

# Increased Detection of Structural Templates Using Alignments of Designed Sequences

Stefan M. Larson,<sup>1</sup> Amit Garg,<sup>2</sup> John R. Desjarlais,<sup>3</sup> and Vijay S. Pande<sup>1\*</sup>

<sup>1</sup>Department of Chemistry and Biophysics Program, Stanford University, Stanford, California

<sup>2</sup>Computer Science Department, Stanford University, Stanford, California

<sup>3</sup>Xencor, Inc., Monrovia, California

**ABSTRACT** Protein structure prediction by comparative modeling benefits greatly from the use of multiple sequence alignment information to improve the accuracy of structural template identification and the alignment of target sequences to structural templates. Unfortunately, this benefit is limited to those protein sequences for which at least several natural sequence homologues exist. We show here that the use of large diverse alignments of computationally designed protein sequences confers many of the same benefits as natural sequences in identifying structural templates for comparative modeling targets. A large-scale massively parallelized application of an all-atom protein design algorithm, including a simple model of peptide backbone flexibility, has allowed us to generate 500 diverse, non-native, high-quality sequences for each of 264 protein structures in our test set. PSI-BLAST searches using the sequence profiles generated from the designed sequences (“reverse” BLAST searches) give near-perfect accuracy in identifying true structural homologues of the parent structure, with 54% coverage. In 41 of 49 genomes scanned using reverse BLAST searches, at least one novel structural template (not found by the standard method of PSI-BLAST against PDB) is identified. Further improvements in coverage, through optimizing the scoring function used to design sequences and continued application to new protein structures beyond the test set, will allow this method to mature into a useful strategy for identifying distantly related structural templates. *Proteins* 2003;51:390–396.

© 2003 Wiley-Liss, Inc.

**Key words:** protein design; BLAST; structural genomics; comparative modeling; sequence space; structure prediction

## INTRODUCTION

A commonly stated goal of structural genomics is to solve enough protein structures to allow comparative modeling of all remaining and future genomic sequences.<sup>1</sup> This rationale relies on the assumption that the total number of unique folds in the protein universe is finite and relatively small (perhaps around 10,000, give or take an order of magnitude)<sup>2</sup> compared to the total number of unique genome sequences (already well into the millions).<sup>3</sup>

If and when a high-resolution structure of at least one representative from each protein fold is available, the onus then shifts to the comparative modeling community to generate a model structure for all remaining protein sequences. The first hurdle in such a large-scale comparative modeling effort is to find a template protein structure for every protein sequence, using sequence-matching algorithms to infer structural similarity. This inference is not perfect, and false positives do occur. As stated in an insightful review of comparative modeling in CASP4,<sup>4</sup> “no model is better than a wrong model.” However, to achieve the goal of full structural coverage of all protein sequences, assuming that the structure of at least one representative of every protein fold is known, a template structure for every sequence must be identified. Once identified, the protein sequence must then be correctly aligned to the template structure, with subsequent refinement of the resulting model. It has been repeatedly shown that these two steps, identifying a template structure and aligning the sequence to that structure, are the largest sources of error in comparative modeling.<sup>4</sup>

Even faced with these challenges, comparative protein structure modeling is currently the most accurate method for structure prediction, and is feasible for significant segments of approximately one-third of all known protein sequences (i.e., one-third of all sequences are recognizably related to at least one protein structure).<sup>5</sup> It is likely that for the remaining two-thirds, a significant fraction *does* have a solved structural homologue, but their structural similarity is not recognized through sequence similarity searching techniques.<sup>5</sup> This is quite significant: improving the fraction of modelable sequences from one-third to, say, one-half would represent a major increase in our understanding of the protein universe. An important step in this direction has been the widely corroborated finding that input from multiple sequence alignments (e.g., searching for template structures with profiles built from sequence alignments as opposed to single target sequences)<sup>6</sup> greatly enhances the efficacy and accuracy of almost all phases of

Grant sponsor: Bing Fellowship Committee.

\*Correspondence to: Vijay S. Pande, Chemistry Department, Stanford University, Stanford, CA 94305-5080. E-mail: pande@stanford.edu

Received 7 August 2002; Accepted 13 November 2002

comparative modeling. However, meaningful sequence alignments can only be built when several natural sequence homologues are available, ideally for both the target sequence and the potential template structures. For protein folds with only one or few representatives, this valuable information is not available.

Multiple sequence alignments are useful in comparative modeling because they represent a natural sampling of a protein fold's sequence space. In previous work by our group and others,<sup>7-9</sup> it has been shown that protein design algorithms can be used to sample protein sequence space in a meaningful way, even in the absence of natural sequence homologues. In this study, we demonstrate the utility of large, diverse, high-quality sequence libraries, generated by large-scale computational protein design, for increasing the coverage of comparative modeling methods for protein structure prediction.

## METHODS

### Genome@home Distributed Computing Cluster

To allow for studies of this scope, a distributed computing project, dubbed Genome@home, was created (see <http://genomeathome.stanford.edu>).<sup>7</sup> During the course of this study, the global cluster of available computers exceeded 3,000 processors. Briefly, the Genome@home server sends out “work units,” a set of protein backbone coordinates and design parameters that are downloaded to the Genome@home client running on a user's computer. The client verifies the work unit, and runs the protein design algorithm,<sup>10</sup> summarized below. Work units of the size used in this study require a few hours to a day on a 400-MHz Intel Celeron processor. Upon completion of the sequence design, the results are verified by the client and sent back to the server, where the data is again verified, stored, and processed.

### Protein Sequence Design

Sequences were designed using SPA,<sup>10</sup> with modifications as previously described.<sup>7</sup> Briefly, protein structures are created by modeling the placement of amino acid side-chain rotamers onto a fixed target backbone. Models are scored using a combination of the Amber potential function<sup>11</sup> with OPLS nonbonded parameters,<sup>12</sup> a surface-area term that accounts implicitly for solvation effects,<sup>13</sup> and a set of amino acid baseline corrections, which are critical for maintaining reasonable amino acid compositions. The models are optimized by a sequence selection process that involves initial filtering of rotamers, and a genetic algorithm for finding an optimal sequence for the target structure. To create an ensemble of 100 target backbones for each structure, a Monte Carlo expansion and contraction algorithm was used to gently perturb the dihedral angles of the target backbone. The algorithm works by creating random perturbations of up to five degrees to the dihedral angles of the target structure, followed by simple Monte Carlo with smaller random perturbations until the target RMSD from the native structure is reached. In this study, the perturbation was

constrained such that no two backbones in the ensemble differ by more than 1.0 Å RMSD. Each work unit of sequence design is done against a fixed backbone (i.e., one of the 100 variants of the target structure), and the designed sequences for all 100 variants are included in the resulting overall sequence set for the target structure.

## RESULTS

As part of an overall effort to generate large diverse sequence libraries, a distributed computing architecture for large-scale protein sequence design was established (see Methods) and tested on an initial set of 264 structures. We sought to define a test set that was (1) unbiased, (2) could give reasonable and representative coverage of the space of protein structures, and (3) for which we could collect sufficient data in a reasonable amount of time. Since CPU-time for protein design algorithms scales supra-linearly with sequence length, we limited the size of proteins in our test set to no more than 100 amino acids in length. Within this limit, to avoid bias, the test set consisted of *all* protein structures in the Protein Data Bank (PDB)<sup>14</sup> that contained one chain, between 30 and 100 amino acids long, solved by X-ray crystallography (roughly 300 records at the time). Sufficient data was returned to complete analyses for 264 of these structures. Table I summarizes the makeup of this diverse test set, which includes 55 different SCOP<sup>15</sup> folds, all 4 protein classes, and represents roughly 10% of known protein folds. Five hundred unique sequences were designed for each protein structure.

### “Reverse” BLAST Searches With Designed Sequence Alignments

The standard method for finding a structural template for a hypothetical genomic protein sequence is to use a sequence similarity searching algorithm, like PSI-BLAST,<sup>16</sup> to search a database of protein sequences with known structures, like the PDB. When available, natural sequence homologues of the target sequence are used to build sequence alignments and profiles; the use of this information in sequence similarity searches greatly increases the probability of finding a significantly close structural template. Conversely, the method presented here, large-scale application of protein design algorithms, allows the generation of sequence “homologues” for the potential *structural templates*. To make use of this new sequence information in searching for structural templates, we applied a process that we refer to as “reverse BLAST searching.” Instead of BLASTing genome sequences against a protein structure database, we used the 264 designed sequence alignments as queries, which were BLASTed against the genome in question. A significant hit indicates that the parent structure from which the query sequence alignment was built could serve as a structural template for the “hit” gene sequence.

To test the accuracy of this method, we used the entire PDB as an “artificial genome,” for which the correct structure of every protein sequence is already known. This

**TABLE 1. Protein Folds Included in Test Set of 264 Protein Structures (Classification According to SCOP)**

Class	Fold
All alpha proteins: 43	Acyl carrier protein-like; Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin; Cytochrome c; DEATH domain; DNA/RNA-binding 3-helical bundle; EF Hand-like; Fungal elicitor, plant pathogen; Histone-fold; lambda repressor-like DNA-binding domains; p53 tetramerization domain; Protozoan pheromone proteins; ROP-like; S15/NS1 RNA-binding domain; SAM domain-like; Uteroglobin-like
All beta proteins: 86	Acid proteases; alpha-Amylase inhibitor tendamistat; Barrel-sandwich hybrid; beta-clip; C2 domain-like; Cupredoxins; gamma-Crystallin-like; Immunoglobulin-like beta-sandwich; OB-fold; PDZ domain-like; Prealbumin-like; SH3-like barrel
Alpha and beta proteins (a/b): 1	BRCT domain
Alpha and beta proteins (a+b): 50	beta-Grasp (ubiquitin-like); CI-2 family of serine protease inhibitors; Cytochrome b5; DNA topoisomerase I domain; DNA-binding domain of Mlu1-box binding protein MBP1; Ferredoxin-like; Histidine-containing phosphocarrier proteins (HPr); IF3-like; IL8-like; KH-domain; Microbial ribonucleases; POZ domain; Ribosomal protein L7/12, C-terminal domain
Small proteins: 73	BPTI-like; Crambin-like; Cysteine-rich domain; HIPIP (high potential iron protein); Knottins (small inhibitors, toxins, lectins); Kringle-like; Ligand-binding domain of low-density lipoprotein receptor; Ovomuroid/PCI-1 like inhibitors; Rubredoxin-like; Snake toxin-like
Coiled coil protein: 9	Parallel coiled-coil; Stalk segment of viral fusion proteins
Peptides: 2	MoMLV p15 fragment (residues 409-426); Peptide hormones

allowed us to explicitly test the coverage and accuracy of reverse BLAST searching for finding structural templates for the “genes” in the PDB. Two hundred fifty-one of the 264 designed sequence alignments produced hits with E-values below 10. Figure 1 displays the distribution of E-values for the most significant hit for each of the 264 designed sequence alignments. One hundred forty-three sequence alignments produced hits at an E-value threshold of  $E < 0.1$ , with only 2 false positives (i.e., a hit against

a non-homologous structure). At a significance level of  $E < 0.01$ , a commonly used threshold in comparative modeling,<sup>4</sup> all hits were against true structural homologues, with 47% (124/264) coverage.

The accuracy of the reverse BLAST searching method is very high, but its coverage is mediocre, with half of the designed sequence alignments not producing a significant hit when BLASTed against the PDB. The most likely reason for this is that the protein design algorithm is, in

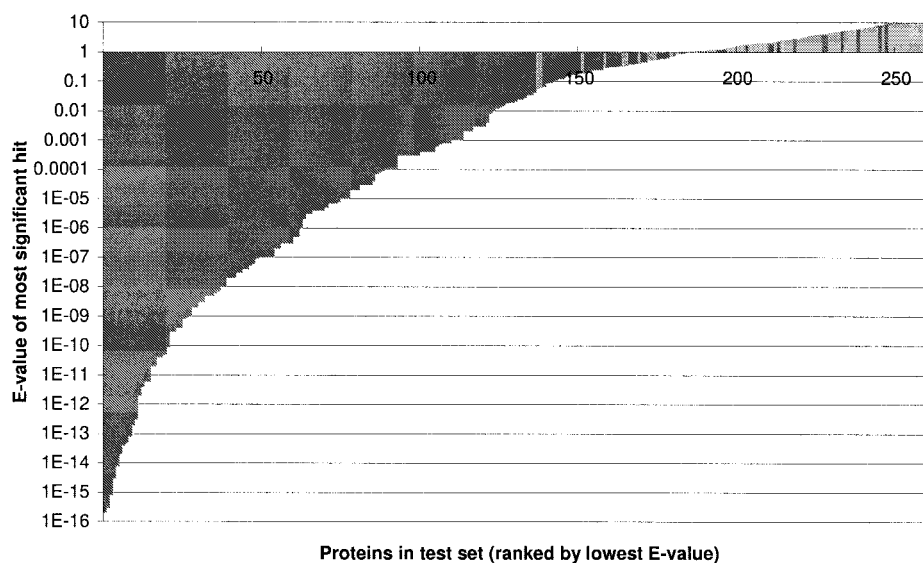


Fig. 1. Accuracy of structural template identification by “reverse BLAST searching” using designed sequence alignments. The most significant (i.e., lowest E-value) hit from each of 264 PSI-BLAST searches (using designed sequence alignments as input) against PDB was used as a predicted structural template. The E-values of the most significant hit for each protein is plotted, ranked by E-value. Dark gray columns represent predictions that were true structural homologues; light gray columns represent false positives. Below a significance level of  $E < 0.01$ , accuracy is 100%, but only 124 of 264 give hits below this threshold.

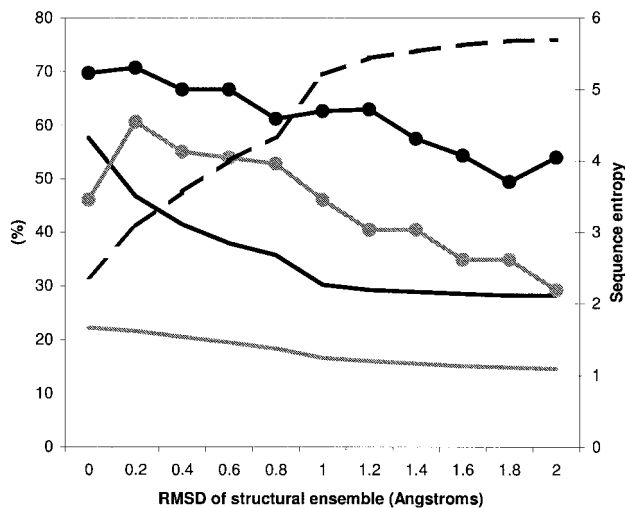


Fig. 2. Effects of increasing structural diversity of backbone ensembles. The primary y-axis plots average pairwise identity (black line) and identity to the native sequence (gray line) of the designed sequence alignments. The accuracy (black circles) and coverage (gray circles) of the reverse BLAST searches against PDB using the designed sequence alignments is also plotted on the primary y-axis. The secondary y-axis shows the sequence entropy (dashed black plot) of the resulting designed sequence alignments.

some cases, failing to produce sequences that maintain key native-like sequence signatures. The design procedure used here was developed to broadly explore sequence space, and tends to conserve only those sequence characteristics that are critical for specifying and maintaining key aspects of the protein structure.<sup>7</sup> The average identity of the designed sequences to the parent structure is only 23%, near the “twilight zone” of sequence homology searching, which can make it difficult to recognize structural homologues by sequence similarity alone.

To understand the effects of sequence diversity on the success of reverse BLAST searching, an additional smaller scale experiment was performed (see Fig. 2). The CPU-time required to generate the large data set used in the original analysis was over  $10^5$  CPU-days, so a smaller subset of one-third (89) of the original 264 structures was randomly selected, and ten structural ensembles were generated for each structure, ranging from 0.2 to 2.0 Å RMSD in 0.2-Å increments. Three hundred sequences were designed for each of these 880 structural ensembles. Each designed sequence alignment was then used to perform a reverse BLAST search against the PDB. The overall coverage of the 1.0-Å subset, 46% (41/89), is almost the same as that of the main test set used in this study (124/264), also generated using 1.0-Å RMSD structural ensembles. The highest coverage (54/89) and accuracy (71% of hits at  $E < 10$  to true structural homologues) was produced by the tightest structural ensemble, 0.2 Å RMSD. Interestingly, though coverage and accuracy tend to decrease with increasing RMSD, the peak for both is not at 0 Å RMSD. Substantially greater coverage is achieved by using a structural ensemble as opposed to a single fixed backbone.

It is interesting to ask if there are certain structural characteristics that this method does not currently handle well. Of the six SCOP classes represented in our test set, the small and coiled-coil proteins, together with the peptides, produced only 21% coverage. The proteins from the alpha, beta, and alpha and beta classes produced 60% coverage. Within these three classes, seven folds never produced a significant hit: Histone-fold, lambda repressor-like DNA-binding domains, ROP-like, S15/NS1 RNA-binding domain, Uteroglobin-like, alpha-Amylase inhibitor tendamistat, IF3-like, IL8-like. Although no common structural characteristics could be identified within this subset of proteins, the lack of hits is likely due to some flaw in the design criteria that does not reproduce native-like sequence characteristics in these cases. It is notable that four of these seven folds are nucleic acid-binding proteins. Since the proteins are designed in isolation, there is no “selective pressure” to conserve ligand-binding residues. We have found (data not shown) that designing proteins based on co-crystal structures with their natural binding partners produces more native-like amino acids at key ligand-binding residues. Since conservation of ligand-binding residues contributes to a native-like sequence profile (although more so in some folds than others), it may be a generally important factor for remote homology detection using designed sequences. Computational design of protein interfaces is, however, a difficult problem (see Shifman and Mayo<sup>17</sup> and references therein), and we are exploring methods to include functional information in our design calculations.

#### Utility of Designed Sequence Alignments in Identifying Structural Templates

The success of comparative modeling is highly dependent on the identification of high-quality structural templates. In fact, in CASP4, it was seen that the final refined model rarely scored better than the unrefined structural template.<sup>4</sup> Unfortunately, at stringent sequence similarity thresholds (e.g.,  $E < 10^{-5}$ ), where the likelihood of identifying an excellent structural template is high, the number of templates identified is often low. The increased accuracy of the resulting structural models is offset by the small number of genes for which such high-quality models can be built. It is known that incorporating information from natural sequence alignments generally increases the number of viable structural templates found for sequences of unknown structure. As described below, this also seems to hold true for designed sequence alignments, especially in cases when stringent thresholds are required.

Again using the PDB as an artificial test “genome,” we compared the number of structural templates identified with and without the use of designed sequence alignments. Each of the 264 proteins in the test set was BLASTed against the PDB, as described above, with and without the inclusion of the designed sequence alignments. Figure 3 shows the fold-increase in the number of genes for which structural templates were identified, and the total number of unique structural templates identified, when designed

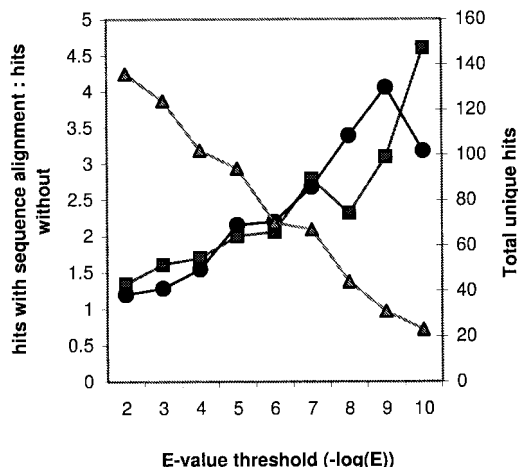


Fig. 3. Increased number of high-quality structural templates identified using reverse BLAST searching. The ratio of correctly identified structural templates found with vs. without the incorporation of designed sequence alignment information in PSI-BLAST searches against the PDB, plotted vs. increasingly stringent search threshold [decreasing  $\log(E)$ ]. Squares: fold-increase in unique structural templates identified; circles: fold-increase in number of genes for which structural templates were identified. The secondary y-axis (triangles) plots the decrease in total number of hits identified with decreasing  $\log(E)$ .

sequence alignments were used in reverse BLAST searching. As the required stringency of sequence matching increases (i.e., E-value decreases), the fold-increase in hits generated by the use of designed sequence alignments

shows a roughly linear increase. The total number of hits decreases with E-value, as expected. Nonetheless, at high sequence similarity thresholds (e.g.,  $E < 10^{-5}$  or stricter), the use of designed sequence alignments allowed for the identification of 2 to 5 times more structural templates than single sequence BLAST searches.

The accuracy and utility of the reverse BLAST searching method is clear from these tests, where the PDB was used as an artificial genome. However, the PDB represents a fairly biased set of protein sequences and does not capture the composition of typical genomes. To evaluate the utility of the method in identifying structural templates for sequences in real genomes, it was applied, again using our test set of 264 designed sequence alignments, against 49 complete microbial genomes (Fig. 4). For comparison against a standard method, all the genomic protein sequences were also queried, using PSI-BLAST (5 iterations), against the sequences of the 264 parent structures (i.e., the corresponding test subset of the PDB). Reverse BLAST searching produced no hits for eight genomes, but identified at least one additional structural template (not identified by the standard PSI-BLAST method against the 264 sequence PDB subset) for the other 41 genomes. In ten cases, the reverse BLAST method more than doubled the number of structural templates identified by PSI-BLAST searching alone (which failed to identify any hits whatsoever in five genomes). Finally, the genome-to-genome variation in the number of additional hits attained by the reverse BLAST method is similar to the variation in the

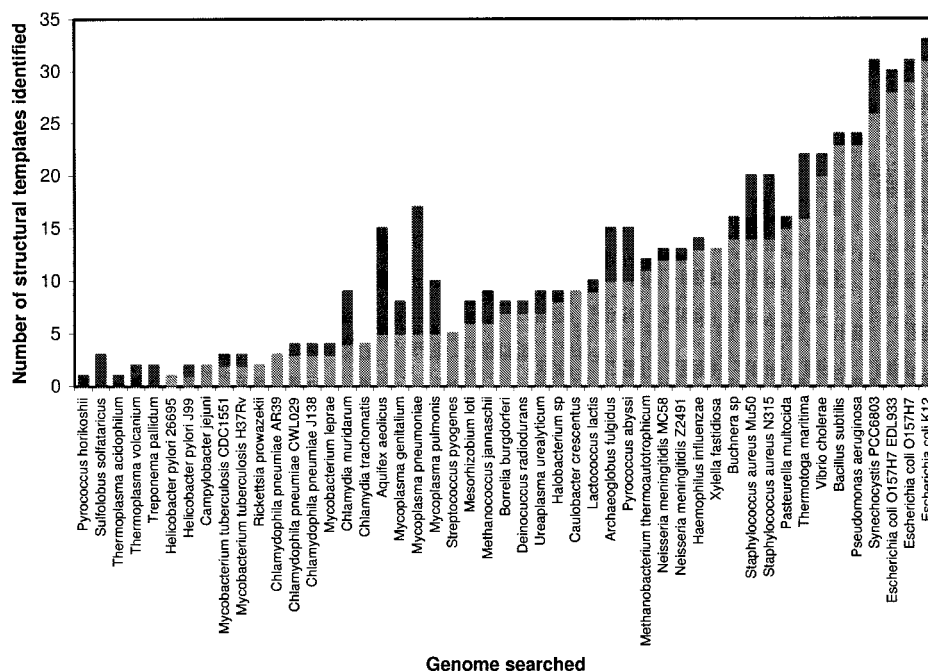


Fig. 4. Number of genes, in 49 genomes, for which structural templates were identified using reverse BLAST searching. Light gray: the number of genes for which structural templates were identified by PSI-BLAST searching against the set of 264 structures in the test set. Dark gray: the number of novel genes for which structural templates were identified by reverse BLAST searching using the 264 designed sequence alignments.

number of hits attained by the standard PSI-BLAST approach, suggesting that there is no systematic bias, arising from genome composition, affecting the performance of the reverse BLAST method in particular.

## DISCUSSION

It is well known that protein structure prediction through comparative modeling can be greatly improved by incorporating sequence homologues of the target sequence and/or additional knowledge about the sequence space of the predicted structural template. We have shown here that it is possible to reproduce many of the benefits of natural sequence homologues in comparative modeling using sequences generated by computational protein design algorithms. We take advantage of the ability of the PSI-BLAST sequence matching algorithm to automatically create position-specific substitution matrices from input alignments, and do our sequence searches in "reverse": the sequence alignments generated by large-scale computational protein design are used to scan a genome for hits. Hits against gene sequences with unknown structure are then targets for comparative modeling with the parent structure of the query sequence alignment. This method can be used to recreate the benefits of natural sequence homologues in those cases where none (or very few) exist.

The maximum sequence length of 100 amino acids in the test set was a necessary limit to this study. The median length of structure prediction targets in CASP5 was 216 amino acids (<http://predictioncenter.llnl.gov/casp5>), well within the range of our method, although not within the bounds of the test set used here; about 10% of the CASP5 targets were below 100 amino acids. However, the results from our test set show no correlation with sequence length or protein size (data not shown). As well, our group and others have found no evidence that the size or characteristics of sequence space are affected by protein length.<sup>7,8,18</sup> Finally, there is little evidence to suggest that the success of remote homology detection or fold recognition techniques is dependent on protein sequence length. Thus, we expect that our results will not be significantly different for larger proteins.

Could this method be applied in large-scale comparative modeling efforts? Since heuristic sequence-matching algorithms such as BLAST are both efficient and scalable, it would be straightforward and tractable to use even very large sets of designed sequence alignments (e.g., up to  $10^4$  or  $10^5$ , one for each protein fold) to search through new genomes for potential comparative modeling targets. Two improvements are necessary to allow full-scale application of this method. First, the coverage provided by reverse BLAST searching is fairly low: only 47% of designed sequence alignments generated a significant hit ( $E < 0.01$ ) when BLASTed against the PDB. This is likely due to shortcomings in the protein design procedure used here. Traditionally, protein design algorithms have been developed to find one optimal (usually meaning most stable) sequence for a specific protein structure. The protein design algorithm used here was designed to broadly ex-

plore sequence space and generate a large diversity of sequences. Other large-scale studies of computational protein design have reported producing more native-like sequences.<sup>8,9</sup> Koehl and Levitt<sup>8</sup> achieve more native-like sequences with the oft-used constraint of fixed (to native) amino-acid composition. Kuhlman and Baker<sup>9</sup> use a single backbone structure for each of their designs, and achieve slightly higher native-like sequence prediction. The design method in this study, on the other hand, has tried to push sequence diversity by using ensembles of backbones with 1-Å RMSD from the parent template structure, which lowers the native character of the sequences (see Fig. 3). If only to steer clear of the twilight zone of sequence matching, the reverse BLAST searching method might well benefit from a protein design algorithm that produces more native-like sequences, while still exploring sequence space to the same extent as a large, diverse natural sequence alignment (e.g., by using a tighter, 0.2-Å RMSD, structural ensemble).

The second obstacle to wide application of this method is the fact that the current set of designed sequence alignments covers only a small subset (264) of all known protein structures (~15,000). The method will not be fully useful until the vast majority of protein structures is covered, approaching one designed sequence alignment for every known protein fold. A practical approach will be to initially concentrate our efforts on designing sequences for protein structures with few or no natural sequence homologues. In the future, a distributed computing architecture such as the Genome@home project will allow the generation of hundreds of diverse designed sequences for every protein in the PDB, and progress is underway towards that goal.

## ACKNOWLEDGMENTS

This work would not have been possible without the enthusiastic participation of thousands of Genome@home users around the world. The authors are greatly indebted to everyone who contributed processor time to this study. A full list of users is available at <http://gah.stanford.edu/userstats.txt>. The authors thank the members of the Pande Group for helpful discussions. S.M.L. is a James Clark Fellow of the SGF program. A.G. thanks the Bing Fellowship Committee for funding his research.

## REFERENCES

1. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
2. Sali A. 100,000 protein structures for the biologist. *Nat Struct Biol* 1998;5:1029–1032.
3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. *Nucleic Acids Res* 2002;30:17–20.
4. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. *Proteins* 2001; 45(Suppl 5):22–38.
5. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 2000;29:291–325.
6. Venclovas C. Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. *Proteins* 2001;45(Suppl 5):47–54.
7. Larson SM, England JL, Desjarlais JR, Pande VS. Thoroughly searching sequence space: large-scale protein design of structural ensembles. *Protein Sci* 2002;11 (in press).

8. Koehl P, Levitt M. De novo protein design. II. Plasticity in sequence space. *J Mol Biol* 1999;293:1183–1193.
9. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA* 2000;97:10383–10388.
10. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. *Protein Sci* 2000;9:1106–1119.
11. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* 1984;106:765–784.
12. Jorgensen WL, Tirado-Rives J. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* 1988;110:1657–1666.
13. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. *Nature* 1986;319:199–203.
14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000;28:235–242.
15. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* 2002;30:264–267.
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
17. Shifman JM, Mayo SL. Modulating calmodulin binding specificity through computational protein design. *J Mol Biol* 2002;323:417–423.
18. Koehl P, Levitt M. Protein topology and stability define the space of allowed sequences. *Proc Natl Acad Sci USA* 2002;99:1280–1285.