Design of Combinatorial Protein Libraries of Optimal Size

Manish C. Saraf, Anshuman Gupta, and Costas D. Maranas*
Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania

ABSTRACT **In this article we introduce a computational procedure, OPTCOMB (Optimal Pattern of Tiling for COMBinatorial library design), for designing protein hybrid libraries that optimally balance library size with quality. The proposed procedure is directly applicable to oligonucleotide ligation-based protocols such as GeneReassembly, DHR, SISDC, and many more. Given a set of parental sequences and the size ranges of the parental sequence fragments, OPTCOMB determines the optimal junction points (i.e., crossover positions) and the fragment contributing parental sequences at each one of the junction points. By rationally selecting the junction points and the contributing parental sequences, the number of clashes (i.e., unfavorable interactions) in the library is systematically minimized with the aim of improving the overall library quality. Using OPTCOMB, hybrid libraries containing fragments from three different dihydrofolate reductase sequences (*Escherichia coli, Bacillus subtilis,* and *Lactobacillus casei*) are computationally designed. Notably, we find that there exists an optimal library size when both the number of clashes between the fragments composing the library and the average number of clashes per hybrid in the library are minimized. Results reveal that the best library designs typically involve complex tiling patterns of parental segments of unequal size hard to infer without relying on computational means. Proteins 2005;60:769–777.** © 2005 Wiley-Liss, Inc.

Key words: combinatorial protein libraries; residue–residue clashes; protein engineering; directed evolution protocols; optimal library design

INTRODUCTION

The directed evolution of variants of a single gene[1] or a family of genes[2] coupled with a screening or selection step has emerged as a dominant strategy for creating proteins with improved or novel properties.[3] Recent developments in methods for directed evolution have led to new approaches[4–6] for creating diverse combinatorial libraries with tunable statistics irrespective of sequence homology. Two of these methods, GeneReassembly[4] and Degenerate Homoduplex Recombination (DHR),[5] use synthesized degenerate oligonucleotides for tailoring the diversity of a library. These oligonucleotides are designed to include coding information for the polymorphisms present in the parental set, while also including "customized" sequence identity at predetermined locations enabling annealing-based recombination. The "customized" sequence identity enables the targeted introduction of crossovers at only desired positions. Alternatively, in sequence-independent site-directed chimeragenesis (SISDC),[6] the exact location of crossovers is predetermined by the use of marker tags for endonuclease recognition. These are two examples out of many currently available protocols that are capable of creating the desired level and type of diversity in a combinatorial library.[3]

Despite these developments, protein engineering remains a formidable task because it is still unclear what should the optimal level and type of diversity be for sampling the sequence space spanned by the parental sequence set.[7,8] Most proteins in nature exhibit complex networks of dynamic interaction for function.[9–12] Therefore, a large number of crossovers between parental sequences is likely to disrupt vital interactions [13–16] rendering most hybrids nonfunctional. In fact, it is commonly observed that the average activity of a library tends to drop off as parental sequence similarity decreases.[1,8] On the other hand, a combinatorial library generated by introducing only a few crossovers will sample only a very small portion of sequence space by retaining many large contiguous parental sequence stretches. Therefore, a key open challenge is how to *a priori* identify the optimal design of a library. This entails the identification of (1) the optimal library size, (2) number and location of junction points, and (3) the parental sequences that contribute a fragment at each one of the junction points (see Fig. 1).

A number of strategies have been developed to assess the quality of a library based on sequence and/or structural information encoded within the parental/family sequences to guide the design of combinatorial libraries.[13–16] Typically, this involves the definition of a scoring metric for evaluating the fitness of hybrid protein sequences against the parental sequences. This concept was pioneered with the development of SCHEMA algorithm[16] that hypothesizes that structural disruptions are introduced when a contacting residue pair in a hybrid has differing parental origins. Hybrids are scored for stability

Grant sponsor: National Science Foundation; Grant number: BES0331047.

*Correspondence to: Costas D. Maranas, Department of Chemical Engineering, The Pennsylvania State University 112 Fenske Laboratory, University Park, PA 16082. E-mail: costas@psu.edu

Received 3 October 2004; Revised 27 December 2004; Accepted 10 January 2005

Published online 6 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20490

© 2005 WILEY-LISS, INC.

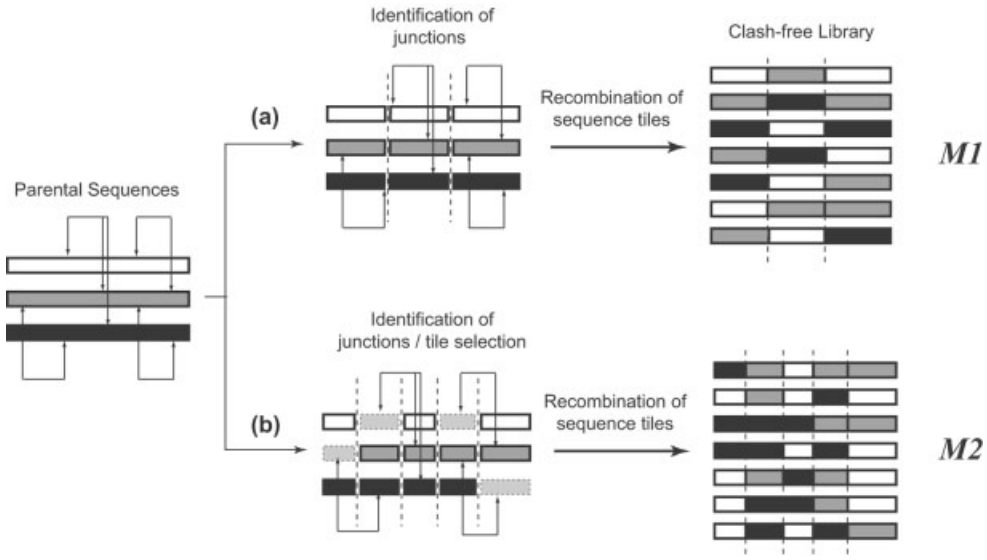

Fig. 1. A pictorial representation of the example where three parental sequences form a combinatorial library through recombination. The clashes between different residues are shown as double-headed arrows. The junction points are shown as dashed lines. The combinatorial libraries are designed using two different design rules: (**a**) all parental sequences contribute fragments at each of the junction points, and (**b**) selective restrictions are imposed on the set of oligomers being contributed by the parents.

by counting the number of disruptions.[17,18] Recently, a dynamic programming algorithm was proposed[18] that identifies the location of junction points that minimize SCHEMA disruption without allowing for parental fragment skipping. Alternatively, a number of methods have been developed in our group based on (1) mean-field energy calculations to infer correlations in substitution patterns (SIRCH[15]), (2) pinpointing property value deviations (i.e., charge, volume, and hydrophobicity) from parental sequences,[14] and (3) family sequence statistics for clash identification (FamClash[13]). Comparisons with experimental studies[13,14,17] have shown that crossovers are indeed preferentially allocated to avoid the predicted clashes among functional hybrids. Interestingly, using FamClash[13] we demonstrated in one case that hybrid activity levels were inversely proportional to the number of clashes in these hybrids.

These methods hint at a design strategy that forms the basis for the computational design procedure OPTCOMB introduced in this article. OPTCOMB pinpoints the location of junctions between fragments as well as their sizes and their parental origins such that the number of clashes between the fragments constituting the library is minimized. Two optimization models are considered abstracting two classes of experimental strategies for combinatorial library generation: (i) no restrictions are imposed on the contributing parental sequences [e.g., SISDC; see Fig. 1(a)], and (ii) restrictions are imposed on the set of oligomers being contributed by the parental sequences in certain locations [e.g., DHR and GeneReassembly; see Fig. 1(b)]. Both optimization models are tested on the computational design of a combinatorial library formed by three dihydrofolate reductase (DHFR) sequences from *E. coli, B. subtilis*, and *L. casei*.

MATHEMATICAL MODELING

The design of a combinatorial library entails a number of discrete decisions such as (1) the placement and the number of junction points to be selected, (2) whether or not a given position along the sequence is a junction point, and (3) if a particular parental sequence contributes a fragment/oligomer at a given junction point. To model these decisions, the OPTCOMB optimization models draw upon mixed-integer linear programming formulations that use binary variables to mathematically represent these discrete decisions. These binary variables act as on/off switches that encode, for instance, the presence/absence of a junction point. The OPTCOMB procedure makes use of models ***M1*** and ***M2*** corresponding to the experimental setups illustrated by Figure 1(a) and (b), respectively. Specifically, model ***M1*** abstracts experimental protocols where all parental sequences contribute a fragment at each one of the junction points. The design variables are binary variables that denote the presence or absence of a junction point along the sequence. On the other hand, model ***M2*** abstracts experimental protocols where "skipping" of parental fragments is permitted. Additional design variables are included in the model to account for whether or not a particular parental sequence contributes a fragment at a junction point. In both ***M1*** and ***M2***, the design variables are adjusted such that the total number of clashing residue pairs between fragments that constitute the library is minimized. These clashes can be identified using many available computational approaches.[13–16]

In addition to the constraints included in the two models that penalize the simultaneous selection of clash forming fragments, other constraints can be included to impose additional requirements. For example, such requirements

may include the preservation of two or more residues to ensure that crucial interactions for catalysis or binding with external molecules are retained.[19,20] Constraints can also be included to guide the selection of junction points based on user-defined requirements. For example, constraints can be used to direct selection of junction points within loop regions[21] so that structural elements (i.e, α-helices, β-sheets, etc.) are not disrupted enabling the swapping of low energy secondary structures.[22] In addition, constraints can be incorporated to minimize bias in family DNA shuffling so that each of the parental sequences contributes a similar number of fragments/oligomers to the library or alternatively to restrict crossover positions to regions of high-sequence identity for proper ligation. The inclusion of such constraints in the current implementation, although not explicitly covered here, is quite straightforward.

The simpler model ***M1*** is applicable when no restrictions are imposed on the contributing parental sequences [Fig. 1(a)]. The only design variables whose values need to be determined are the locations of junction points. The sets, parameters, and variables used in model ***M1*** are described below.

Sets:

$k, k_1, k_2 \in \{1,2,\ldots,K\}$ = set of parental sequences
$i, i_1, i_2 \in \{1,2,\ldots,I\}$ = set of aligned positions

Parameters:

N = Number of oligomers
$L_{\min}$ = Length of shortest allowable oligomer
$L_{\max}$ = Length of longest allowable oligomer
$C_{i_1 i_2}^{k_1 k_2}$ = 1 if a clash exists between residue i_1 of parental sequence k_1 and residue i_2 of parental sequence k_2; $i_1 < i_2$; $k_1 \neq k_2$
= 0 otherwise

Variables:

Y_i = 1 if an oligomer starts at position i (i.e., a junction point)
= 0 otherwise

$Z_{i_1 i_2}$ = 1 if there exists at least one pair of parental sequences for which there is a clash between residues at positions i_1 and i_2.
= 0 otherwise

Note that here the values assigned to parameters $C_{i_1 i_2}^{k_1 k_2}$ are either 1 or 0, depending on whether there exists a clash between the two residues. Alternatively, continuous values (e.g., between 0 and 1) that quantify the severity of the clashes could also be used. Based on the above defined sets, parameters, and variables, the model ***M1*** of OPTCOMB yields an optimization problem implemented as the following mixed-integer linear programming (MILP) formulation.

$$\underset{Y_i \in \{0,1\}}{\text{minimize}} \sum_{i_1=1}^{I} \sum_{\substack{i_2=i \\ i_2>i_1}}^{I} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} Z_{i_1 i_2} \cdot C_{i_1 i_2}^{k_1 k_2} \quad (1)$$

$$\sum_{i=1}^{I} Y_i \geq N \quad (2)$$

$$\sum_{i'=i}^{i+L_{\min}-1} Y_{i'} \leq 1, \qquad \forall\, i = 1,2,\ldots,I - L_{\min}+1 \quad (3)$$

$$\sum_{i'=i}^{i+L_{\max}-1} Y_{i'} \geq 1, \qquad \forall\, i = 1,2,\ldots,I - L_{\max} + 1 \quad (4)$$

$$Z_{i_1 i_2} \leq \sum_{i=i_1+1}^{i_2} Y_i, \qquad \forall\, (i_1, i_2 > i_1, k_1, k_2) \text{ such that } C_{i_1 i_2}^{k_1 k_2} = 1 \quad (5)$$

$$Z_{i_1 i_2} \geq Y_i, \qquad \forall\, (i_1,\ i_2 > i_1, k_1, k_2) \text{ and } i = i_1 + 1, i_1 + 2, \ldots, i_2 \text{ such that } C_{i_1 i_2}^{k_1 k_2} = 1 \quad (6)$$

$$\sum_{I-L_{\max}+1}^{I-L_{\min}+1} Y_i = 1 \quad (7)$$

$$0 \leq Z_{i_1 i_2}, \leq 1; \quad Y_{i=1} = 1 \quad (8)$$

The objective function [Eq. (1)] of model ***M1*** entails the minimization of the number of clashes between oligomers/fragments selected for library design. Constraint 2 ensures that the number of oligomers present is greater than or equal to some specified target, thus establishing the library size. The lower and upper bounds on the lengths of all oligomers is enforced by constraints 3 and 4, respectively. Typically, these lengths are determined based on the specifics of the ligation protocol [e.g., GeneReassembly (39–60 nucleotides or 13–20 amino acids),[4] DHR (54–72 nucleotides or 18–24 amino acids[5])]. Note that the oligomer size ranges ($L_{\min}$, $L_{\max}$) determine the range of values that N can take, and therefore indirectly determine the library size. For a given value of $L_{\min}$ and $L_{\max}$, the values of N can range between $N_{\min} = K \times \lfloor I/L_{\max} \rfloor$ and $N_{\max} = K \times \lfloor I/L_{\min} \rfloor$, where $\lfloor \bullet \rfloor$ corresponds to the floor function. Therefore, the library size will range between $K^{N\min/K}$ and $K^{N\max/K}$. Clearly, as the oligomer sizes reduce, the parental sequences can be divided into larger number of fragments allowing a larger number of combinations of these fragments to be available for the construction of hybrids. Equation (5) in conjunction with Equation (6) determines whether a clash is formed between any two positions (i_1, i_2) of the selected fragments from parents (k_1, k_2). Equation (7) ensures that the last fragment of each parental sequence falls within the allowable range of fragment lengths.

Note that in model ***M1***, the included constraints ensure that all parental sequences must contribute a fragment at all junction points without skipping. Therefore, the only means of clash relief is the judicious selection of junction

points such that the minimum number of clashes is formed while ensuring that minimum and maximum fragment size limits are satisfied. Alternatively, model ***M2*** allows for more flexibility as it accounts for the "skipping" of parental fragments. Clashes are relieved based on the selection of crossover positions and also on the choice of parental fragments at each one of the junction points [Fig. 1(b)]. This additional complexity requires additional variables and constraints to capture information on the selection/rejection of fragments of different parental sequences at each one of the junction points. Note that by restricting the contributing parents at each one of the junction points many more clashes can be relieved for the same number of junction points. Model ***M2*** retains all the variables defined for model ***M1*** in addition to the following new ones:

New variables:

y_{ik} = 1 if a new oligomer starts at position i for parent k
= 0 otherwise
Y_i = 1 if a new oligomer starts at position i for at least one parent
= 0 otherwise
$Z_{i_1 i_2}^{k_1 k_2}$ = 1 if residues i_1 of parent k_1 and i_2 if parent k_2 are selected and $C_{i_1 i_2}^{k_1 k_2} = 1$
= 0 otherwise

$$\underset{y_{ik}, Y_i \in \{0,1\}}{\text{minimize}} \sum_{i_1=1}^{I} \sum_{\substack{i_2=1 \\ i_2 > i_1}}^{I} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} Z_{i_1 i_2}^{k_1 k_2} \cdot C_{i_1 i_2}^{k_1 k_2} \tag{9}$$

$$\sum_{k=1}^{K} \sum_{i=1}^{I} y_{ik} \geq N \tag{10}$$

$$\sum_{i'=i}^{i+L_{\min}-1} Y_{i'} \leq 1, \forall \qquad i = 1,2,\ldots,I - L_{\min} + 1 \tag{11}$$

$$\sum_{i'=i}^{i+L_{\max}-1} Y_{i'} \geq 1, \forall \qquad i = 1,2,\ldots,I - L_{\max} + 1 \tag{12}$$

$$Y_i \geq y_{ik}, \forall \qquad i = 1,2,\ldots,I \text{ and } k = 1,2,\ldots,K \tag{13}$$

$$Y_i \leq \sum_{k=1}^{K} y_{ik}, \forall \qquad i = 1,2,\ldots,I \tag{14}$$

$$Z_{i_1 i_2}^{k_1 k_2} \leq \sum_{i=i_1+1}^{i_2} y_{ik_1} \cdot y_{ik_2}, \qquad \forall\, (i_1, i_2 > i_1, k_1, k_2) \text{ such that } C_{i_1 i_2}^{k_1 k_2} = 1 \tag{15}$$

$$Z_{i_1 i_2}^{k_1 k_2} \geq y_{ik_1} \cdot y_{ik_2}, \qquad \forall\, (i_1, i_2 > i_1, k_1, k_2) \text{ and } i = i_1 + 1, i_1 + 2, \ldots, i_2 \text{ such that } C_{i_1 i_2}^{k_1 k_2} = 1 \tag{16}$$

$$\sum_{I - L_{\max}+1}^{I - L_{\min}+1} Y_i = 1 \tag{17}$$

$$0 \leq Z_{i_1 i_2}^{k_1 k_2} \leq 1; \qquad y_{i=1,k} = 1 \tag{18}$$

Note that Equations (15) and (16) involve the product of binary variables. This is linearized by introducing a new set of variables $w_{ik_1k_2} = y_{ik_1} \cdot y_{ik_2}$ to exactly recast the product as a set of linear constraints:[23]

$$\begin{gathered} y_{ik_1} \cdot y_{ik_2} = w_{ik_1k_2} \\ w_{ik_1k_2} \leq y_{ik_1}; \quad w_{ik_1k_2} \leq y_{ik_2}; \quad w_{ik_1k_2} \geq y_{ik_1} + y_{ik_2} - 1; \\ 0 \leq w_{ik_1k_2} \leq 1 \end{gathered} \tag{19}$$

The objective function [Eq. (9)] entails the minimization of the number of clashes between fragments that constitute the library. Equation (10) ensures that the total number of oligomers selected for library design is greater than some specified lower bound. The lower and upper bounds on the lengths of all oligomers is enforced by constraints 11 and 12, respectively. Equation (13) ensures the presence of a junction point if a particular parent contributes a fragment starting at that position. Equation (14) ensures that at least one parental sequence contributes a fragment at any given junction point. Equation (15) in conjunction with Equation (16) determines whether a clash is formed between any two positions (i_1, i_2) of the selected fragments (k_1, k_2). Finally, Equation (17) ensures that the length of the last segment of each parental sequence falls between $L_{\min}$ and $L_{\max}$.

The solution of the OPTCOMB models (***M1*** or ***M2***) provides the complete design of the combinatorial library of a given specified size that minimizes the presence of clashes. By successively varying the number of junction points or fragments (N), a tradeoff curve between library size and percent of clash-free variants (or the average number of clashes per hybrid) can be generated. This curve provides a systematic way for determining the optimal library size given the set of parental sequences and the residue clash map. Note that in this study we have used the percent of clash-free hybrids in a library as a surrogate measure of library quality. However, the OPTCOMB model can also be used for cases where the metric of library quality is different. In such a case, the objective and scoring ($C_{i_1 i_2}^{k_1 k_2}$ functions will need to be appropriately defined. For example, when the metric of quality is the average stability of the library, the scoring function (or the number of clashes here) should be a descriptor of stability[14] rather than of activity.

RESULTS

The optimal tradeoff between library size and clashes is examined using OPTCOMB for combinatorial libraries composed of the well studied dihydrofolate reductase (DHFR) proteins from *E. coli*, *B. subtilis*, and *L. casei*. Clashes between residues of different parental sequences are first derived using the FamClash[13] procedure. According to the FamClash procedure clashes occur when a statistically significant deviation in the properties (such as charge, volume, and hydrophobicity) of pairs of residues in the hybrids are observed from the values observed in the protein family.[13] Similar results are observed when clashes

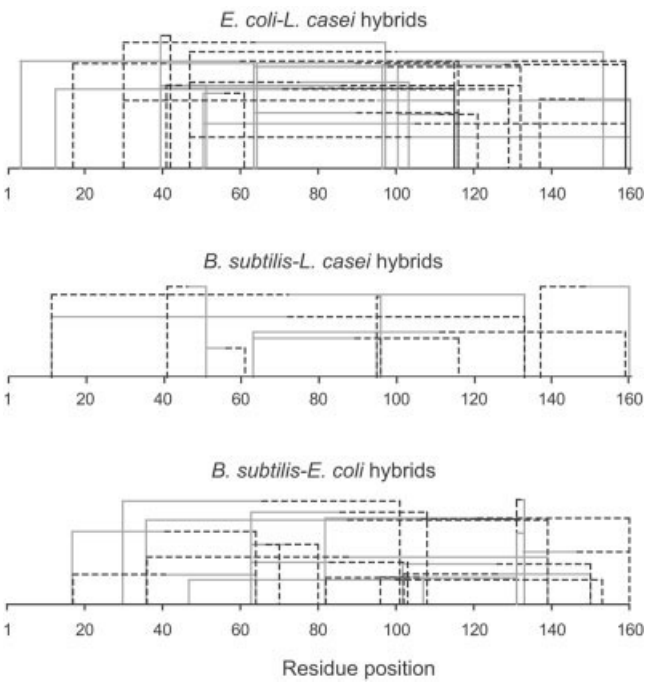

Fig. 2. Clash maps determined using the FamClash procedure[13] corresponding to the three different sequence combinations [*E. coli–L. casei* (black-gray), *B. subtilis–L. casei* (black-gray), and *B. subtilis–E. coli* (black-gray)]. Note that the color shown in the parentheses alongside each pair of sequences correspond to the corresponding pair of parental sequences. Residues in the hybrids retained from parental sequences with the same color as the arc connecting them lead to a clash.

are identified based on steric hindrance, charge repulsion, and hydrogen bond disruption.[14] The DHFR protein family sequence data required for clash prediction is downloaded from the PFAM[24] database including 300 sequences in total. Out of the total 50 clashes identified, 20 clashes are between *E. coli–B. subtilis* (sequence identity = 44.0%), 9 clashes are between *B. subtilis–L. casei* (sequence identity = 36.10%), and 21 clashes are between *L. casei–E. coli* (sequence identity = 28.4%) sequence pairs (see Fig. 2). Notably, most of the clashes (41 out of 50) are associated with the *E. coli* sequence even though it is not the most divergent of the three sequences. These clashes are encoded using the $C_{i_1 i_2}^{k_1 k_2}$ parameters and imported into the OPTCOMB procedure to guide the design of the combinatorial library. The OPTCOMB optimization models (***M1*** and ***M2***) are solved using the CPLEX solver[25] accessed via the GAMS[26] modeling environment. This computational base enables us to explore the following questions:

1. How many clashes remain in the combinatorial library designs obtained using models ***M1*** and ***M2*** as a function of library size and how does this number compare with randomly generated libraries?
2. What are the oligomer/fragment tiling characteristics of the optimally designed libraries?
3. Is there an optimal library size that leads to a minimum of retained clashes per hybrid?
4. What is the effect of library size on the relative contribution of fragments by the three parental sequences, clash distribution, and the tiling combinations?

To answer the first question, model (***M1*** and ***M2***) driven designs are first contrasted against randomly generated libraries to assess whether the systematic selection of junction points affords significant gains over random choices. The optimal designs obtained using models ***M1*** and ***M2*** are also compared against each other to infer the extent of improvement achieved by disallowing fragments from participating in library design. Both OPTCOMB models (***M1*** and ***M2***) are solved for different values of N (number of oligomers) allowing for a minimum and maximum oligomer length of 15 and 30 residues, respectively, covering the range of length of oligonucleotides used in the GeneReassembly and DHR protocols.[4,5] Library designs of increasing size are generated computationally for N equal to 15, 18, 21, 24, 27, and 30. In addition, random tiling combinations are generated for the same number and length of oligomers using the same design constraints outlined for models ***M1*** and ***M2*** [see Fig. 1(a) and (b)] and the average number of clashes per hybrid are calculated for different library sizes. As expected, we find that in both cases the libraries designed using OPTCOMB include much fewer clashes than the randomly generated libraries. Figure 3 depicts the number of clashes (◆) retained between optimally designed oligomers using model ***M1*** [Fig. 3(a)] and model **M2** [Fig. 3(b)] against library size. These clashes are contrasted against the average number of clashes (▲) between oligomers for randomly generated tiling combinations for the two cases. These results clearly demonstrate that substantial improvement in library design can be made by pro-actively minimizing clash retentions. Comparisons between optimal designs obtained with models ***M1*** and ***M2*** reveal that the additional flexibility of "skipping" of certain parental fragments at key junction positions reduces clash retention by approximately 50% (see Fig. 3) for the same library size.

The second question focuses on the tiling characteristics of optimal library designs. We find that, in general, the optimal designs obtained using model ***M2*** involve fragments of roughly similar lengths with, however, widely varying contributions from different parental sequences. In contrast, optimal designs using model ***M1*** typically employ nonuniform fragment lengths. For example, Figure 4 shows the optimal tiling pattern obtained using model ***M2*** for $N = 21$. Only a small portion of the *E. coli* sequence is present while most of *L. casei* and the entire *B. subtilis* sequence are participating in the optimal library design reflecting that OPTCOMB systematically disallows fragments from the *E. coli* sequence implicated in clash formation. The concatenation of the oligomers shown in Figure 4 yields a library composed of 1536 hybrids that avoid 44 out of the 50 clashes identified using FamClash. The remaining six clashes are shown as arcs connecting the two implicated residues (see Fig. 4). In contrast, libraries designed by random selection of junction points and sequence tiles involve on average 26 clashes. Notably,

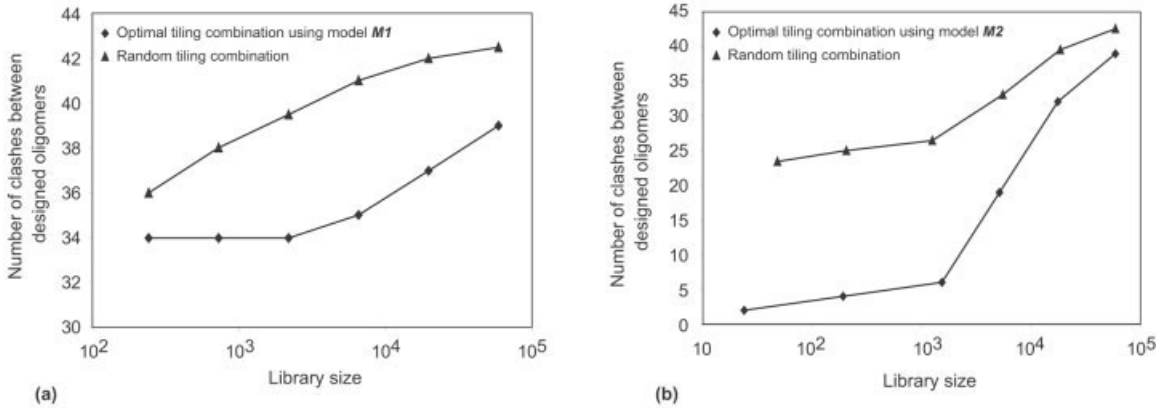

Fig. 3. Plot of the number of clashes between optimally designed oligomers (◆) using models (**a**) ***M1*** and (**b**) ***M2*** against library size. The average numbers of clashes between randomly generated designs (▲) for various library sizes are also shown.

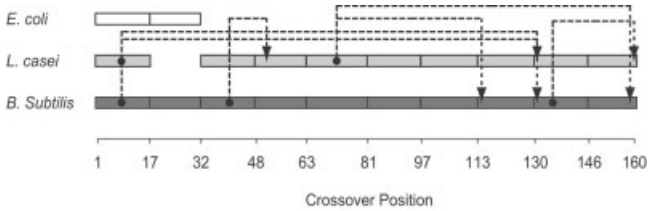

Fig. 4. Results obtained using model ***M2*** for minimum and maximum fragment lengths of 15 and 30 residues respectively and $N = 21$. The clashes that are retained are shown as dashed arcs with the position of the first residue of a clashing pair in the hybrid being represented by a dot (●).

the designed crossover positions do not follow any easily discernable patterns in terms of the underlying secondary structure. Although many of the designed crossovers fall within the loop regions, many of them are found to be within α-helices and β-sheets. The crossover positions also seem to be equally distributed between conserved and nonconserved stretches of parental sequences.

The third question examines the optimal tradeoff between library size and quality exemplified by the number of clashes between fragments chosen for the library design, the percent of clash-free hybrids and the average number of remaining clashes per hybrid. Clearly, the number of both the clash-free and clash-containing hybrids increases with increasing library size. However, because there is a limit to the number of sequences that can be screened, we use the percent of clash-free hybrids as a metric of quality. Tradeoff curves for these three different library quality metrics are generated using model ***M2*** to assess library quality (see Fig. 5). Figure 5(a) shows the tradeoff curve between library size and number of clashes between fragments that constitute the library for different values of N. The number of clashing residue pairs is, as expected, monotonically increasing with library size. Interestingly, we find that the rate of increase, beyond a library size of approximately 1.6×10^3 [shown as a dashed line in Fig. 5(a)], is dramatically enhanced. It appears that beyond this size threshold OPTCOMB runs out of nearly clash-free fragment combinations, and thus clash-forming fragments must be used to meet the increased library size requirements. The same behavior is observed for libraries designed using varying ranges of fragment length implying a global trend. This transition point also shows prominently in the tradeoff curves between (1) the percent of clash-free hybrids and the library size [see Fig. 5(b)], and (2) the average number of clashes per hybrid versus the library size [Fig. 5(c)]. We find that the percent of clash-free hybrids increase up to this transition point and afterwards it begins to decline [Fig. 5(b)]. Accordingly, the average number of clashes per hybrid decreases up to this point and begins to rise again [Fig. 5(c)]. The reason for this trend is that for small library sizes the OPTCOMB model chooses the junction points and the contributing parental sequences such that most of the clash-forming fragments are avoided. However, there is only a limited number of clash-free fragment combinations, all of which are selected before the threshold library size. Therefore, to obtain library sizes beyond this threshold size, the model is forced to choose fragments involving increasingly higher number of clashes resulting in the decline in the percent of clash-free hybrids (or alternatively resulting in the increase in the percent of hybrids with clashes) in the library. It is noteworthy that this transition point is at approximately the same library size (or value of N) for all library quality metrics [see Figure 5(a)–(c)]. The *a priori* identification of this optimal library size is of considerable importance to the application of directed evolution protocols by answering the question of what is the appropriate library size that best balances diversity with quality for a given protein engineering task.

As expected, the optimal library size is a strong function of the fragment/oligomer sizes, and is found to increase substantially with decreasing fragment length ranges. Figure 6 depicts the optimal library size for different ranges of fragment sizes. Smaller fragment sizes afford more fragment choices for library design and significantly more tiling combinations to choose from. Because different experimental protocols for directed evolution have different requirements on fragment lengths, the tradeoff curves

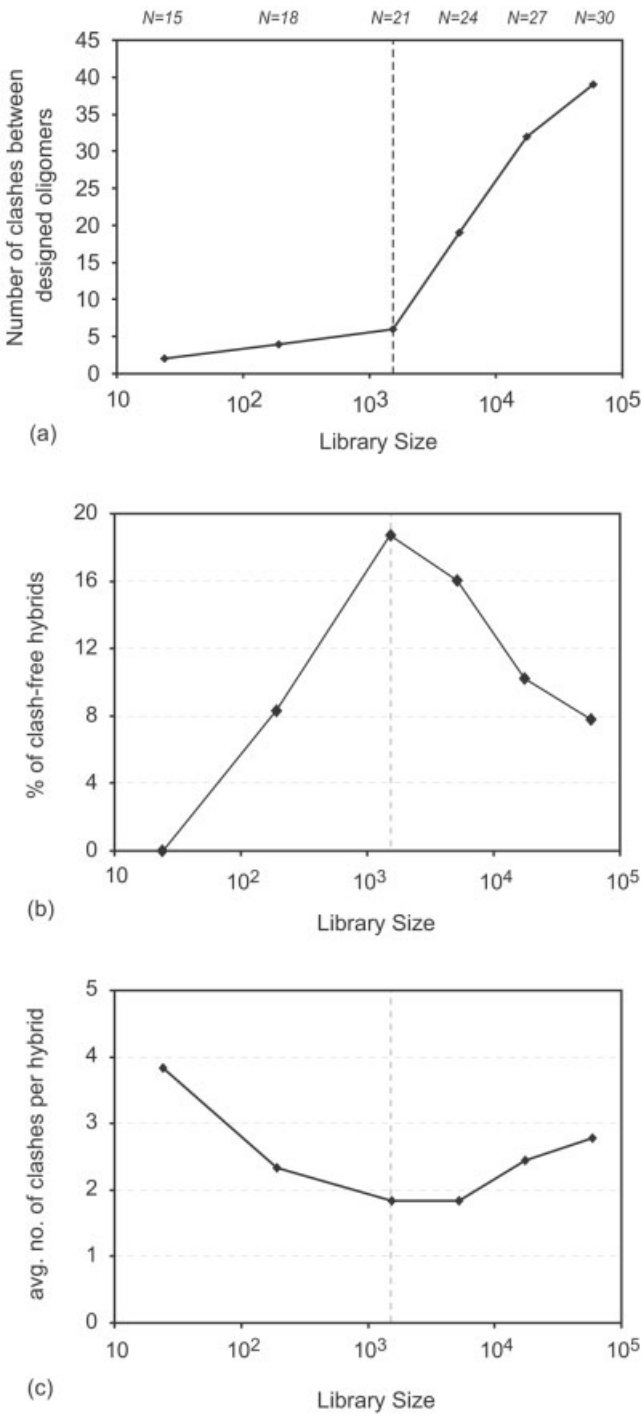

Fig. 5. (**a**) Plot of the number of clashes between selected parental fragments (corresponding to $N = 15, 18, 21, 24, 27$, and 30; $L_{min} = 15$ and $L_{max} = 30$) forming the library against library size. There is an optimal library size $\sim 1.6 \times 10^3$ (shown with a dashed line) beyond which the number of clashes increases significantly. (**b**) Plot of the percent of clash-free hybrids versus library size. Notably, at the transition point/optimal library size (1.6×10^3) the percent of clash-free hybrids is at a maximum. (**c**) Plot of the average number of clashes per hybrid versus library size. Again the minimum number of clashes is observed at the optimal library size (1.6×10^3).

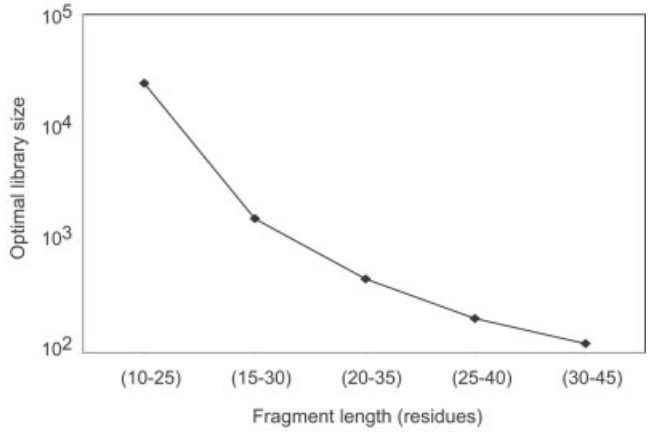

Fig. 6. Plot of the optimal library size for different ranges (10–25, 15–30, 20–35, 25–40, and 30–45) of fragment lengths. The optimal library size decreases with increasing fragment sizes.

such as the one shown in Figure 6 can aid in selecting the correct protocol based on library size or the sequence space to be explored.

The last question explores the effect of combinatorial library size (or N) on the tiling combination, the clash distribution, and the relative contribution of the three parental sequences towards the library. We find that the optimal tiling combination and the relative contribution of the parental sequences change significantly when N is varied (see Fig. 7) and that there exists persistently "skipped" fragments (e.g., residues 80–130 of the *E. coli* sequence) in the tiling combinations. For example, we observe that the contribution of the *B. subtilis* and *L. casei* sequences to the library increases with N. Interestingly, we find that although initially the *E. coli* sequence contribution to the library is equal to the one from *L. casei* ($\sim$40% each for $N = 15$), it rapidly drops to 10% (for $N = 18$), after which it increases to meet the increasingly higher required numbers of oligomers (see Fig. 7). At the end ($N = 30$), there are no skipped fragments, thus recovering the solution of model ***M1***. Although, the fragment sizes are allowed to vary from 15–30 residues, we find that the fragment size chosen in the library design are fairly uniform and range between 15–18 residues. Clearly, smaller fragments allow for more flexibility, and therefore enhance the chances of avoiding the clashes. The largely nonvarying fragment sizes imply that the location of the junction points as well do not change significantly (see Fig. 7). The distribution of number of hybrids based on the number of clashes follow a log-normal distribution with the number of clashes in the hybrids varying from 0–10. The distribution of clashes is narrow for small values of N and broadens with increasing N. Note that the total number of clashes present in the hybrids of a given library vary between 0–10 and is significantly lower than the total number of clashing residue pairs that can be formed between the fragments that constitute the library (as many as 39 for $N = 30$).

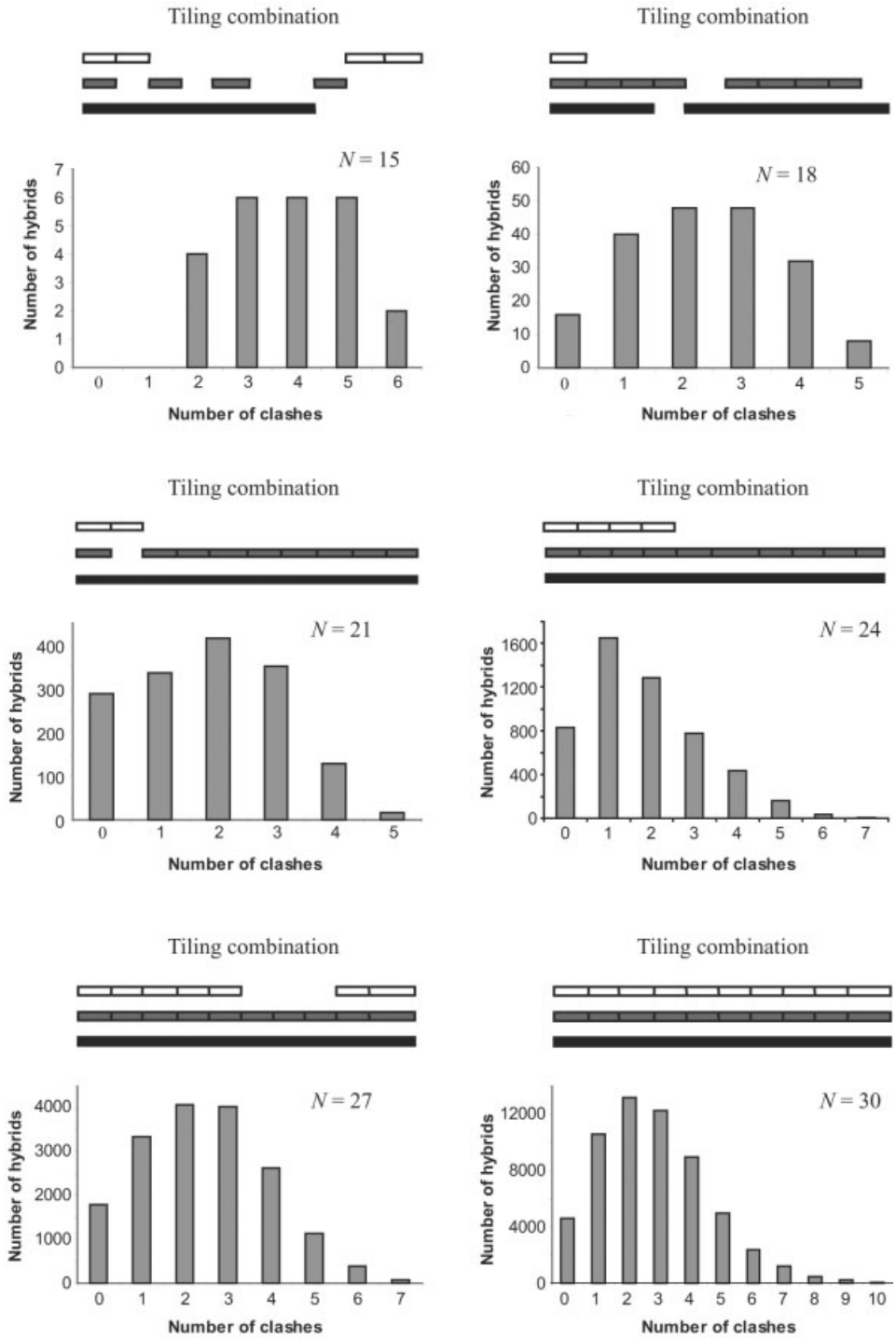

Fig. 7. The tiling choices and the clash distributions for the hybrids for $N = 15$, 18, 21, 24, 27, and 30.

SUMMARY

In this article, the OPTCOMB procedure was introduced for the optimal design of synthetic oligomer ligation based protocols.[4–6] The capabilities of OPTCOMB were demonstrated by computationally designing recombinant libraries composed of sequences from *E. coli, B. subtilis*, and *L. casei* DHFR proteins. The key result of this study is the computational verification of the existence of an optimal library size that best balances library diversity and quality. The optimal library size was found to be a strong function of fragment size and involved the coordinated skipping of certain parental fragments.

Clearly, the obtained results depend on the accuracy of the clash prediction frameworks.[13–16] We expect that

more accurate clash prediction methods will become available in the future, which can capture backbone movement in the hybrids through the use of sophisticated potential energy/scoring functions.[27–29] Nevertheless, OPTCOMB provides a versatile framework that can handle the information generated by various clash prediction methods.[13–16]

REFERENCES

1. Stemmer WP. Rapid evolution of a protein in vitro by DNA shuffling. Nature 1994;370:389–391.
2. Joern JM, Meinhold P, Arnold FH. Analysis of shuffled gene libraries. J Mol Biol 2002;316:643–656.
3. Moore GL, Maranas CD. Computational challenges in combinatorial library design for protein engineering. AICHE J 2004;50:262–272.
4. Ness JE, Kim S, Gottman A, Pak R, Krebber A, Borchert TV, Govindarajan S, Mundorff EC, Minshull J. Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. Nat Biotechnol 2002;20:1251–1255.
5. Coco WM, Encell LP, Levinson WE, Crist MJ, Loomis AK, Licato LL, Arensdorf JJ, Sica N, Pienkos PT, Monticello DJ. Growth factor engineering by degenerate homoduplex gene family recombination. Nat Biotechnol 2002;20:1246–1250.
6. Hiraga K, Arnold FH. General method for sequence-independent site-directed chimeragenesis. J Mol Biol 2003;330:287–296.
7. Patrick WM, Firth AE, Blackburn JM. User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. Protein Eng 2003;16:451–457.
8. Ostermeier M. Synthetic gene libraries: in search of the optimal diversity. Trends Biotechnol 2003:21:244–247.
9. Osborne MJ, Schnell J, Benkovic SJ, Dyson HJ, Wright PE. Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. Biochemistry 2001;40:9846–9859.
10. Rod TH, Radkiewicz JL, Brooks CL 3rd. Correlated motion and the effect of distal mutations in dihydrofolate reductase. Proc Natl Acad Sci USA 2003;100:6980–6985.
11. Sawaya MR, Kraut J. Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. Biochemistry 1997;36:586–603.
12. Agarwal PK, Billeter SR, Rajagopalan PT, Benkovic SJ, Hammes-Schiffer S. Network of coupled promoting motions in enyzme catalysis. Proc Natl Acad Sci USA 2002;99:2794–2799.
13. Saraf MC, Horswill AR, Benkovic SJ, Maranas CD. FamClash: a method for ranking the activity of engineered enyzmes. Proc Natl Acad Sci USA 2004;101:4142–4147.
14. Saraf MC, Maranas CD. Using a residue clash map to functionally characterize protein recombination hybrids. Protein Eng 2003;16:1025–1034.
15. Moore GL, Maranas CD. Identifying residue–residue clashes in protein hydrids by using a second-order mean-field approach. Proc Natl Acad Sci USA 2003;100:5091–5096.
16. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. Protein building blocks preserved by recombination. Nat Struct Biol 2002;9:553–558.
17. Meyer MM, Silberg JJ, Voigt CA, Endelman JB, Mayo SL, Wang ZG, Arnold FH. Library analysis of SCHEME-guided protein recombination. Protein Sci 2003;12:1686–1693.
18. Endelman JB, Silberg JJ, Wang ZG, Arnold FH. Site-directed protein recombination as a shortest-path problem. Protein Eng Des Sel 2004;17:589–594.
19. Saraf MC, Moore GL, Maranas CD. Using multiple sequence correlation analysis to characterize functionally important protein regions. Protein Eng 2003;16:397–406.
20. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
21. Nagi AD, Regan L. An inverse correlation between loop length and stability in a four-helix-bundle protein. Fold Des 1997;2:67–75.
22. Bogarad LD, Deem MW. A hierarchical approach to protein molecular evolution. Proc Natl Acad Sci USA 1999;96:2591–2595.
23. Glover F. Improved linear integer programming formulations of nonlinear integer problems. Manage Sci 1975;22:455–460.
24. Bateman A, Coin L, Durbin R, Finn RD, Holich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32(Database issue)D138–D141.
25. Brooke A, Kendrick D, Meeraus A, Raman R. GAMS: the solver manuals. Washington, DC: GAMS Developemnt Corporation; 2005.
26. Brooke A, Kendrick D, Meeraus A, Raman R. GAMS: a user's guide. Washington, DC: GAMS Development Corporation; 2003.
27. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature 2003;423:185–190.
28. Dwyer MA, Looper LL, Hellinga HW. Computational design of a biologically active enzyme. Science 2004;304:1967–1971.
29. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.