGART), (ii) human Mu class glutathione *S*-transferase (GST) M1-1 and M2-2, (iii) β-lactamase TEM-1 and PSE-4, (iv) catechol-2,3-oxygenase (C23O) *xylE* and *nahH*, and (v) dioxygenases *todC1C2* (toluene dioxygenase), *tecA1A2* (tetrachlorobenzene dioxygenase) and *bhpA1A2* (biphenyl dioxygenase).

These systems vary considerably not only in terms of pairwise sequence identity and number of functional hybrids, but also in the directed evolution protocol used for generating crossovers. All possible residue pairs with different parental origin that are brought in contact in one (or more) of the resultant hybrids are screened for all three forms of clashes. These clashes are then shown as arcs composing the residue clash map (Figure 2). This representation is used for hybrids with a single crossover (GART) while a generalized representation (i.e. bicolored arcs) is used for hybrids with multiple crossovers (GST, β-lactamase, C23O, and dioxygenases). A detailed comparison of the available experimental data using the proposed (i) residue clash map, (ii) residue contact map, and (iii) randomly generated clashes is presented. A randomly generated clash map is constructed by randomly choosing an Arbitrary number of pairs of non-conserved residue positions from the structural alignment. Note that conserved residue positions are not of interest here since they are also conserved in the hybrids and therefore will not form a clash. These results are examined in terms of %ACC (percent of avoided calculated clashes), defined as the percentage of the predicted clashes avoided by the functional hybrids present in the data set, and %CFC (percent of clash free crossovers), defined as the percentage of the observed functional crossovers that do not lead to any of the identified clashes. The %ACC of the randomly generated clash map is obtained by averaging these values over 100 000 such randomly generated samples. Alternatively, these values can be calculated as the ratio of all pairs of non-conserved residue positions that have residues at these positions in the functional hybrids that are both simultaneously retained from either one of the parental sequences to the total number of combinations of such residue pairs.

Glycinamide ribonucleotide transformylase

In this case study we identify all clashing residue pairs for the two single-crossover incremental truncation libraries encoding purN/hGART and hGART/purN hybrids. These hybrids are constructed using purN (209 residues) and hGART (201 residues) sequences whose structures (PDB i.d.: 1GAR, 1MEO, respectively) are obtained from the PDB. Structural alignment of the two structures using the CE method results in a root mean square distance (r.m.s.d.) value of 1.30 Å and a sequence identity of 38.20%. The residue clash map is constructed after identifying all common contacting residues based on the structural alignment. The purN/hGART library includes eight steric clashes (shown as gray arcs in Figure 3a) and five repulsive residue pairs (shown as black arcs), while the hGART/purN library exhibits nine steric clashes, three cases of charge repulsion and one hydrogen bond disruption (shown as a broken arc in Figure 3b).

Lutz *et al*. (2001) generated incremental truncation libraries with crossovers in the sequence window from residues 53 to 144. The functional characterization results are superimposed onto the residue clash map (Figure 3) along with the experimental count of each one of these hybrids. The purN/hGART library includes 68 functional members and as seen in Figure 3a most of the functional crossover positions avoid disrupting any arcs. Note that most functional crossovers fall in the regions between residues 79 and 114 and 120 and 138 that are free of any type of clashes. Out of 68 functional members present in the library, only four involve crossovers (i.e. positions 70 and 144) that disrupt any arcs [i.e. (4, 31)–80 and 140–145, respectively] resulting in 94.12% of functional members being free of predicted clashes (Table I). The hGART/purN library, on the other hand, includes 56 functional members with only one (i.e. crossover position 83) disrupting an arc (i.e. 81–84). Interestingly, most of the crossover positions (82%) in the hGART/purN library are found in the

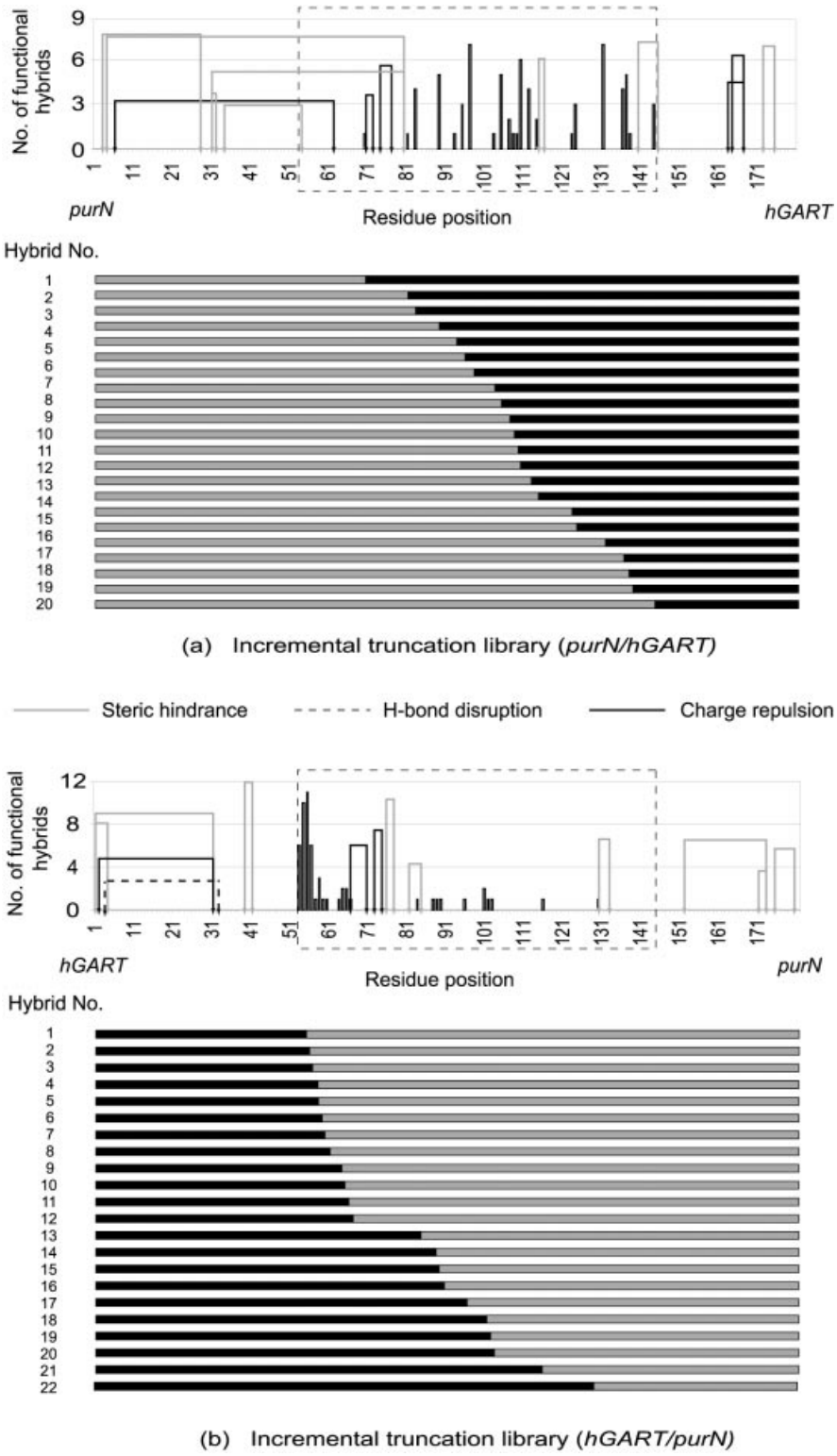

Fig. 3. Different types of clashes for (**a**) purN/hGART and (**b**) hGART/purN are shown as arcs linking the two positions. Functional crossover positions (Lutz *et al*., 2001) are shown as vertical bars whose heights represent their number. Shown below these clash maps are the functional hybrids with the gray region corresponding to *purN* and the black region to *hGART*. Notably, the crossover distribution and directionality in both cases is such that most functional hybrids are free of the identified clashes.

region 53–65, whereas none are observed in this region for the purN/hGART library (Figure 3), alluding to the strong effect of crossover directionality. Note that crossovers generated using ITCHY are uniformly distributed over the desired truncation range (Ostermeier, 2003) without exhibiting any directionality bias. Therefore, we believe that the crossover directionality in hGART/purN versus purN/hGART in region 53–65 is not likely to be due to bias in library generation, but rather an

Table I. Summary of statistical analysis for the five protein families

Protein system	Sequence identity (%)	r.m.s.d.[a] (Å)	Total number of identified clashes	Correct clashes[b]	Residue clash map (RCM)		Residue contact map		Random clashes	$\%ACC^{RCM}/\%ACC^{Rnd}$
					%ACC[c]	%CFC[d]	%ACC	%CFC	%ACC	
GART	38.20	1.30	13[e]	8[e]	61.54	90.91	30.20	0.00	9.74	6.31
GST	85.25	0.50	7	7	100.00	100.00	56.33	0.00	13.10	7.63
β-Lactamase[f]	43.17	1.30	57	46	80.70	35.00	65.00	0.00	14.68	5.50
β-Lactamase[g]	43.17	1.30	57	44	77.19	31.03	62.31	0.00	13.07	5.90
C23O	84.70	0.10	6	6	100.00	100.00	70.86	0.00	25.86	3.87
Dioxygenases	~71.10	–	94	93	98.90	96.80	71.20	9.70	54.08	1.83

[a]Root mean square distances (Å) between the crystal structures of the two parental sequences.
[b]Number of identified clashes absent in the functional hybrid library.
[c]%ACC is defined as the percentage of arcs representing clashes or contact pairs that are not disrupted by the crossover pattern found in the functional hybrid library.
[d]%CFC is defined as the percentage of functional crossovers that do not disrupt any of the arcs representing clashes.
[e]These values are based on clashes found in the region between residues 53 and 144.
[f]These results are based on the crossover data taken from results published by Voight *et al*. (2002).
[g]These results are based on the crossover data published by Hiraga and Arnold (2003).

outcome of the selection pressure. By superimposing the residue clash maps on the corresponding functional hybrid libraries (i.e. purN/hGART and hGART/purN), we find that 61.54% of the predicted clashes are absent in the set of functional hybrids (%ACC) and 90.91% of the functional hybrids included none of our predicted residue clashes (%CFC). In contrast, comparisons of the residue contact map and randomly distributed clashes against the functional library yield much smaller %ACCs of only 30.20 and 9.74%, respectively.

Glutathione S-transferase

The two GST parental sequences (i.e. human Mu class glutathione *S*-transferases, GST M1-1 and M2-2) share a relatively high sequence identity of 84% and align well both at the sequence and structural level. Both sequences are 217 residues in length, and have available structures (PDB i.d.: 1GTU and 2GTU). Even though they share only a 16% difference in the sequence at the protein level, their specific activities with the substrate aminochrome and 2-cyano-1,3-dimethyl-1-nitrosoguanidine (cyanoDMNG) differ by more than 100-fold (Hansson *et al*., 1999a). The chimeric GSTs in the experimental study were modified so that the first 32 bp (~10 amino acids) of each were from GST M1-1 (Figure 4). The two segments vary only at two positions (i.e. 3 and 8) implying that the modified DNA shuffled parental sequences have a slightly increased sequence identity of 85.25% at the protein level. The 20 functional hybrid sequences involving multiple crossovers (Hansson *et al*., 1999b) are shown in Figure 4 with gray denoting fragments retained from GST M1-1 and black denoting fragments from GST M2-2. All recombinant sequences have a number of identical stretches of undetermined parental origin, shown in white. The hybrids are listed in decreasing order of activities with respect to aminochrome and CDNB.

The residue clash map for the GST hybrids is modified to account for multiple crossovers (Figure 4). Each arc in Figure 4 is bicolored to encode the origin of the clashing residues. Therefore, only if the residues joined by an arc originate from the parental sequences with the same color designation as the arc, a clashing interaction is introduced. As shown in Figure 4, we find five cases of charge repulsion corresponding to pairs 91–96 (–/–), 93–91 (+/+), 128–125 (–/–), 129–125 (–/–) and 167–165 (+/+), with the first position retained from GST M1-1 and the second position from GST M2-2. The signs within the parentheses indicate the type of interaction that is present in the hybrid. Steric clashes are found between residues 106 and 107 and 159 and 103 with the first entry of each pair originating from GST M1-1 and the second from GST M2-2. Comparison of our results with the 20 functional hybrids (Hansson *et al*., 1999b) show that most crossover positions in the functional hybrids lie outside the range where clashes are found (i.e. regions 1–90 and 170–217) (Figure 4). Interestingly, even though some crossovers exist between these arcs, their directionality is such that no clash is formed. None of the 20 hybrids contain any predicted clashing pairs resulting in a %ACC of 100%. Residue contact map based and randomly distributed clashes yielded much lower %ACC values of 56.33 and 13.10%, respectively (Table I).

β-Lactamases

Surprisingly, even though the sequence identity between the two β-lactamase parental sequences [PDB i.d.: 1G68 (PSE-4) and 1BTL (TEM-1)] is 43.17%, slightly more than the GART system, the number of identified clashes is significantly higher. The total number of clashes in the TEM-1/PSE-4 directionality is found to be 27 while the reverse directionality involved 30 clashes (Figure 5). Hybrids for both directions contained 14 cases of charge repulsion while the remaining clashes resulted from steric clashes. Crossover sequence data for functional hybrids are taken from the *in vitro* recombination experiments conducted by Voigt *et al*. (2002) where 10 functional hybrids (Figure 5) are reported. These crossovers were generated between residue positions 26 and 290. Notably, by superimposing the residue clash map against the crossover distribution, we find that 80.70% of the predicted clashes share such directionalities so that they are not found in any of the functional members of the library. Figure 5 shows that most of the predicted clashes fall in the range between positions 25 and 125 and are present in only four out of the 19 functional crossovers. On the other hand, residue contact map and random clash distributions yielded much lower %ACC values of only 65.00 and 14.68%, respectively (Table I). Recently, Hiraga and Arnold (Hiraga and Arnold, 2003) published additional

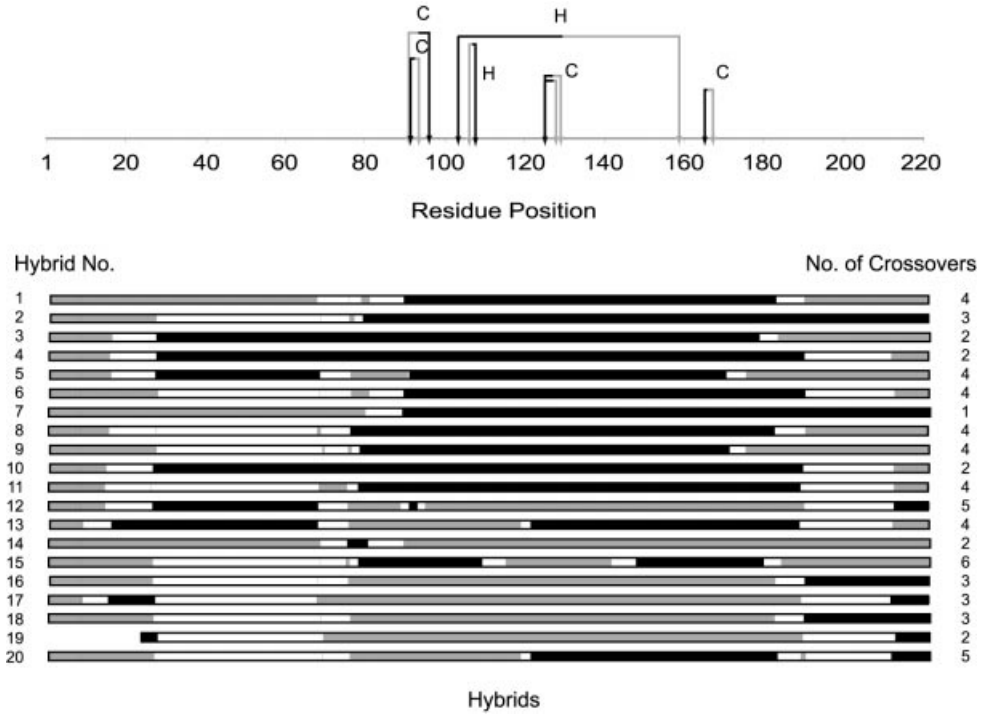

Fig. 4. Residues in the hybrids retained from parental sequences with the same color (gray, GST M1-1; black, GST M2-2) as the arc connecting them, lead to an unfavorable interaction. The arcs indicate steric hindrance (H) or electrostatic repulsion (C) between the two residues. Shown below these arcs are the functional hybrids, constructed using DNA shuffling, of GST M1-1 (gray) and GST M2-2 (black). They are ordered in decreasing ratios of activities with respect to aminochrome and CDNB (Hansson *et al*., 1999b). White segments represent conserved stretches of unknown origin. Numbers to the right of each hybrid indicate the number of crossovers.

crossover results for functional β-lactamase hybrids constructed using SISDC. These new data were also compared with the predicted clash map shown in Figure 5 and the results of these comparisons are summarized in Table I.

Catechol-2,3-oxygenase

Kikuchi *et al*. (2000) obtained seven thermally stable hybrids using single-stranded DNA shuffling on the parental sequences *xylE* (catechol-2,3-dioxygenase from *Pseudomonas putida*, PDB i.d.: 1MPY) and *nahH* (synthetic construct). Because no structure is currently available for *nahH*, we used an estimated structure obtained using Swiss-Model (Peitsch, 1996) with the structure of *nahH* (IMPY) as the template. This was subsequently used to obtain the structural alignment using the CE method (Shindyalov and Bourne, 1998). The two sequences share 84.7% sequence identity at the protein level. A total of six clashes are identified for both directions, all of which resulted from electrostatic repulsion (Figure 6). Five of these have *xylE/nahH* directionality [79–80 (+/+), 82–83 (–/–), 183–184 (–/–), 183–286 (–/–) and 285–286 (–/–)] and only one with *nahH/xylE* directionality [80–83 (+/+)]. The residue clash map identified three clashes located in the region around residue 80 which is the region retained from the same parental sequence in all of the hybrids, thus, preventing the formation of clashes. Interestingly, all the functional hybrids in the library have different parental origins for the contacting residue pair 183–286; however, none have *xylE/nahH* directionality, thus avoiding the charge clash that could be formed in the hybrids with reverse (*xylE/nahH*) directionality (Table I).

Dioxygenases

All four protein systems analyzed so far included hybrids constructed from two parental sequences. The dioxygenase hybrids involve three parental sequences and have a relatively higher number of crossovers per sequence. The active library was created (Joern *et al*., 2002) by recombining the α and β subunits of toluene dioxygenase (*todC1C2*), tetrachlorobenzene dioxygenase (*tecA1A2*) and biphenyl dioxygenase (*bhpA1A2*). *tod* and *tec* are 89.16% identical at the protein level. The *bhp* sequence is less similar, exhibiting 62.30 and 61.85% pairwise sequence identity with *tec* and *tod*, respectively. No structures are available for any of these protein sequences, thus an estimated structure for each one of them is used. The dimeric state of the dioxygenases requires the use of Swiss-Model in Optimize mode (Schwede *et al*., 2003) for structure prediction. Naphthalene dioxygenase (PDB i.d.: 1O7G), a distant homolog of the three dioxygenases was found using the ExPDB database (Schwede *et al*., 2000) and was used as the template. Figure 7 shows the clash maps for the three different sequence combinations (i.e. *tec-tod*, *tod-bhp* and *bhp-tec*) contrasted against the eight active clones with one to eight crossovers per sequence. Comparisons of these results are summarized in Table II. A total of 94 clashes are identified of which 94.68% result from the *tod-bhp* and *bhp-tec* combinations alone, a consequence of low sequence identity between these sequences. Notably, out of the 94 identified clashes only one clash is present in the hybrids [arising from charge repulsion (+/+) between residues 13 and 385 with a *tec-bhp* directionality] resulting in a high %ACC of 98.9% and a %CFC of 96.8%. Alternatively, we calculated a total of 3685 non-conserved contacting residues with different parental origins using the estimated structures out of which 84.42% result from the *tod-bhp* and *bhp-tec* combinations. Of these contacts, 1063 are found to be present in the active hybrids, resulting in %ACC and %CFC values of 71.2 and 9.7%, respectively (Table I).

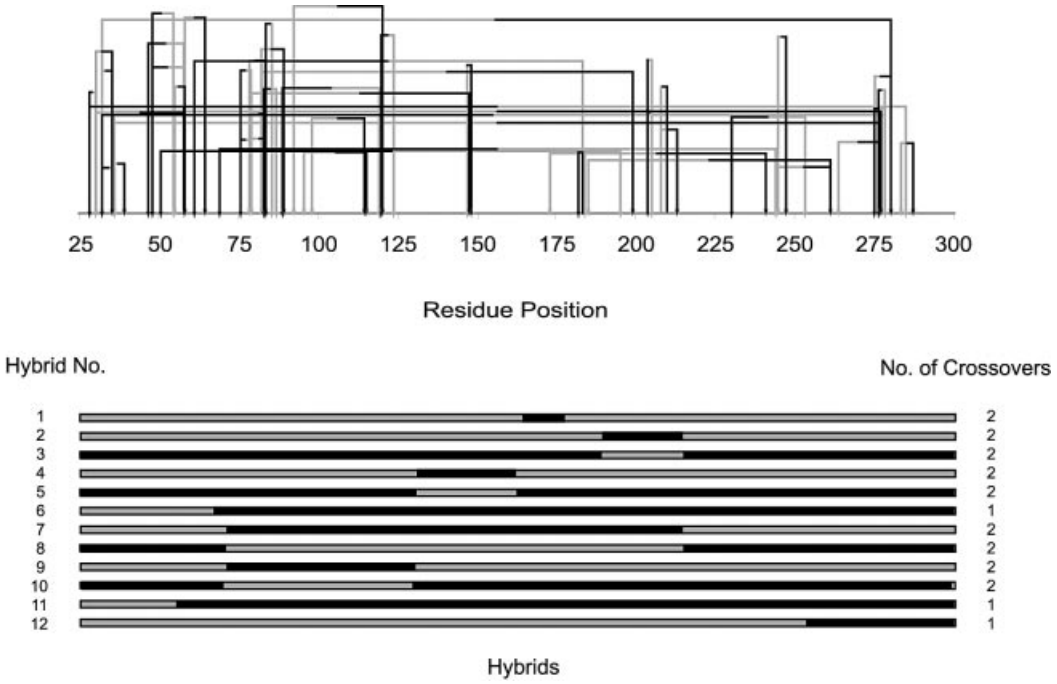

Fig. 5. The identified residue clashes are shown against the 10 active β-lactamase [TEM-1 (black), PSE-4 (gray)] hybrids identified experimentally (Voigt *et al*., 2002). The total number of clashes in the TEM-1/PSE-4 directionality is found to be 27 while the reverse directionality has 30 clashes. Hybrids in either directionality contain 14 cases of charge repulsion while the remaining resulted from steric clashes.

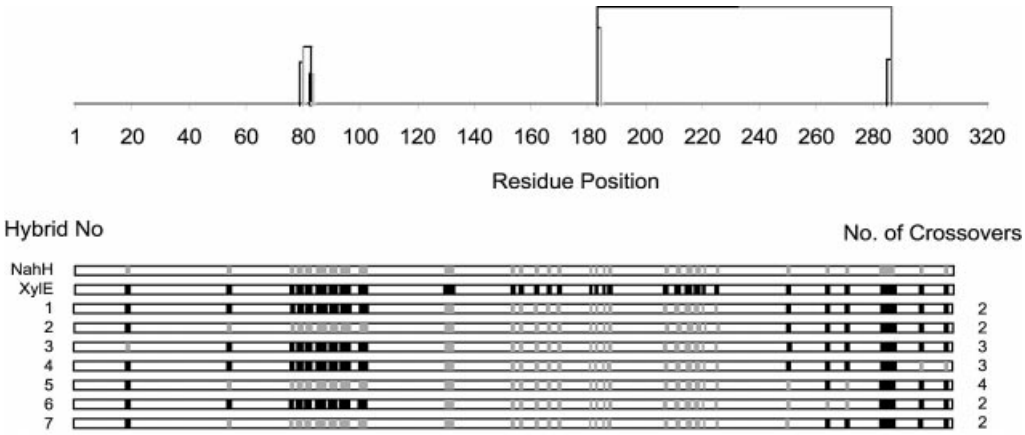

Fig. 6. Seven different thermally stable C23O hybrids obtained by shuffling ssDNA are shown above (Kikuchi *et al*., 2000). The residues derived from *NahH* and *XylE* are shown in gray and black, respectively, while conserved residue positions of ambiguous origin are colored white. Only six clashes, all of which result from charge repulsion, are identified.

Discussion

In this paper, we introduced a rapid procedure for checking for three different types of clashes (i.e. electrostatic repulsion, steric hindrance, cavity formation, and hydrogen bond disruption) that could be introduced in protein hybrids. This approach was used to identify clashes between contacting residue pairs of the hybrids that have different parental origins for a number of experimental systems. The identified clashing residue pairs between pairs of parental proteins were then contrasted against functionally characterized hybrid libraries. Results of these comparisons, summarized in Table I, show that the patterns of identified clashing residue pairs are consistent with experimentally found patterns of functional crossover combinations. The clash map *p*-values (i.e. the fraction of randomly generated clash maps with %ACC greater than or equal to an observed value) were computed for some of the systems. A sample of 100 000 randomly generated clash maps was used with the average number of clashes in each sample equal to those predicted for that particular system. These *p*-values were found to be in the order of 10^{-2}–10^{-3}, implying that the predictions are statistically meaningful.

Note also that we find that the residue clash maps are on average 1.55 times more specific (i.e. ratio of %ACCs) than residue contact maps and 5.03 times more specific than

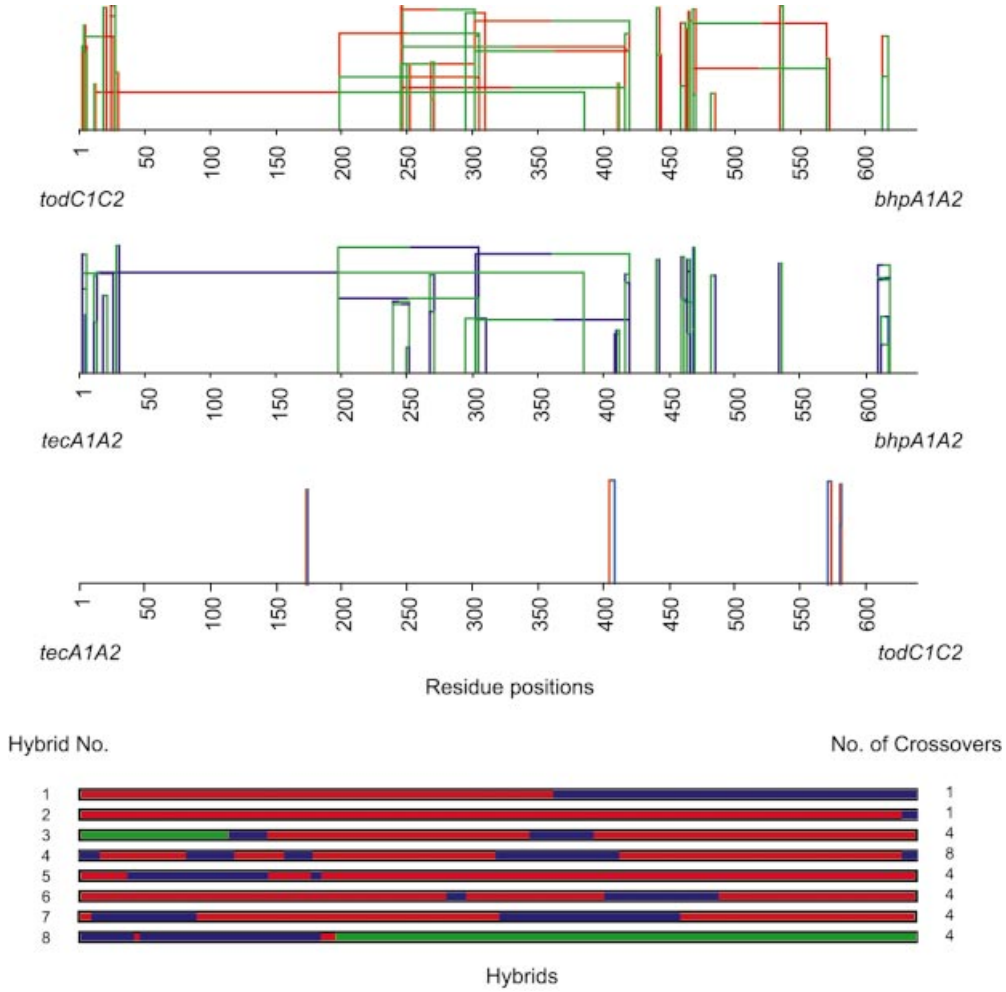

Fig. 7. Eight toluene-active members of the hybrid library obtained by shuffling genes encoding the α and β subunits of three dioxygenases are shown as horizontal bars (Joern *et al*., 2002). Sequence elements from *tecA1A2*, *todC1C2* and *bhpA1A2* are colored blue, red and green, respectively. Shown above these are the clash maps corresponding to the three different sequence combinations (i.e. *tod-bhp*, *tec-bhp* and *tod-tec*) whose details are given in Table II.

Table II. Clash map based analysis for the dioxygenase system

Clash map based analysis for the dioxygenase system.		
Crossover type	Total number of clashes (see Figure 7)	Clashes present in hybrids 1-8.
bhp-tod	26	0
bhp-tec	21	0
tod-bhp	21	0
tec-bhp	21	1
tec-tod	2	0
tod-tec	3	0

randomly generated clashes at explaining observed functional crossovers. While residue contact maps do capture some information on residue pairs that result in unfavorable interaction in the hybrids, not all disrupted contact pairs are detrimental to functionality. The proposed residue clash map improves prediction by filtering out many of the incorrectly predicted pairs. The clash map categorizes these clashes into three distinct types (i.e. electrostatic repulsion, steric clash and hydrogen bond disruption). By pinpointing the cause of these clashes one can then perform site-directed mutagenesis to ameliorate clashes by replacing problematic residues with ones that do not form any clashes. Admittedly, the residue clash map does not account for the possibility of relieving some of the identified clashes through side chain and/or backbone movement. This simplification is reflected in the results as the accuracy in crossover classification is reduced as the sequence identity and thus similarity between the parental sequences is reduced (Table I). Therefore, some of the residues that are in contact in the parental sequences may not necessarily remain in contact in the hybrid, thus relieving some of the predicted clashes. Alternatively, new clashes may be introduced due to new contacts formed or altered side chain conformations. Nevertheless, the proposed approach enables the rapid prescreening of an entire protein family for revealing favorable recombination partners that can subsequently be analyzed by more detailed molecular modeling methods that capture side

chain and backbone movement. So far the clash map based method can only classify hybrids as functional or non-functional but cannot rank hybrids with respect to their activity. We are currently developing methods for overcoming this limitation by ranking the hybrids with respect to their activity based on the identified clashes.

Acknowledgements

We thank Professor Stephen Benkovic and Gregory Moore for helpful discussions and the reviewers for many useful suggestions. Financial support from National Science Foundation Award BES0331047 and hardware support from the IBM-SUR program are gratefully acknowledged.

References

Agarwal,P.K., Billeter,S.R., Rajagopalan,P.T., Benkovic,S.J. and Hammes-Schiffer,S. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 2794–2799.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Bogarad,L.D. and Deem,M.W. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 2591–2595.
Chakrabarti,P. (1993) *J. Mol. Biol.*, **234**, 463–482.
Copley,R.R. and Barton,G.J. (1994) *J. Mol. Biol.*, **242**, 321–329.
Dupraz,P., Oertle,S., Meric,C., Damay,P. and Spahr,P.F. (1990) *J. Virol.*, **64**, 4978–4987.
Fischer,D., Wolfson,H., Lin,S.L. and Nussinov,R. (1994) *Protein Sci.*, **3**, 769–778.
Glusker,J.P. (1991) *Adv. Protein Chem.*, **42**, 1–76.
Gobel,U., Sander,C., Schneider,R. and Valencia,A. (1994) *Proteins*, **18**, 309–317.
Hansson,L.O., Bolton-Grob,R., Massoud,T. and Mannervik,B. (1999a) *J. Mol. Biol.*, **287**, 265–276.
Hansson,L.O., Bolton-Grob,R., Widersten,M. and Mannervik,B. (1999b) *Protein Sci.*, **8**, 2742–2750.
Hiraga,K. and Arnold,F.H. (2003) *J. Mol. Biol.*, **330**, 287–296.
Joern,J.M., Meinhold,P. and Arnold,F.H. (2002) *J. Mol. Biol.*, **316**, 643–656.
Kikuchi,M., Ohnishi,K. and Harayama,S. (2000) *Gene*, **243**, 133–137.
Loll,B., Raszewski,G., Saenger,W. and Biesiadka,J. (2003) *J. Mol. Biol.*, **328**, 737–747.
Lutz,S., Ostermeier,M., Moore,G.L., Maranas,C.D. and Benkovic,S.J. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 11248–11253.
Meyer,M.M., Silberg,J.J., Voigt,C.A., Endelman,J.B., Mayo,S.L., Wang,Z.G. and Arnold,F.H. (2003) *Protein Sci.*, **12**, 1686–1693.
Moore,G.L. and Maranas,C.D. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 5091–5096.
Moore,J.C., Jin,H.M., Kuchner,O. and Arnold,F.H. (1997) *J. Mol. Biol.*, **272**, 336–347.
Munson,M., Balasubramanian,S., Fleming,K.G., Nagi,A.D., O'Brien,R., Sturtevant,J.M. and Regan,L. (1996) *Protein Sci.*, **5**, 1584–1593.
Oldfield,T.J. (2002) *Proteins*, **49**, 510–528.
Ostermeier,M. (2003) *Biotechnol. Bioeng.*, **82**, 564–577.
Ostermeier,M., Nixon,A.E., Shim,J.H. and Benkovic,S.J. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 3562–3567.
Peitsch,M.C. (1996) *Biochem. Soc. Trans*, **24**, 274–279.
Ratnaparkhi,G.S. and Varadarajan,R. (2000) *Biochemistry*, **39**, 12365–12374.
Saraf,M.C., Moore,G.L. and Maranas,C.D. (2003) *Protein Eng.*, **16**, 397–406.
Schwede,T., Diemand,A., Guex,N. and Peitsch,M.C. (2000) *Res. Microbiol.*, **151**, 107–112.
Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C. (2003) *Nucleic Acids Res.*, **31**, 3381–3385.
Shindyalov,I.N. and Bourne,P.E. (1998) *Protein Eng.*, **11**, 739–747.
Sieber,V., Martinez,C.A. and Arnold,F.H. (2001) *Nat. Biotechnol.*, **19**, 456–460.
Song,K.S., Park,Y.S., Choi,J.R., Kim,H.K. and Park,Q. (1999) *Exp. Mol. Med.*, **31**, 47–51.
van Gunsteren,W.F., Billeter,S.R., Eising,A.A., Hünenberger,P.H., Krüger,P.K., Mark,A.E., Scott,W.R.P. and Tironi,I.G. (1996) *Biomolecular Simulations: The GROMOS96 Manual and User Guide*. Verlag der Fachvereine, Zurich, pp. 1–1024.
Voigt,C.A., Mayo,S.L., Arnold,F.H. and Wang,Z.G. (2001) *J. Cell. Biochem. Suppl.*, **37**, 58–63.
Voigt,C.A., Martinez,C., Wang,Z.G., Mayo,S.L. and Arnold,F.H. (2002) *Nat. Struct. Biol.*, **9**, 553–558.
Vriend,G. (1990) *J. Mol. Graph.*, **8**, 52–56.
Wallace,A.C., Borkakoti,N. and Thornton,J.M. (1997) *Protein Sci.*, **6**, 2308–2323.
Wang,P.L. (2000) *Dis. Markers*, **16**, 3–13.
Westbrook,J. *et al.* (2002) *Nucleic Acids Res.*, **30**, 245–248.

Received August 1, 2003; revised October 21, 2003; accepted October 23, 2003