PROTEINS: Structure, Function, and Bioinformatics 55:656–677 (2004)

Modeling Structurally Variable Regions in Homologous Proteins With Rosetta

Carol A. Rohl,[1]* Charlie E.M. Strauss,[2] Dylan Chivian,[3] and David Baker[3]
[1]*Department of Biomolecular Engineering, University of California, Santa Cruz, California*
[2]*Biosciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico*
[3]*Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, Washington*

ABSTRACT **A major limitation of current comparative modeling methods is the accuracy with which regions that are structurally divergent from homologues of known structure can be modeled. Because structural differences between homologous proteins are responsible for variations in protein function and specificity, the ability to model these differences has important functional consequences. Although existing methods can provide reasonably accurate models of short loop regions, modeling longer structurally divergent regions is an unsolved problem. Here we describe a method based on the de novo structure prediction algorithm, Rosetta, for predicting conformations of structurally divergent regions in comparative models. Initial conformations for short segments are selected from the protein structure database, whereas longer segments are built up by using three- and nine-residue fragments drawn from the database and combined by using the Rosetta algorithm. A gap closure term in the potential in combination with modified Newton's method for gradient descent minimization is used to ensure continuity of the peptide backbone. Conformations of variable regions are refined in the context of a fixed template structure using Monte Carlo minimization together with rapid repacking of side-chains to iteratively optimize backbone torsion angles and side-chain rotamers. For short loops, mean accuracies of 0.69, 1.45, and 3.62 Å are obtained for 4, 8, and 12 residue loops, respectively. In addition, the method can provide reasonable models of conformations of longer protein segments: predicted conformations of 3Å root-mean-square deviation or better were obtained for 5 of 10 examples of segments ranging from 13 to 34 residues. In combination with a sequence alignment algorithm, this method generates complete, ungapped models of protein structures, including regions both similar to and divergent from a homologous structure. This combined method was used to make predictions for 28 protein domains in the Critical Assessment of Protein Structure 4 (CASP 4) and 59 domains in CASP 5, where the method ranked highly among comparative modeling and fold recognition methods. Model accuracy in these blind predictions is dominated by alignment quality, but in the context of accurate alignments, long protein segments can be accurately modeled. Notably, the method correctly predicted the local structure of a 39-residue insertion into a TIM barrel in CASP 5 target T0186. Proteins 2004;55:656–677.**

Key words: comparative protein structure modeling; homology modeling; fragment assembly; CASP; loop modeling; structurally variable region

INTRODUCTION

Comparative modeling is based on the observation that proteins with similar sequences almost always share similar structures (for review, see Ref. 1). Structure prediction by comparative modeling is initiated by aligning the query sequence to a parent sequence of known structure. For residues that can be aligned, the backbone coordinates of the model are based closely on the coordinates of the parent structure. Residues in the query sequence that cannot be aligned to the parent sequence because of insertions and deletions cannot, by definition, be modeled by using the parent structure as a template. Models for such segments of the protein must be constructed by alternate prediction methods. In addition, regions where sequence similarity is weak and/or alignment uncertain are also candidates for methods targeted at predicting conformations for protein segments corresponding to alignment gaps. Currently, ~30% of known sequences have sufficient sequence similarity to a known structure for current comparative modeling methods. One third of these sequences are similar over <80% of their length; consequently, complete three-dimensional (3D) models cannot be generated by homology-based methods alone.[2] Because sequence and structure divergence between homologous family members is responsible for changes in protein function and specificity, accurately modeling the struc-

Grant sponsor: Howard Hughes Medical Institute; Grant sponsor: Interdisciplinary Training in Genomic Sciences; Grant number: T32 HG00035-06.

*Correspondence to: Carol A. Rohl, Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064. E-mail: rohl@ucsc.edu

Received 8 October 2003; Accepted 14 July 2003

Published online 1 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.10629

tural differences between similar structures is an important goal of protein structure prediction.

Traditionally, loop modeling is defined as the problem of constructing 3D atomic models for short protein segments corresponding to loops on the proteins surface that connect regular secondary structure elements. Much attention has been focused on this problem, and several methods have been described that predict loop conformations up to about 8–12 residues with accuracies comparable to the accuracy of models obtained by homology-based methods. Modeling of longer segments of protein structures has received significantly less attention and remains generally an unsolved problem. Here, we use the term structurally variable region (SVR) modeling to refer to prediction of the conformation of any protein segment in the context of a framework or template structure, regardless of the segment length, secondary structure content, or surface exposure. We describe a method based on the successful de novo structure prediction method Rosetta[3,4] for modeling SVRs. In combination with an alignment algorithm to generate template structures, the SVR modeling method allows complete atomic models of proteins to be generated by combining both homology-based and de novo strategies. Complete models for comparative modeling and fold recognition targets were predicted by using this combined strategy and submitted to the Critical Assessment of Structure Prediction (CASP 4 and CASP 5), where the method was ranked highly.[5,6]

A thorough review of the extensive literature on loop modeling methodologies is beyond the scope of this article. Instead, we focus here on general approaches to and distinctions among loop modeling strategies to place the Rosetta-based strategy in context with respect to other methods, as well as to highlight novel contributions. Loop modeling methods primarily differ in the method of conformation generation and in the evaluation or scoring of alternate conformations. Algorithms can be generally grouped into knowledge-based methods, de novo or ab initio strategies, and combined approaches. The knowledge-based approach uses the database of experimental protein structures as a source of loop conformations.[7–14] Generally, such loop conformations are evaluated by using a knowledge-based potential or rule-based filters, evaluating such criteria as geometric fit and sequence similarity to select likely loop conformations. In the de novo approach, loop conformations are generated by a variety of methods including molecular dynamics,[15,16] simulated annealing,[17,18] exhaustive enumeration or heuristic sampling of a discrete set of (ϕ, ψ) angles,[19–23] random tweak,[24,25] or analytical methods.[26,27] Such de novo generated conformations are often evaluated by using components of molecular mechanics force fields, with a variety of treatments of electrostatics and solvation.[18,25,28] Knowledge-based potentials have also been used in combination with conformational sampling methods, as have energy functions that combine molecular mechanics force-field terms with statistical potentials.[17]

Several studies have also combined knowledge-based and de novo methods in a hybrid approach to loop modeling. Mas et al.[29] used a combination of database and conformational search methods to model the hypervariable loops in an antibody; the conformational search method was used both to verify conformations selected from the set of canonical structures and to model de novo the conformation of one loop for which canonical conformations could not be reliably identified. Martin et al.[30] proposed a method that relied on conformational search for short (<5 residues) loops and database search for medium (6–7 residues) loops; long loops were predicted by a hybrid approach in which the central residues in a database-selected conformation were reconstructed by conformational search. In a sequential approach to combining database and conformational searching, VanVlijmen and Karplus[11] demonstrated that performance of a database method could be improved by subsequent optimization and ranking with a molecular mechanics potential. Deane and Blundell[13] described a combined approach that uses the consensus predictions of a knowledge-based and de novo loop modeling method. Sudarsanam et al.[31] used exhaustive sampling of dimers of discrete set of (ϕ, ψ) angles but derived this angle set from angles sampled in known protein structures.

The Rosetta-based method described here is a novel approach to combining database-derived conformations and de novo prediction for loop modeling. In the Rosetta method, originally developed for de novo prediction of entire protein domains, structures sampled by local sequences are approximated by the distribution of structures seen for those short sequences and related sequences in the Protein Data Bank (PDB). These fragments are then assembled in a Monte Carlo search strategy using a scoring function that favors nonlocal properties of native protein structures such as hydrophobic burial, compactness, and pairing of β-strands. Using only primary sequence information, successful de novo Rosetta predictions of entire protein domains yield models on the order of 3–7 Å Cα root-mean-square deviation (RMSD) to native for substantial fragments (>60 residues) of the query sequence.[32,33]

The fragment assembly strategy used by Rosetta is currently perhaps the most successful method for de novo structure prediction, and it may be particularly well suited to modeling SVRs in proteins. By building conformations from smaller fragments, the problem of adequate sampling in the database for longer loops encountered in knowledge-based methods can be potentially overcome, while still restricting the conformational search to a tractable size—a problem encountered by de novo loop modeling methods for longer segments. Furthermore, the fragment buildup strategy allows regular secondary structure to be easily incorporated in predictions for longer SVRs, overcoming a limitation of many de novo loop modeling strategies. Consequently, the method is not limited to protein loops but is applicable to SVRs of any size. A final novel approach used in the current method is the simultaneous modeling of side-chain and backbone conformations using idealized geometry and a rotamer approximation of side-chain conformation. The use of rotamer representations of

the side-chains during optimization of backbone conformations further reduces the complexity of the search space while allowing an atom-based potential function to be used for optimization.

MATERIALS AND METHODS

The SVR modeling method described here uses the Rosetta scoring function and fragment insertion methodologies developed for de novo structure prediction.[3,4] In brief, a customized library of fragments for each three-and nine-residue window in the protein sequence is selected from a database of known protein structures on the basis of local sequence similarity and similarity between the known and predicted secondary structure. These fragments are then assembled by using a Monte Carlo simulated annealing search strategy in which fragments are randomly inserted into the protein chain by replacing the backbone torsion angles in the protein chain with those in the fragment. The resulting protein conformation is then evaluated according to a protein database-derived scoring function that rewards native-like protein properties (see below). In the standard Rosetta protocol for de novo structure prediction, a reduced representation of the protein is used: backbone heavy atoms and Cβ atoms are explicitly included, whereas side-chains are represented by a single centroid. As described below, structure prediction simulations used a combination of this reduced protein representation and an all heavy atom representation with explicit side-chain rotamers.

Database Search

Like the de novo Rosetta protocol, the modeling strategy used here also uses a combination of database-derived fragments that approximate local conformational preferences and a Monte Carlo simulated annealing minimization of a target energy function. Given a sequence alignment between the query and a parent homologue of known structure, the protein structure is divided into template regions and SVRs, which are defined as sections of the chain whose torsion angles cannot be approximated by using those of the parent structure and may include loops, larger insertions, regions of uncertain alignment, and aligned regions where significant structural perturbations are expected. Template regions include all residues whose backbone torsion angles and Cartesian coordinates are taken directly from the parent structure and held fixed throughout the simulations. Cofactors and ligands present in the homologue structure are included in the fixed template coordinates. As in the standard Rosetta protocol, a customized library of three- and nine-residue fragments is selected for the protein sequence and used as described below.

For each SVR of 15 residues or less, an additional customized library of 200–300 possible conformations for the SVR is extracted from the protein structure database. The scoring function used to evaluate these initial loop conformations is a modified form of the scoring function used to generate fragment libraries in the de novo Rosetta protocol and ranks protein segments according to four criteria: 1) sequence profile–profile similarity over the SVR, 2) similarity of the predicted and known secondary structure over the SVR, 3) similarity between secondary structure of template residues adjacent to the SVR in the query and the candidate database conformation, and 4) geometric fit of the database conformation to the template. The process proceeds in two stages. First, a large database representative of the diversity in the nonredundant PDB is coarsely screened for the top 2000 segments that score well by a composite of the four criteria. To select a final set from this pool of 2000, the segments are ranked first by one of the criteria listed above; the top 250 conformations are then reranked by a second criteria, and the top 25 conformations are retained. The culling process is then repeated with use of other criteria. A variety of orders of ranking criteria are used in the culling sequence, and then all the sets are combined into the final library with duplicates removed. The resulting database of initial conformations is comprised of a narrow set of segments when there is a consensus among the methods and a diverse set when there is a lack of consensus, consistent with the philosophy that a diverse set is preferable to a narrower but potentially incorrect set.

Conformational Search

Multiple independent Monte Carlo-simulated annealing optimizations are conducted from different random seeds for each SVR. For each individual simulation, an initial database conformation is selected randomly from the customized library and built onto the fixed template by requiring chain connectivity at either the N- or C-terminal template-SVR junction and allowing discontinuities in the protein backbone at the other junction. The selection of the junction for chain discontinuity is random for each simulation. Initial conformations for SVRs $>$ 15 residues in length are generated by using the standard Rosetta de novo protocol of randomly inserted nine-residue fragments from the customized library into an initially extended protein chain.[3,4] The generation of these initial conformations is conducted in the context of the template but without evaluation of the geometric fit of the variable region to the template.

SVRs greater than seven residues in length are then subjected to Monte Carlo optimization by using a move set of three- and nine- (for SVRs longer than 15 residues) residue fragments. Fragments are either selected randomly from the library or prescreened to bias selection toward fragments that improve the geometric fit of the SVR to the template stems as measured by a gap penalty (see below). Fragment insertions are also combined with a "wobble" operation in which backbone (ϕ, ψ) angles within or adjacent to the fragment insertion site are perturbed to minimize a cost function consisting of the gap penalty and the torsion potential (see below). In addition to fragment insertions, backbone conformations of SVRs are also modified by using random small changes in ρ, ψ angle pairs of individual residues or compensating changes of (ψ_{i-1}, ϕ_i) pairs. These random angle perturbation moves are also combined with the wobble operation. The combination of

TABLE I. Short Loop Reconstruction Results

				Best score[b]					Best RMSD-G		
Protein	Length	Residues	Native score	RMSD-L (Å)	RMSD-G (Å)	Rank[c]	Score	Enrichment	RMSD-L (Å)	RMSD-G (Å)	Score
2act	8	198–205	−66	2.38	3.79	192	−694	2.84	1.42	2.10	−677
2apr	8	76–83	−914	1.06	2.54	226	−930	1.33	0.33	0.53	−912
2fb4	7	H26–H32	−949	1.12	1.79	15	−961	4.80	0.64	0.97	−958
2fbj	7	H100–H106	−1772	0.34	0.98	1	−1744	4.89	0.34	0.98	−1744
3blm	5	131–135	−1191	0.18	0.43	84	−1215	4.89	0.18	0.21	−1200
3dfr	4	20–23	−1215	0.44	0.80	84	−1237	3.20	0.19	0.34	−1215
3dfr	5	89–93	−1215	0.78	0.96	21	−1256	0.71	0.42	0.83	−1234
3dfr	5	120–124	−1215	0.64	0.76	98	−1231	1.69	0.27	0.34	−1181
3grs	7	83–89	−1447	0.61	0.97	23	−1484	6.04	0.29	0.30	−1464
3sgb	9	E199–E211	−1422	0.80	1.10	6	−1393	1.24	0.66	0.90	−1371
5cpa	7	231–237	−824	0.89	1.22	36	−847	3.38	0.54	0.77	−821
8abp	6	203–208	−913	0.44	0.56	35	−949	4.44	0.27	0.31	−933
8tln	7	E32–E38	−1180	2.10	2.62	71	−1220	0.44	0.84	1.24	−1168
8tln	8	E248–E255	−1221	0.75	1.52	11	−1250	5.78	0.42	0.76	−1239

[a]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations in the 15% best scoring population compared with the entire population.
[b]Best-scoring conformation of 500 independent optimizations.
[c]Rank order by RMSD-G of the best-scoring conformation.

TABLE II. Accuracy of 4-, 8- and 12-Residue Segment Predictions[†]

	Best score[a]			Best RMSD-G	
Length	Mean (median) RMSD-L (Å)	Mean (median) RMSD-G (Å)	Mean enrichment[b]	Mean (median) RMSD-L (Å)	Mean (median) RMSD-G (Å)
4	0.42 ± 0.05 (0.31)	0.69 ± 0.06 (0.54)	2.9 ± 0.3	0.21 ± 0.02 (0.18)	0.30 ± 0.03 (0.25)
8	0.97 ± 0.10 (0.79)	1.45 ± 0.14 (1.20)	3.6 ± 0.2	0.50 ± 0.03 (0.47)	0.67 ± 0.05 (0.59)
12	2.23 ± 0.15 (2.29)	3.62 ± 0.31 (3.65)	2.6 ± 0.2	1.28 ± 0.08 (1.30)	1.66 ± 0.10 (1.76)

[†]Reported uncertainties are the standard error of the mean.
[a]Best-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations in the 15% best-scoring population compared with the entire population.

the various types of conformation modification operators is selected so that moves become progressively more local and less globally perturbing during the course of the simulation. The conformational search is conducted by using a Monte Carlo search followed by a two-stage Monte Carlo minimization strategy.[34] In the first stage, a single line minimization along the gradient is conducted for each attempted move, whereas in the second, the variable metric method of Davidon–Fletcher–Powell is used to find the nearest local minimum of the potential energy surface following each initial conformation modification.[35]

Following the optimization using centroid side-chain representations, full-atom coordinates of the side-chains are generated by using a simulated annealing algorithm and a backbone-dependent rotamer library.[36,37] Additional optimization using small backbone torsion angle perturbations and the full-atom potential (see below) is conducted by using the Monte Carlo minimization strategy, iteratively updating the backbone and side-chain conformations. After modification of the backbone torsion angles, side-chain coordinates are updated by adjusting χ angles to their preferred values for the particular rotamer given the new backbone torsion angles. Rotamers at each position in the SVR and spatially adjacent template regions are then updated, in a randomly selected order, by using the rotamer at each position that gives the best energy according to the full-atom potential (see below). At the conclusion of the energy minimization protocol, the side-chains at all positions are completely repacked by using the simulated annealing protocol.

Energy Function

The standard Rosetta potential is derived from a Bayesian treatment of native protein structures and is comprised of two general classes of terms.[3,4] The first class of terms, which describe the probability of a structure independent of sequence, reward native-like arrangements of secondary structure and overall compactness. A second class of terms, describing the probability of a particular sequence given a structure, reward burial of hydrophobic residues and specific pair interactions and penalize van der Waals clashes. For the portions of simulations using reduced side-chain representations, this standard Rosetta potential is modified to include a gap penalty that penalizes chain discontinuities. This gap penalty is calculated as the RMSD between the fixed coordinates of the first

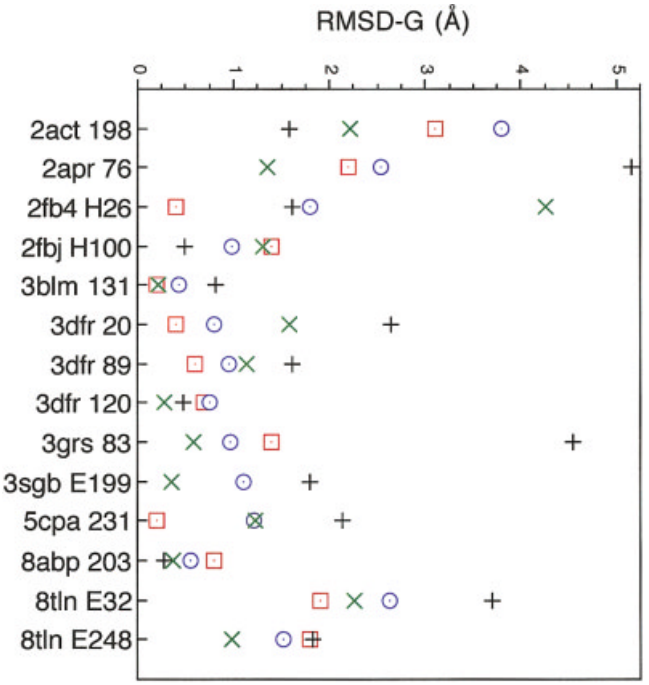

Fig. 1. Comparison of accuracies for loops in Table I predicted by four different methods. The accuracies of previously published predictions by three methods are shown as black plus symbols (Van Vlijmen and Karplus[11]), green x symbols (Fiser et al.[17]), and red squares (Deane and Blundell[13]). Prediction accuracies from the current work are shown as blue circles. RMSD-G is the RMSD of loop residues after superposition of the stem residues (see text). For the Fiser et al. and Rosetta predictions, all backbone heavy atoms (N, CA, C, O) are included in RMSD-G evaluations. For the predictions of Van Vlijmen and Karplus and Deane and Blundell, only N, CA, and C atoms are included in RMSD-G evaluations. Modeled segments are identified by PDB code and first residue.

template residue at each discontinuous template-variable segment junction and the coordinates of this residue determined from the dihedral angles and coordinates of the adjacent variable residue. This same gap penalty score is used in evaluating geometric fit of database conformations to the template.

For all backbone conformation modifications that introduce (ϕ, ψ) angles not taken directly from the fragment library (i.e., random angle perturbation, "wobble" operations, and minimization), torsion angles are evaluated by using a secondary structure-dependent torsion potential.[38] This knowledge-based potential is derived from a nonredundant set of protein X-ray structures of >2.5 Å resolution. For each of the 20 amino acid types in each of three secondary structure types (helix, strand, and other as defined by DSSP[39]), the frequency of (ϕ, ψ) pairs was determined for 10° × 10° bins. Probability distributions were smoothed by using pseudocounts, and the potential was calculated by taking the logarithm of the interpolated probabilities. Randomly selected small angle perturbations, which move backbone conformations away from those represented in the fragment libraries, are discarded according to a Boltzman criterion if they represent an increase in this torsion energy. For moves involving perturbation of backbone angles to minimize a cost function, the torsion potential was included in the target cost function. Backbone ω angles are only modified by fragment insertion and are not evaluated in the torsion potential.

The rotamer packing and rotamer replacement algorithms use the full-atom potential of Kuhlman and Baker[37] with the following modifications: The hydrogen bond potential term used by Kuhlman and Baker is replaced with a hydrogen bond potential derived from PDB statistics. The energies of main-chain–main-chain, side-chain–side-chain, and main-chain–side-chain hydrogen bonds are estimated as a function of the donor and acceptor hybridization and the geometry of the hydrogen bond based on the observed distribution of these parameters in protein crystal structures.[40] The full-atom potential was also supplemented with the torsion potential and gap penalty that was incorporated into the standard Rosetta potential (see above). The complete full-atom potential is thus comprised of 1) the attractive portion of the 12-6 Lennard–Jones potential, 2) a linear repulsive term used in place of the repulsive portion of the 12-6 potential, 3) backbone-dependent internal free energies of the rotamers estimated from PDB statistics, 4) solvation energies calculated by using the model of Lazaridis and Karplus,[41] 5) a knowledge-based pair potential, 6) the hydrogen-bonding potential described above, 7) the knowledge-based backbone torsion potential described above, and 8) the gap penalty described above. This potential is used both for iterative optimization of the SVR backbone and all rotamers and to rank the final population of conformations.

CASP Predictions

For CASP predictions, alignments between the query and parent homologue sequences were generated by using a Smith–Waterman algorithm using PSI-BLAST[42] profile–profile scores, similarity of predicted and known secondary structure, and structural and functional constraints implied by FSSP/DALI topological family sequence profiles.[43] Penalties for insertions and deletions were assigned in a structure-dependent manner using known protein structures to assess the probability of an insertion or deletion of a particular length given the spatial and geometric constraints imposed by flanking residues in the parent structure.[44] Given the alignment between the target sequence and a homologous parent, gaps, insertions, and regions of low-confidence alignment were treated as SVRs.

All SVRs in the target were simultaneously optimized. From the set of resulting models, conformations for each SVR were ranked independently in the context of the fixed template, discarding any conformations that resulted in knots or large-chain discontinuities and retaining the lowest-energy conformations. Combinations of low-energy conformations for each SVR were then evaluated simultaneously to identify low-energy combinations of conformations for all SVRs. The modeling strategy used for CASP 4 targets was an earlier version of the current method and differs from the method described above in several aspects. The primary differences are as follows: 1) the Monte Carlo plus minimization strategy was not used, and all optimization occurred by Monte Carlo search, 2) optimizations

TABLE III. Four-Residue Segment Reconstruction Predictions

Protein	Residues	Sequence	Best score[a]		Enrichment[b]	Best RMSG	
			RMSD-L	RMSD-G		RMSD-L	RMSD-G
1aaj	82–85	FTEA	0.14	0.25	4.58	0.13	0.21
1ads	99–102	LKLD	0.22	0.28	2.44	0.16	0.18
1bam	92–95	PIDV	0.61	1.07	5.82	0.62	1.03
1bgc	40–43	HKLC	0.89	1.03	1.47	0.51	0.55
1cbs	21–24	VLGV	0.12	0.29	3.64	0.16	0.18
1fkf	42–45	RNKP	0.16	0.28	3.51	0.11	0.13
1frd	59–62	DQSD	1.07	1.75	0.40	0.24	0.29
1gpr	123–126	NVPS	0.55	0.97	3.20	0.28	0.35
1iab	100–103	FYHE	0.58	0.75	2.89	0.17	0.28
1mba	97–100	GFGV	0.23	1.01	3.56	0.15	0.22
1nfp	37–40	EDTS	1.20	1.49	1.56	0.53	0.58
1pbe	117–120	GATT	0.29	0.57	3.47	0.19	0.25
1pda	139–142	RRPD	0.23	0.32	2.71	0.15	0.17
1pgs	226–229	LGAL	0.83	1.38	0.76	0.17	0.28
1plc	74–77	LSNK	0.37	0.44	3.47	0.26	0.27
1ppn	42–45	TGNL	0.19	0.23	1.29	0.15	0.19
1prn	66–69	GNAA	0.26	0.39	2.98	0.21	0.24
1rcf	111–114	QRGG	0.16	0.25	2.67	0.16	0.25
1tca	287–290	AGPK	0.27	0.43	1.11	0.17	0.22
1thw	194–197	PGSS	1.09	1.28	0.53	0.18	0.28
1tib	46–49	KADA	1.18	1.38	1.24	0.10	0.16
1tml	42–45	FAHH	0.36	0.50	4.22	0.36	0.50
1tys	131–134	SAWN	0.67	1.15	3.07	0.21	0.56
1xif	82–85	TGMK	0.30	0.41	1.87	0.14	0.19
1xnb	30–33	WSNT	0.20	0.51	5.20	0.39	0.44
2cmd	163–166	GKQP	0.28	0.60	2.00	0.19	0.22
2cy3	101–104	KDKK	0.33	0.55	2.53	0.16	0.25
2cyp	127–130	RCGR	0.47	0.81	2.18	0.20	0.33
2cyr[c]	69–71	HAK	0.23	1.12	3.07	0.16	0.45
2exo	161–164	DPTA	0.48	1.03	4.67	0.38	0.40
2sga	44–47	LGFN	0.33	0.43	5.24	0.15	0.25
2sil	220–223	LPSG	0.32	0.66	1.16	0.19	0.26
2tgi	72–75	ASAS	0.34	0.50	5.02	0.15	0.19
3cla	27–30	HRLP	0.13	0.39	0.58	0.11	0.25
4enl	335–338	EKKA	0.25	0.54	6.62	0.17	0.28
4gcr	116–119	FHLT	0.41	0.58	2.49	0.15	0.21
5fd1	81–84	ITEK	0.21	0.53	0.62	0.21	0.31
5p21	75–78	GEGF	0.46	0.87	1.51	0.16	0.28
7rsa	47–50	VHES	0.11	0.18	5.02	0.12	0.17
8abp	55–58	ASGA	0.13	0.20	3.82	0.12	0.16

[a]Top-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations found in the 15% best-scoring population compared with the entire population.
[c]Three residues only; conformation A of Lys 71 was used as native reference.

generally used only centroid representations of side-chains, although complete heavy atom side-chain coordinates were generated for the final models using the simulated annealing rotamer-packing algorithm, and 3) coding errors present at the time of CASP 4 limited the effectiveness of the optimization. CASP 5 targets used the standard protocol described here, but final loop conformations were selected manually from the top ranked conformations (ranked by energy or cluster size in single-linkage cluster analysis) to eliminate loop combinations resulting in models with steric clashes and/or knots. In addition, although homologous proteins were excluded from the structure database for segment reconstruction tests, homologous proteins were used when available for CASP predictions.

Evaluation of Model Accuracy

To evaluate both the accuracy of the SVR itself, as well as the accuracy of the SVR orientation with respect to the rest of the protein, we report two metrics of model accuracy. RMSD-L is a measure of the model accuracy in a local context and is the RMSD between the model and native over all backbone heavy atoms in the SVR after optimal superposition of the SVR. RMSD-G reports the correctness of both the predicted SVR conformation and its orientation with respect to the template and is the RMSD between the

TABLE IV. Eight-Residue Segment Reconstruction Predictions

Protein	Residues	Sequence	Best score[a]		Enrichment[b]	Best RMSG	
			RMSD-L	RMSD-G		RMSD-L	RMSD-G
1351	84–91	LSSDITAS	1.37	1.63	2.44	0.59	0.66
1alc	34–41	SGYDTQAI	0.79	1.09	5.29	0.30	0.46
1art	88–95	FGKGSALI	2.08	3.16	3.82	1.00	1.46
1btl	50–57	DLNSGKIL	0.43	0.63	6.04	0.27	0.41
1cbs	55–62	STTVRTTE	0.52	0.76	4.40	0.35	0.50
1clc	313–320	FRPYDPQY	0.29	1.01	6.27	0.38	0.50
1ddt	127–134	FGDGASRV	2.10	2.94	1.02	0.44	0.69
1fnd	262–269	LKKDNTYV	0.36	0.51	4.22	0.25	0.28
1gky	72–79	QFSGNYYG	0.38	0.72	5.64	0.41	0.48
1gof	606–613	VPSDSGVA	0.76	1.11	3.07	0.54	0.60
1hbq	31–38	DPEGLFLQ	1.17	2.37	1.91	1.21	1.32
1hfc	142–149	SNVTPLTF	0.56	0.68	1.33	0.46	0.49
1iab	48–55	RTTESDYV	2.11	2.92	2.18	0.77	0.83
1ivd	413–420	EGKSCINR	0.97	1.36	4.84	0.51	0.65
1lst	101–108	PIQPTLES	0.47	1.02	2.93	0.35	0.54
1mpp	74–81	TYGTGGAN	1.57	2.55	2.76	0.67	0.73
1nar	192–199	FSNQQKPV	1.04	1.27	4.71	0.50	0.73
1oyc	80–87	GGYDNAPG	0.60	0.68	6.09	0.41	0.51
1phf	85–92	CPFIPREA	0.71	1.12	2.31	0.71	1.12
1poa	71–78	CSQGTLTC	1.15	1.80	4.40	0.50	0.89
1prn	150–157	DPDQTVDS	2.38	2.76	3.20	0.63	0.69
1sbp	107–114	KQIHDWND	0.32	1.07	3.87	0.30	0.45
1thw	18–25	SKGDAALD	0.62	1.01	0.67	0.62	1.01
1tml	187–194	NTSNYRWT	0.75	1.51	3.78	0.37	0.50
1tys	83–90	WADENGDL	0.46	0.86	2.22	0.37	0.47
1xnb	99–106	KSDGGTYD	0.34	0.72	3.87	0.24	0.39
2ayh	123–130	YTNGVGGH	1.32	1.61	3.29	0.31	0.37
2cmd	270–277	LGKNGVEE	1.49	2.55	6.62	0.64	0.95
2ctc	89–96	DYGQDPSF	0.89	1.35	4.53	0.87	1.18
2dri	161–168	PADFDRIK	1.19	1.40	3.60	0.33	0.51
2exo	262–269	MQVTRCQG	0.35	0.41	1.96	0.35	0.41
2ran	26–33	MKGLGTDE	2.33	3.26	2.67	0.90	1.17
2sga	32–43	TTGGSRCS	0.88	1.41	4.93	0.48	0.65
2sns	17–24	AIDGDTVK	0.49	0.59	4.98	0.51	0.57
3cox	109–116	GRGVGGGS	0.78	0.84	3.42	0.38	0.44
3grs	424–431	ANKEEKVV	1.62	3.20	2.84	0.41	0.49
4enl	24–31	TTEKGVFR	0.85	1.43	1.29	0.47	0.69
4fxn	88–95	YGWGDGKW	1.61	1.66	4.18	0.59	1.22
5p21	45–52	VIDGETCL	0.28	0.45	3.47	0.19	0.26
8dfr	65–72	RPLKDRIN	0.41	0.67	3.64	0.51	0.63

[a]Top-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations found in the 15% best-scoring population compared with the entire population.

model and native of all heavy backbone atoms in the SVR after optimal superposition of three adjacent stem residues on each side of the SVR. For short loops, RMSD-G is the critical measure of accuracy. Most interactions of atoms in short loops are with the template portion of the protein, and correctly predicting the orientation of the loop with respect to the protein core is the primary goal of modeling. For longer SVRs, including insertions comprising intact structural modules, RMSD-L becomes an increasingly relevant metric of model accuracy. Although correct prediction of both the structure of the segment itself and its orientation with respect to the protein core is the end goal of SVR modeling, this goal is generally beyond the capabilities of current methods. Consequently, the accuracy with which SVR structure can be predicted without requiring correct global orientation is a relevant quality indicator. Furthermore, models with correct structure but incorrect orientation likely still include useful structural information.

For purposes of evaluating SVR modeling in CASP targets, a third metric, RMSD-E, is also evaluated to quantify the structural accuracy of the local environment in which the SVR is predicted. RMSD-E is the RMSD between the model and native conformations evaluated over the three stem residues N- and C-terminally adjacent to the SVR after optimal superposition of these residues. For the segment reconstruction tests, the "template" corresponds exactly to the native protein backbone structure, and all RMSD-E values are 0 Å. For CASP targets and, in fact any realistic comparative modeling problem, both

TABLE V. Twelve-Residue Segment Reconstruction Predictions

Protein	Residues	Sequence	Best score[a]		Enrichment[b]	Best RMSG	
			RMSD-L	RMSD-G		RMSD-L	RMSD-G
1541	153–164	NVRSYARMDIGT	0.99	1.51	2.89	0.67	0.97
1arp	201–212	LDSTPQVFDTQF	0.49	0.77	0.93	0.60	0.76
1ctm	9–12	YENPREATGRIV	3.81	5.64	2.00	1.16	2.31
1dts	41–52	SGSEKTPEGLRN	1.58	4.97	1.02	1.33	2.52
1eco	35–46	MAKFTQFAGKDL	3.13	4.15	2.53	0.64	0.94
1ede	150–161	CLMTDPVTQPAF	0.86	0.89	4.84	0.86	0.89
1ezm	122–133	FGDGATMFYPLV	2.06	4.46	3.11	2.06	2.18
1hfc	165–176	RGDHRDNSPFDG	2.33	3.80	1.20	1.79	2.46
1ivd	365–376	TISKDLRSGYET	2.72	4.23	1.38	1.04	1.26
1msc	9–20	LVDNGGTGDVTV	2.75	9.18	2.84	1.85	2.43
1onc	23–34	MSTNLFHCKDKN	2.82	4.03	1.11	1.61	1.72
1pbe	129–140	LHDLQGERPYVT	1.83	2.90	3.38	0.76	0.92
1pmy	77–88	KCAPHYMMGMVA	3.03	4.08	4.00	1.28	1.58
1prn	15–26	VEDRGVGLEDTI	3.16	6.44	3.91	1.98	2.31
1rcf	88–99	TGDQIGYADNFQ	2.15	3.60	1.87	1.83	2.04
1rro	17–28	ECQDPDTFEPQK	2.05	2.66	2.00	0.77	1.02
1scs	199–210	IKSPDSHPADGI	1.85	3.17	2.40	1.06	1.18
1srp	311–322	SDVGGLKGNVSI	1.12	1.16	4.00	0.97	1.10
1tca	305–316	AVGKRTCSGIVT	2.42	3.75	3.91	1.65	1.84
1thg	127–138	WIYGGAFVYGSS	2.89	4.17	2.04	1.89	2.29
1thw	178–189	PDAFSYVLDKPT	2.28	2.83	0.58	1.52	2.09
1tib	99–110	EINDICSGCRGH	2.69	3.12	1.73	0.78	0.94
1tml	243–254	STTNTGDPMIDA	2.97	5.80	3.64	1.75	2.19
1xif	203–214	IERLERPELYGV	1.34	1.64	3.78	0.64	1.08
2cpl	145–156	FGSRNGKTSKKI	3.64	7.45	1.07	1.79	2.07
2cyp	191–202	WGAANNVFTNEF	2.18	2.84	2.13	1.61	2.29
2ebn	136–147	YQTPPPSGFVTP	2.56	3.28	2.09	0.64	0.94
2exo	293–304	LVWDASYAKKPA	1.00	1.51	0.84	0.66	0.96
2pgd	361–372	WRGGCIIRSVFL	2.61	4.32	2.44	1.44	2.04
2rn2	90–101	WKTADKKPVKNV	4.23	7.09	5.47	1.55	2.26
2sil	255–266	ETKDFGKTWTEF	0.53	0.68	5.64	0.53	0.68
2sns	111–122	VAYVYKPNNTHE	1.89	3.14	4.40	2.37	3.02
2tgi	48–59	CPYLWSSDTQHS	2.19	2.86	4.13	1.72	1.96
3b5c	12–23	IQKHNNSKSTWL	3.05	5.22	2.27	0.87	1.04
3cla	176–187	AKYQQEGDRLLL	1.20	1.49	4.49	1.20	1.49
3cox	478–489	VPGNVGVNPFVT	1.65	1.97	1.38	1.26	1.60
3hsc	72–93	RLIGRRFDDAVV	0.55	0.70	2.53	0.51	0.64
451c	16–27	HAIDTKMVGPAY	3.47	5.59	1.51	1.75	2.53
4enl	372–383	SHRSGETEDTFI	2.30	3.69	1.96	1.36	1.79
4ilb	46–57	FVQGEESNDKIP	2.92	4.02	2.04	1.48	2.20

[a]Top-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations found in the 15% best-scoring population compared with the entire population.

alignment errors and template perturbations contribute to the accuracy of the template from which SVRs are modeled, and these modeling errors result in non-zero RMSD-E values. RMSD-E measures the accuracy only of the stem residues sequentially adjacent to the SVR and does not reflect the structural accuracy of other spatially adjacent residues. Consequently, small RMSD-E values for regions modeled as SVRs in homology models indicate only that the local geometry constraining the ends of the SVR is approximately correct.

RESULTS

The SVR modeling method described here is intended to comprise part of a complete modeling strategy for structure prediction by comparative modeling and fold recognition and was, in fact, applied in combination with an alignment algorithm to generate complete models for all targets in CASP 4 and CASP 5 for which a homologous protein of known structure could be identified. The double-blind CASP experiment offers a realistic test of comparative modeling methods because both alignment errors and structural deviations between a query sequence and the parent structure degrade the accuracy of the local environment in which SVRs must be modeled. However, the blind evaluation of CASP targets is conducted without knowledge of which portions of the model were generated by alignment and which were modeled as structurally divergent. To supplement the analysis of model quality pro-

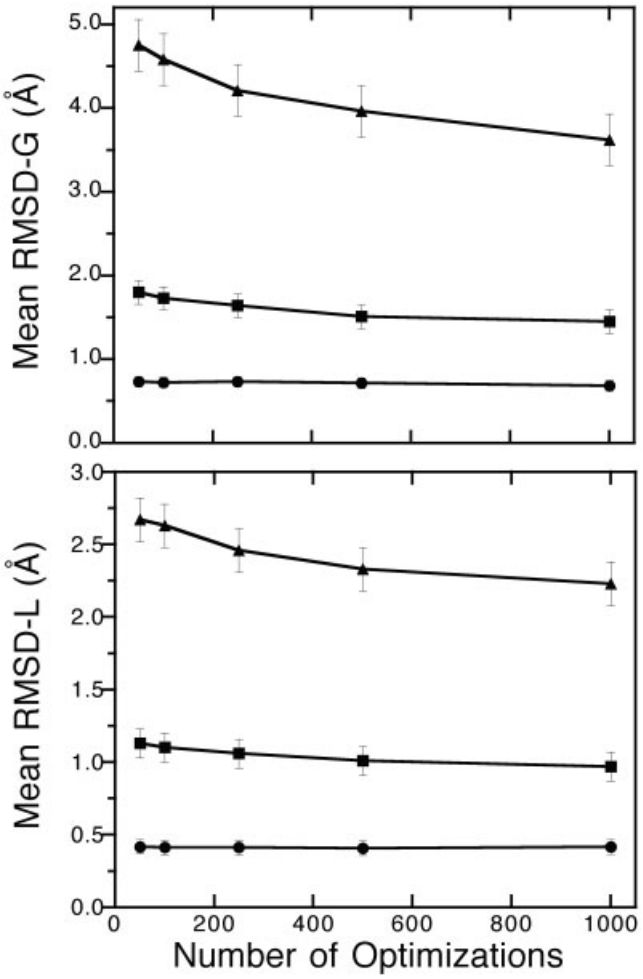

Fig. 2. Mean prediction accuracy as a function of number of independent optimizations.

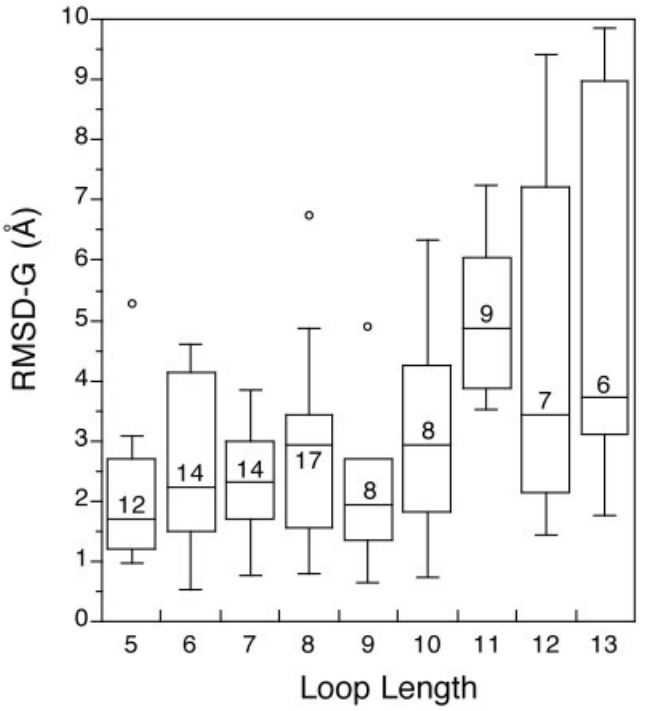

Fig. 3. Box plots of distributions of RMSD-G values for SVRs of lengths 5–13 residues in CASP 5 targets. Only SVRs modeled in the context of reasonably accurate environments (RMSD-E < 1.5Å; see Materials and Methods) are included in the figure. The number in each box indicates the number of modeled SVRs contributing to each distribution.

vided by the CASP assessors and to assess the performance of the Rosetta-based method in the context of realistic modeling errors, we report here the accuracy of the CASP 4 and CASP 5 predictions specifically for segments modeled as SVRs. Complete lists of all regions modeled as SVRs in both CASP 4 and CASP 5 targets for which structures have been released, along with the template and prediction accuracies, are reported here.

In addition to blind CASP predictions of SVRs made in the context of realistic modeling errors, we also present results of predictions in which segments of proteins of known structure are reconstructed in the context of exact templates. Segment reconstruction, although artificial in the sense that it does not represent a realistic structure prediction problem, does allow the SVR method to be assessed in the absence of propagated errors resulting from incorrect alignment and template perturbation. In addition, segment reconstruction has been used as a standard method for assessment of loop modeling methods and allows direct comparison of different modeling strategies. Notably, in the segment reconstruction predictions here, none of the native side-chain conformations are retained; instead, all side-chains are replaced by using the simulated anneal rotamer-packing algorithm. Consequently, although the template backbone is exact, the template side-chain conformations are not, making the segment reconstruction test somewhat more realistic.

Prediction of Short Protein Loops

Results of segment reconstruction predictions made for sets of surface-exposed protein loops, selected and previously predicted by other authors, are given in Tables I and II. The fourteen loops in Table I, varying in length from four to nine residues, are provided as representative examples of predictions for short to medium loops. Several other groups have made predictions for these same segments, allowing direct comparison of several methods on identical examples (Fig. 1). Table II summarizes results obtained for 40 loops each of 4, 8, and 12 residues. Results for all individual predictions in these sets are given in Tables III–V.

For short loops, the Rosetta method effectively samples low RMSD-G conformations. For 38 of 44 loops in the 4- to 5-residue range, conformations <0.5 Å RMSD-G are sampled; in the 7- to 9-residue range, conformations <1 Å are sampled in 40 of 49 cases; and for 30 of 40 12-residue loops, conformations < 2.2 Å are sampled. In most cases, conformations that have energies equal or better than the native loop conformation are sampled (Table I). The effectiveness of the sampling

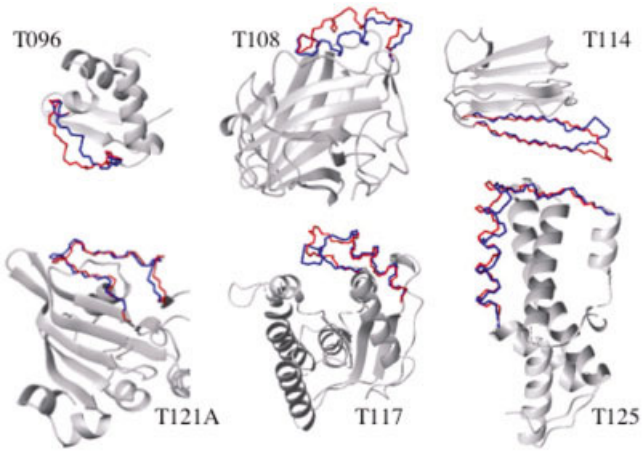

Fig. 4. Top scoring conformations for representative long segment reconstructions. The backbone of the modeled region is shown in blue (native conformation) and red (predicted conformation). The remainder of the backbone structure is shown in gray as a ribbon diagram. Protein structure diagrams were made by using MolMol.[49]

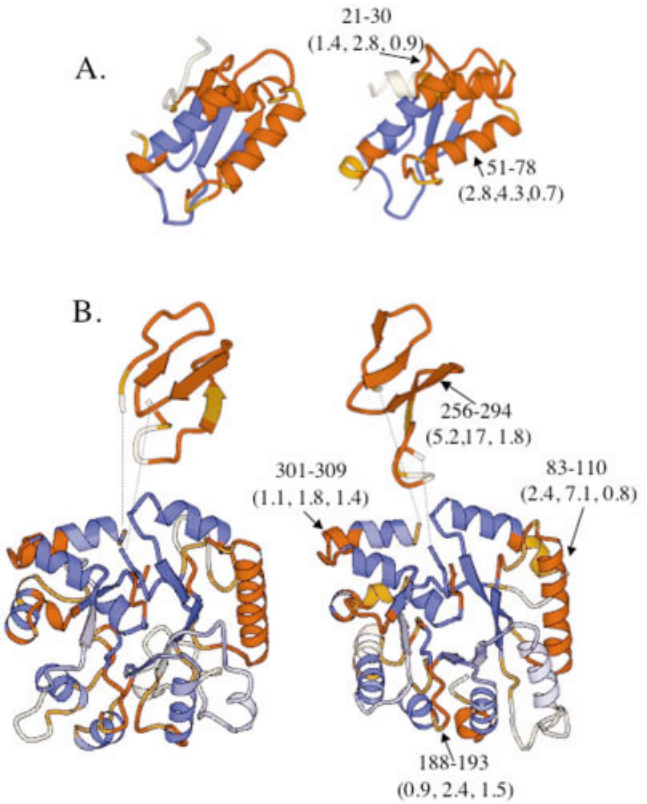

Fig. 6. Selected CASP 5 comparative modeling predictions. Structure diagrams of CASP 5 targets T0130 and T186, residues 44–332, are shown in panels A and B, respectively. The experimental structure is shown on the left, and the first-ranked model is on the right. Regions colored in shades of blue were modeled by using coordinates of a homologue of known structure, whereas regions in shades of orange were modeled as SVRs. For each target, an optimal subset of superimposable residues was found by using the LGA algorithm.[46] Given this structural superposition, the CA deviation between the model and native structure at each position is indicated by color intensity. Regions in dark orange/dark blue have CA deviations of <2 Å after superposition; regions in medium orange/medium blue have CA deviations between 2 and 4 Å, and regions in pale orange/pale blue have CA deviations > 4 Å. Residues are colored identically in the predicted model and experimental structure diagrams. For T0186 (**B**), residues 256–294 have been independently superimposed by using the LGA algorithm. The dotted lines indicate the stem regions to which the SVR termini are connected. Selected SVRs, indicated by arrows, are identified by residue number. Prediction accuracies for these SVRs are given in parenthesis (RMSD-L, RMSD-G, RMSD-E). See text for details. Protein structure diagrams were generated by using Molscript.[50]

method is further illustrated by examining the mean prediction accuracy as a function of the number of independent optimizations conducted (Fig. 2). An increase in mean prediction accuracy on doubling the number of optimizations from 500 to 1000 is seen only for the 12-residue loops. For short loops, the accuracy of prediction is generally limited by discrimination, although ranking of conformations by the potential function does result in significant enrichment (Tables I and II).

Although accurate predictions are made in the context of the native protein, significantly poorer performance is seen for short loop modeling in CASP targets where local template geometries are less than perfect. In CASP 5, 59 domains were modeled by using homology to a protein of known structure. In the targets for which structures are available, 215 regions of $\leq$13 residues were modeled as SVRs. Of these, 177 are nonterminal segments, with template-imposed geometric constraints similar to those of the segments reconstructed in native proteins. Ninety-seven of these SVRs were modeled in the context of reasonably accurate local templates (RMSD-E $<$ 1.5 Å). The distribution of prediction accuracies for loops meeting these criteria are shown in Figure 3. The mean accuracies of loop predictions are significantly worse than those seen in the segment reconstruction tests, indicating, as noted by many previous authors, that the accuracy of loop modeling in real comparative modeling applications is determined almost entirely by alignment accuracy and template distortions. In addition, loop modeling in real homology models is complicated by the fact that multiple, potentially interacting, loops must be modeled within the same structure.

Prediction of Long SVRs

A motivating goal in developing Rosetta for SVR modeling is to provide a modeling method that is not limited only to short loops but is also applicable to predicting longer insertions and structural differences between homologous proteins. To examine the accuracy of the Rosetta method in predicting conformations of longer SVRs, 10 segments ranging from 13 to 34 residues were selected from CASP 4 comparative modeling targets to be reconstructed in the context of the native protein. For each of the proteins, the region of greatest structural divergence with respect to the closest structural match in the PDB, as determined by the CASP 4 assessors,[45] was selected as the segment to be reconstructed. Unlike the shorter protein loops discussed above, these segments do not necessarily correspond to surface-exposed protein loops. Results for these 10 predictions are given in Table VI, and structures of low-energy conformations for some successful predictions are given in Figure 4. For these longer protein segments, the accuracy

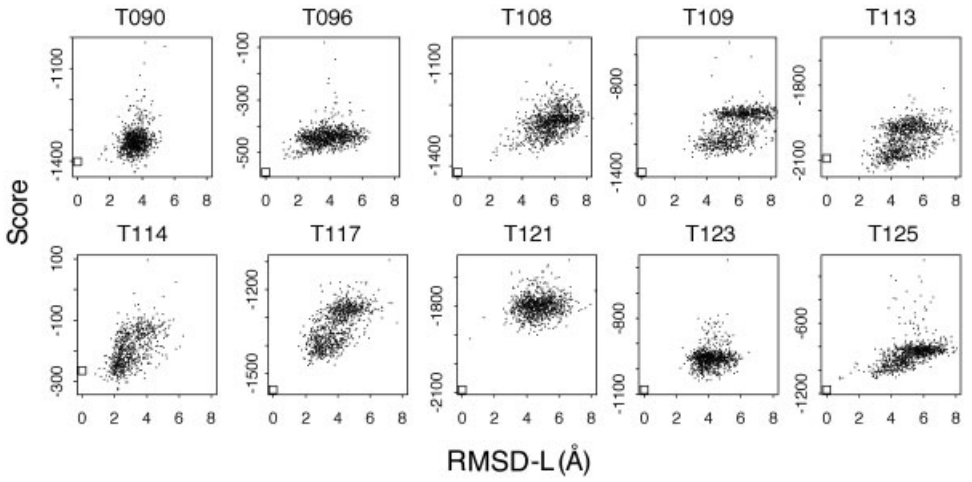

Fig. 5. Conformation discrimination for long SVR reconstructions. The correlation between the final score and RMSD-L is shown for independent optimizations of each predicted segment in Table III. The score of the native segment conformation in each case is indicated by the open square.

TABLE VI. Long Segment Reconstruction Results

				Best score[a]					Best RMSD-L		
Protein	Residues	Length	Native score	RMSD-L (Å)	RMSD-G (Å)	Rank[c]	Score	Enrichment[b]	RMSD-L (Å)	RMSD-G (Å)	Score
T090	77–91	15	−1402	3.54	6.11	474	−1434	1.6	1.33	3.45	−1394
T096	19–31	13	−572	1.21	2.42	3	−521	1.8	1.12	2.82	−460
T108	139–155	17	−1452	2.05	2.84	1	−1415	2.8	2.05	2.84	−1415
T109	48–81	34	−1393	4.00	20.4	49	−1302	2.7	2.62	9.72	−1242
T113	203–223	21	−2092	2.91	3.91	19	−2145	3.0	2.21	2.62	−2080
T114	51–65	15	−264	2.22	3.08	163	−325	3.1	0.84	1.28	−275
T117	138–159	22	−1557	2.19	2.39	24	−1471	2.4	1.60	3.93	−1401
T121	65–82	18	−2089	0.51	0.88	1	−1913	1.5	0.51	0.88	−1913
T123	28–41	14	−1082	3.88	6.95	360	−1047	1.2	2.29	3.78	−996
T125	94–118	25	−1165	0.84	2.48	2	−1076	4.3	0.82	2.23	−1058

[a]Best-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-L conformations in the 15% best-scoring population compared with the entire population.
[c]Rank order by RMSD-L of the best-scoring conformation.

of the predictions is limited both by the conformational search and by discrimination. Native structures show significantly better scores than all sampled conformations in 7 of the 10 examples. In most cases, some correlation between the accuracy of the predicted segment conformations and the evaluated scores is observed (Fig. 5), with an average enrichment of 2.5 ± 0.9 (Table VI), suggesting that additional sampling might result in improved prediction accuracies.

The Rosetta method was also used to make predictions for long SVRs in CASP 4 and CASP 5 targets. Fifty SVRs ranging in length from 14 to 78 residues were predicted in CASP 4 targets for which structures are available, and 74 SVRs ranging in length from 14 to 123 residues were predicted in CASP 5 targets for which structures are available. Table VII gives results of the long SVR predictions in CASP 5 targets that were modeled in the most accurate local template environments (RMSD-E < 2.5Å) and the identity and prediction accuracies for all SVRs in all CASP 4 and CASP 5 targets are given in Tables VIII and IX. As with short loops, performance on long SVRs degrades significantly in the context of realistic modeling errors. In segment reconstruction, 7 of 10 examples have RMSD_L < 3 Å and 5 have RMSD-G < 3 Å. Of 32 long SVRs in CASP 5 targets (Table VII), 12 have RMSD-L < 3 Å, and only 2 have RMSD-G < 3 Å. As noted above (see Materials and Methods), RMSD-E only measures the correctness of stem geometry, not the overall accuracy of the environment. Because longer segments generally have more nonlocal contacts than short, surface-exposed loops, RMSD-E significantly underestimates the true environment error for long SVR predictions. Consequently, examining predictions that have correct local structures, even in the absence of correct orientation is warranted. However, it is important to note that many of the predictions with best local accuracy correspond to single regular secondary structure elements (e.g., a single helix in a TIM barrel that was modeled as an SVR because of alignment uncertain-

TABLE VII. Long SVR Predictions in CASP 5 Targets[†]

Target	Region	Length	RMSD-L (Å)	RMSD-G (Å)	RMSD-E (Å)	End-to-end distance (Å)
T0147	7–20	14	3.85	5.42	1.90	16.3
T0168	298–311	14	3.51	7.29	0.44	8.5
T0149	19–33	15	3.18	6.33	1.35	15.3
T0168	249–263	15	3.54	6.47	1.06	8.0
T0168	279–293	15	4.23	12.15	1.78	4.5
T0169	124–138	15	5.53	12.46	0.85	19.9
T0184	108–122	15	4.25	7.77	0.39	12.6
T0185	176–190	15	4.61	7.68	0.98	17.3
T0186	197–211	15	4.42	7.75	1.37	10.6
T0134	161–176	16	3.18	6.30	2.11	22.1
T0151	84–99	16	0.73	1.46	0.56	5.8
T0154	15–30	16	0.59	1.89	0.36	20.8
T0185	248–263	16	2.07	4.13	1.57	13.4
T0195	58–73	16	3.07	8.02	2.49	21.0
T0165	224–240	17	3.90	8.42	2.17	13.5
T0168	222–238	17	2.14	4.72	2.49	9.3
T0183	96–112	17	0.68	2.39	0.99	21.7
T0184	35–51	17	4.40	8.16	0.92	14.6
T0186	116–132	17	2.63	11.08	2.28	12.4
T0189	16–33	18	4.67	12.38	1.21	4.6
T0193	149–166	18	1.04	3.40	0.56	15.7
T0160	94–112	19	2.30	6.52	0.92	6.2
T0172	56–75	20	1.97	3.08	0.42	11.6
T0133	228–251	24	0.87	1.18	0.41	12.7
T0141	86–111	26	6.42	18.99	2.01	15.2
T0149	98–124	27	3.04	5.39	0.48	12.3
T0130	51–78	28	2.81	4.34	0.65	8.6
T0142	45–72	28	3.41	4.59	0.47	19.7
T0186	83–110	28	2.38	7.12	0.78	11.0
T0165	120–150	31	7.45	12.82	1.23	13.0
T0195	91–124	34	6.58	20.45	1.84	11.4
T0186	256–294	39	5.20	17.27	1.71	9.3

[†]Predictions for SVRs of length 14 and greater submitted as part of first-ranked models in CASP 5. Only predictions made in the context of the most accurate local environments (RMSD-E < 2.5Å) are included in the table.

ties). The end-to-end distance of each SVR in the native protein is reported to help identify those SVRs whose conformations are highly constrained by stem locations.

Despite the difficulty in drawing general conclusions from SVRs in CASP targets, these predictions illustrate the promise of the method for long SVR modeling. Examples from two CASP 5 targets are shown in Figure 6. The template portion of T0130 was generated by alignment to 1fbaA [blue region in Fig. 6(A)]. The two proteins are 23% identical over the structurally superimposable portions, permitting a reasonably accurate alignment to be obtained. Relative to the optimal structural superposition of 1fbaA and experimental T0130 structure, the alignment in the CASP 5 model is 76% accurate and 29% complete. By intent, our alignment algorithm was biased for high specificity at the expense of sensitivity, and we relied on SVR modeling with Rosetta to complete the models. Two internal segments of T0130 were modeled as SVRs: residues 21–30 comprise the C-terminus of the first helix, the N-terminus of the first strand, and the intervening loop; residues 51–78 comprise the second helix and the two long loops connecting this helix to the sheet. Both of these loops are among the best predictions made for loops of their size in CASP 5 targets [Fig. 6(A)].

The template portion of T0186, residues 44–332, was generated by alignment to 1gkpA. The two proteins are 15% identical over structurally superimposable regions, and the alignment used to generate the template is only 46% accurate and 50% complete with respect to the structural superposition. Despite significant alignment errors, four SVRs were modeled in the context of reasonably accurate stem geometries (RMSD-E $\leq$ 1.8 Å). Residues 83–100 comprise one helix on the surface of the TIM barrel along with the connecting loops; residues 188–193 are a loop connecting a helix-strand pair, and residues 301–309 are a loop connecting a helix–helix pair on one end of the barrel. As in T0130, these three SVR predictions are among the most accurate predictions for SVRs of their size in the CASP 5 targets. In addition, when the entire protein model is compared with the experimental structure without concern for the modeling method used, these three SVRs, as well as the two internal SVRs in T0130 discussed above, are of approximately the same accuracy as regions of the model generated by accurate alignment (see Fig. 6).

TABLE VIII. All SVR Predictions In First-Ranked Models of CASP 4 Targets

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0089	1–10	10	1.41	4.07	2.98
T0089	26–31	6	2.55	5.26	3.12
T0089	46–54	9	3.94	10.39	0.25
T0089	64–93	30	10.35	19.83	0.10
T0089	119–159	41	11.16	20.95	0.28
T0089	166–170	5	2.01	4.64	1.73
T0089	198–206	9	2.61	4.47	1.19
T0089	223–230	8	1.94	3.49	1.16
T0089	249–253	5	0.56	1.36	7.25
T0089	263–290	28	4.68	7.69	0.26
T0089	312–331	20	3.70	6.78	0.43
T0089	359–419	61	6.88	16.68	5.97
T0090	1–57	57	13.94	26.44	0.86
T0090	66–70	5	0.58	1.61	1.26
T0090	78–91	14	4.35	14.36	3.99
T0090	149–158	10	2.05	10.17	1.24
T0090	177–209	33	4.55	21.44	2.01
T0092	1–38	38	4.82	20.58	0.45
T0092	51–57	7	1.47	2.97	2.03
T0092	74–82	9	2.56	4.35	0.14
T0092	98–111	14	2.43	5.83	2.31
T0092	116–127	12	1.97	4.14	0.09
T0092	132–144	13	2.37	2.67	0.63
T0092	162–210	49	5.51	26.03	3.20
T0092	218–222	5	0.66	1.47	0.89
T0092	229–234	6	2.66	7.43	2.01
T0096	1–9	9	1.67	4.71	1.09
T0096	21–34	14	4.02	8.14	0.67
T0096	41–46	6	0.45	0.62	0.58
T0096	64–70	7	2.74	4.97	7.82
T0100	25–44	20	8.85	16.36	0.09
T0100	54–58	5	1.48	5.57	1.99
T0100	69–73	5	0.88	1.76	2.18
T0100	75–78	4	1.87	2.48	0.78
T0100	92–114	23	5.94	8.05	1.99
T0100	118–121	4	0.91	1.92	0.54
T0100	131–154	24	3.94	6.69	3.94
T0100	158–166	9	2.50	3.61	3.26
T0100	176–179	4	2.02	3.46	2.02
T0100	182–186	5	1.56	2.14	0.89
T0100	196–199	4	1.68	4.36	2.86
T0100	202–206	5	1.58	2.79	2.16
T0100	216–232	17	3.73	7.69	4.76
T0100	253–262	10	2.83	4.42	3.55
T0100	266–287	22	3.82	13.42	2.78
T0100	299–315	17	5.85	12.50	4.00
T0100	319–323	5	1.91	5.53	2.14
T0100	332–352	21	4.98	24.82	2.63
T0100	361–366	6	2.80	8.65	1.30
T0101	26–47	22	5.69	34.03	1.75
T0101	57–64	8	2.45	6.16	2.70
T0101	70–75	6	1.81	4.34	2.77
T0101	84–94	11	1.93	6.32	0.67
T0101	96–105	10	4.81	8.87	2.31
T0101	117–134	18	3.67	4.35	2.01
T0101	150–178	29	6.52	23.77	1.03
T0101	181–190	10	1.43	2.00	0.76
T0101	196–232	37	8.47	25.72	0.90
T0101	239–244	6	1.61	2.02	1.98
T0101	260–283	24	9.71	16.07	7.95
T0101	300–307	8	2.12	3.08	1.94
T0101	317–328	12	3.77	9.20	3.97

TABLE VIII. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0101	342–362	21	5.26	20.54	4.65
T0101	419–425	7	2.71	9.82	1.60
T0103	1–23	23	3.73	17.70	0.50
T0103	34–36	3	1.23	1.52	2.24
T0103	54–58	5	3.09	5.20	1.04
T0103	66–71	6	1.75	8.55	2.01
T0103	124–128	5	1.63	1.94	1.06
T0103	137–147	11	3.23	3.98	3.26
T0103	172–187	16	5.36	13.04	0.69
T0103	204–213	10	4.22	8.17	0.25
T0103	229–285	57	15.77	17.49	0.37
T0103	318–372	55	10.18	21.59	0.14
T0108	1–39	39	3.58	18.76	1.35
T0108	44–48	5	1.44	3.37	3.55
T0108	53–59	7	1.84	6.27	1.19
T0108	72–86	15	5.79	9.22	1.48
T0108	94–103	10	2.66	2.95	2.73
T0108	147–158	12	2.41	13.81	1.45
T0108	164–174	11	3.31	5.84	0.46
T0108	190–196	7	2.04	5.34	1.12
T0109	1–8	8	1.88	9.59	2.60
T0109	32–44	13	4.48	11.02	0.15
T0109	50–85	36	4.75	13.61	5.45
T0109	118–129	12	1.92	6.40	2.70
T0109	134–158	25	6.29	13.43	3.25
T0109	177–182	6	3.47	9.16	0.31
T0111	1–1	1	0.63	2.04	0.72
T0111	30–33	4	2.09	2.88	0.50
T0111	79–85	7	1.25	1.94	0.64
T0111	139–142	4	1.81	3.78	2.63
T0111	199–203	5	0.46	1.04	3.51
T0111	234–239	6	0.61	1.23	10.19
T0111	261–267	7	1.64	4.20	4.68
T0111	306–310	5	0.58	0.85	0.40
T0112	11–15	5	2.00	3.62	2.86
T0112	48–53	6	2.03	5.39	1.39
T0112	113–122	10	3.79	7.61	0.29
T0112	151–154	4	0.17	1.28	2.61
T0112	160–165	6	0.68	1.92	1.94
T0112	190–194	5	0.34	0.84	0.95
T0112	212–216	5	2.02	2.87	2.16
T0112	220–228	9	3.23	6.16	0.76
T0112	261–264	4	0.98	1.10	1.43
T0112	273–283	11	2.68	4.44	0.77
T0112	336–342	7	2.31	4.00	2.18
T0112	349–352	4	0.61	3.92	1.67
T0113	1–12	12	1.84	5.84	4.65
T0113	96–110	15	4.79	6.14	3.97
T0113	137–146	10	3.37	8.72	2.63
T0113	202–227	26	2.83	4.94	2.14
T0113	241–247	7	0.99	1.25	0.81
T0113	256–261	6	2.29	15.73	0.21
T0114	1–15	15	5.70	24.34	1.25
T0114	59–62	4	2.44	5.77	0.46
T0114	70–72	3	1.58	3.72	0.22
T0115	1–4	4	0.97	3.92	2.81
T0115	9–13	5	1.28	5.27	0.73
T0115	29–95	67	10.47	20.40	1.30
T0115	136–168	33	8.69	13.72	2.60
T0115	181–186	6	1.98	10.69	4.02
T0115	194–222	29	9.44	26.52	1.70
T0116	1–18	18	6.31	45.22	0.13
T0116	43–59	17	7.46	12.99	2.54

TABLE VIII. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0116	72–82	11	3.54	5.48	0.78
T0116	104–117	14	4.86	9.14	4.70
T0116	124–130	7	3.29	4.33	2.77
T0116	136–151	16	3.73	10.57	1.14
T0116	158–164	7	1.63	2.37	0.64
T0116	168–174	7	2.61	8.28	0.84
T0116	180–252	73	2.49	21.74	0.46
T0117	1–23	23	1.56	4.82	1.99
T0117	36–46	11	1.97	3.44	1.51
T0117	71–78	8	2.04	2.88	1.54
T0117	89–101	13	3.77	5.90	1.94
T0117	135–146	12	1.30	3.17	0.90
T0117	173–176	4	1.88	3.91	1.64
T0117	191–200	10	3.11	7.43	1.60
T121	1–3	3	N/A	N/A	N/A
T0121	67–76	10	3.89	5.94	1.56
T0121	102–112	11	0.61	1.07	2.78
T0121	132–136	5	2.16	4.69	0.62
T0121	188–191	4	1.94	2.64	3.94
T0122	1–2	2	0.76	6.87	4.41
T0122	26–33	8	0.77	1.05	14.68
T0122	77–81	5	1.06	1.66	3.00
T0122	173–180	8	1.42	5.10	4.00
T0122	241–248	8	2.63	4.48	1.03
T0125	1–10	10	1.85	3.04	4.76
T0125	32–43	12	1.12	1.67	0.15
T0125	67–83	17	3.08	5.09	6.51
T0125	99–117	19	3.75	5.89	7.95
T0125	135–141	7	1.31	8.13	1.00
T0127	1–23	23	2.56	7.25	0.23
T0127	41–47	7	1.60	2.71	1.85
T0127	68–145	78	13.14	14.26	7.23
T0127	153–161	9	0.59	1.39	0.62
T0127	170–185	16	2.84	9.34	4.74
T0128	1–12	12	0.62	2.39	1.95
T0128	66–72	7	1.73	4.39	2.02
T0128	147–151	5	0.52	1.89	0.87
T0128	212–222	11	4.01	11.02	0.86

[a]Not adjusted for missing density in experimental PDB files. Superposition and RMSD calculations use only atoms for which density is reported in the experimental PDB file.
[b]Root-mean-square deviation of residues in the SVR following optimal superposition of the SVR residues.
[c]Root-mean-square deviation of residues in the SVR following optimal superposition of the three stem residues N- and C-terminally adjacent to the SVR.
[d]Root-mean-square deviation of the three stem residues N- and C-terminally adjacent to the SVR following optimal superposition of these stem residues.

The fourth SVR in T0186, residues 256–294, is a small subdomain inserted into the TIM barrel. Although the relative orientation of this SVR was not predicted correctly (RMSD-G = 17Å), the local structure of four-stranded meander is correctly predicted with an RMSD-L of 5.2Å (Fig. 6). If the distortions at the SVR termini are disregarded, the local RMSD significantly improves: a sequence-dependent iterative superposition with a 4 Å cutoff using the LGA algorithm[46] yields an optimal fragment match of 30 residues with an RMSD-L of 2.4 Å. Notably, the prediction of this SVR by the Rosetta-based method was significantly better than any other submitted prediction.

DISCUSSION

The Rosetta-based method for SVR modeling represents a new approach to combining database and de novo strategies for modeling protein segments, both short loops and longer SVRs. The assembly of conformations from smaller fragments allows the benefits of database methods and de novo loop modeling methods to be combined. Iterative optimization of the backbone and side-chain conformations, using a rotamer approximation for side-chains, which to our knowledge has not been previously applied to loop modeling, allows detailed atomic interactions to be evaluated, while significantly restricting the complexity of the conformational search. Allowable confor-

TABLE IX. All SVR Predictions in First-Ranked Models of CASP 5 Targets

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0130	1–13	13	2.98	6.55	0.22
T0130	21–30	10	1.44	2.84	0.90
T0130	51–78	28	2.81	4.34	0.65
T0130	81–114	34	5.02	9.34	0.37
T0132	1–18	18	7.33	15.51	0.56
T0132	52–57	6	3.45	5.11	2.58
T0132	101–112	12	2.76	7.21	0.51
T0132	122–130	9	2.34	2.77	1.59
T0132	133–154	22	1.50	3.63	0.38
T0133	1–29	29	2.04	20.19	0.13
T0133	59–66	8	2.43	3.45	1.05
T0133	98–109	12	1.71	3.42	1.42
T0133	117–126	10	3.03	6.33	1.22
T0133	145–179	35	6.18	16.17	3.02
T0133	197–215	19	4.82	8.89	3.49
T0133	228–251	24	0.87	1.18	0.41
T0133	270–279	10	1.15	1.83	1.25
T0133	287–312	26	9.20	18.34	1.57
T0134	878–882	5	0.87	11.92	0.56
T0134	899–905	7	2.00	2.86	1.10
T0134	928–943	11	3.44	5.17	0.98
T0134	966–976	12	1.15	3.40	0.54
T0134	982–993	3	0.41	1.29	1.74
T0134	1003–1005	13	5.01	7.97	4.66
T0134	1020–1032	16	3.18	6.30	2.11
T0134	1038–1053	5	2.46	6.78	4.29
T0134	1060–1064	7	4.08	8.52	5.39
T0134	1070–1076	6	2.00	3.15	2.76
T0134	1082–1087	7	0.46	0.90	0.23
T0137	41–49	9	0.80	1.36	0.44
T0137	97–102	6	2.21	4.20	0.61
T0137	108–112	5	0.26	2.03	0.41
T0137	119–123	5	0.30	1.06	0.25
T0138	1–4	4	1.94	5.28	0.52
T0138	46–53	8	2.25	2.94	1.11
T0138	58–63	6	1.69	4.13	1.37
T0138	84–89	6	2.55	9.55	3.86
T0138	96–103	8	3.26	7.12	4.09
T0138	106–116	11	2.20	3.21	1.65
T0138	132–135	4	1.24	7.76	1.42
T0141	1–30	30	8.06	22.10	0.31
T0141	55–75	21	4.25	9.05	3.37
T0141	86–111	26	6.42	18.99	2.01
T0141	118–128	11	2.86	3.87	0.69
T0141	144–150	7	2.75	4.07	2.12
T0141	154–171	18	5.27	10.23	2.62
T0141	175–187	13	4.00	13.20	0.14
T0142	1–8	8	1.60	2.35	0.16
T0142	45–72	28	3.41	4.59	0.47
T0142	91–103	13	3.10	3.87	2.05
T0142	106–114	9	2.73	3.19	1.70
T0142	136–144	9	2.66	6.58	2.85
T0142	155–164	10	3.13	5.64	2.10
T0142	200–208	9	1.31	1.73	0.49
T0142	234–239	6	0.86	1.46	0.86
T0142	248–257	10	2.76	3.48	0.92
T0142	262–269	8	2.26	2.73	0.97
T0142	279–282	4	0.35	4.27	0.18
T0147	1–3	3	0.80	3.95	0.69
T0147	7–20	14	3.85	5.42	1.90

TABLE IX. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0147	38–52	15	2.66	5.79	2.65
T0147	62–68	7	2.93	6.16	2.56
T0147	72–82	11	2.93	9.76	2.53
T0147	90–94	5	1.12	2.70	2.21
T0147	98–116	19	4.47	10.77	2.71
T0147	121–127	7	1.90	3.37	3.31
T0147	131–141	11	3.67	5.41	2.63
T0147	149–154	6	1.57	3.29	1.63
T0147	158–175	18	2.54	5.29	2.08
T0147	182–186	5	0.64	2.52	1.61
T0147	190–202	13	4.40	8.07	2.12
T0147	210–216	7	2.01	4.39	2.66
T0147	219–245	27	5.01	14.47	1.75
T0149	1–5	5	0.81	7.72	0.21
T0149	19–33	15	3.18	6.33	1.35
T0149	37–43	7	2.95	5.19	3.53
T0149	58–77	20	6.36	20.28	4.95
T0149	84–94	11	3.59	5.16	2.74
T0149	98–124	27	3.04	5.39	0.48
T0149	148–154	7	2.65	4.66	2.21
T0149	174–184	11	3.01	4.56	1.96
T0149	186–193	8	4.11	7.12	4.35
T0149	195–318	124	15.98	33.44	1.76
T0150	–2–6	8	3.03	4.88	0.07
T0150	94–100	7	0.37	2.25	0.38
T0151	1–6	6	2.10	3.04	0.34
T0151	21–28	8	2.02	3.08	0.76
T0151	36–52	17	1.84	2.64	0.75
T0151	84–99	16	0.73	1.46	0.56
T0151	103–164	62	6.35	9.32	0.19
T0153	30–35	6	1.11	1.28	0.85
T0153	52–58	7	0.42	0.77	0.56
T0153	95–103	9	1.10	2.14	0.49
T0153	119–154	36	6.60	24.15	0.26
T0154	1–11	11	3.67	18.61	0.12
T0154	15–30	16	0.59	1.89	0.36
T0154	54–62	9	1.83	2.36	0.32
T0154	110–117	8	2.57	3.43	0.77
T0154	241–248	8	2.26	3.52	1.82
T0154	254–266	13	3.93	9.87	0.44
T0154	286–309	24	2.13	9.09	0.37
T0155	84–91	8	0.43	1.13	0.37
T0155	119–133	15	0.69	2.82	0.67
T0157	1–2	2	0.47	4.60	0.51
T0157	21–26	6	0.46	2.37	1.73
T0157	35–42	8	1.71	6.74	1.21
T0157	59–71	13	2.95	4.97	1.94
T0157	95–121	27	4.46	5.61	1.70
T0157	133–138	6	0.45	1.06	0.21
T0159	1–8	8	2.30	11.74	0.31
T0159	12–18	7	2.01	2.54	2.55
T0159	31–40	10	3.37	4.58	2.96
T0159	54–59	6	1.43	3.31	1.72
T0159	63–73	11	3.43	4.02	1.40
T0159	77–88	12	3.15	5.46	2.11
T0159	103–111	9	3.53	6.84	1.58
T0159	114–124	11	1.55	7.51	2.50
T0159	146–153	8	2.74	5.61	3.26
T0159	186–193	8	2.37	6.04	1.86
T0159	211–229	19	5.79	15.98	2.55
T0159	265–282	18	3.95	14.68	4.28
T0159	291–296	6	2.11	3.97	1.73
T0159	298–309	12	1.39	11.66	1.46

TABLE IX. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0160	−3–7	10	3.84	8.91	0.22
T0160	76–84	9	3.27	4.91	0.65
T0160	94–112	19	2.30	6.52	0.92
T0165	1–54	54	18.78	41.91	0.29
T0165	64–68	5	1.02	1.35	0.80
T0165	75–82	8	2.17	4.88	0.80
T0165	87–91	5	1.05	1.27	1.29
T0165	95–101	7	2.74	3.53	3.20
T0165	105–111	7	0.40	1.24	1.98
T0165	120–150	31	7.45	12.82	1.23
T0165	167–171	5	1.36	3.08	1.27
T0165	189–199	11	1.79	3.63	0.37
T0165	206–212	7	1.84	2.98	0.79
T0165	224–240	17	3.90	8.42	2.17
T0165	252–260	9	1.01	2.70	0.76
T0165	284–289	6	1.94	2.55	0.43
T0165	298–304	7	0.78	1.15	0.46
T0167	1–4	4	2.05	3.38	0.12
T0167	111–123	13	3.01	3.11	0.28
T0167	127–146	20	5.66	6.82	0.58
T0167	183–185	3	1.21	4.77	0.61
T0168	56–59	4	0.50	1.40	0.51
T0168	63–69	7	2.83	3.84	1.08
T0168	91–129	39	11.76	22.63	5.98
T0168	150–164	15	5.36	10.73	6.24
T0168	196–209	14	5.40	10.41	9.40
T0168	222–238	17	2.14	4.72	2.49
T0168	249–263	15	3.54	6.47	1.06
T0168	271–275	5	1.94	3.46	2.20
T0168	279–293	15	4.23	12.15	1.78
T0168	298–311	14	3.51	7.29	0.44
T0168	323–327	5	1.65	11.86	0.28
T0169	5–10	6	0.43	1.49	0.92
T0169	23–28	6	1.41	7.53	2.90
T0169	36–42	7	1.17	1.74	1.36
T0169	62–67	6	2.65	3.71	0.56
T0169	110–115	6	2.23	4.60	0.75
T0169	124–138	15	5.53	12.46	0.85
T0172	1–7	7	2.43	13.43	1.71
T0172	19–27	9	2.85	4.62	3.06
T0172	45–49	5	0.49	0.99	0.50
T0172	56–75	20	1.97	3.08	0.42
T0172	80–85	6	0.85	1.43	2.00
T0172	107–218	112	13.48	17.59	3.33
T0172	245–249	5	2.06	3.55	2.00
T0172	264–282	19	6.25	9.20	4.53
T0172	293–299	7	0.53	4.96	1.17
T0182	1–5	5	0.77	2.16	0.12
T0182	47–52	6	0.31	0.52	0.51
T0182	249–250	2	0.55	1.86	0.73
T0183	1–26	26	4.79	43.44	0.60
T0183	40–47	8	0.49	0.78	0.31
T0183	56–63	8	0.45	1.15	0.36
T0183	78–84	7	1.36	1.70	0.57
T0183	96–112	17	0.68	2.39	0.99
T0183	142–148	7	0.73	1.18	0.25
T0183	155–166	12	0.47	1.44	0.43
T0183	183–189	7	1.98	2.54	0.38
T0183	197–206	10	0.53	0.74	0.59
T0183	221–229	9	0.51	0.64	0.40
T0183	235–248	14	1.21	15.27	0.34
T0184	1–9	9	0.48	1.74	0.20
T0184	35–51	17	4.40	8.16	0.92
T0184	70–79	10	1.11	1.87	0.74

TABLE IX. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0184	108–122	15	4.25	7.77	0.39
T0184	139–146	8	1.13	1.57	0.81
T0184	162–172	11	4.21	6.92	4.26
T0184	178–185	8	1.66	2.23	0.75
T0184	193–199	7	0.95	2.40	0.71
T0184	207–212	6	1.55	4.07	0.37
T0184	235–240	6	0.83	1.87	0.14
T0185	1–3	3	0.33	1.83	0.22
T0185	15–27	13	2.65	3.75	2.46
T0185	51–60	10	1.73	3.01	0.39
T0185	88–103	16	2.04	4.93	2.67
T0185	127–133	7	2.39	3.09	2.00
T0185	160–172	13	2.91	3.65	0.56
T0185	176–190	15	4.61	7.68	0.98
T0185	217–221	5	1.78	2.48	0.60
T0185	236–243	8	3.08	8.28	2.77
T0185	248–263	16	2.07	4.13	1.57
T0185	306–312	7	0.62	2.23	0.32
T0185	317–322	6	0.82	1.94	0.38
T0185	329–341	13	0.83	1.77	0.93
T0185	346–369	24	2.39	3.47	2.15
T0185	374–407	34	4.92	11.48	1.49
T0185	416–421	6	2.16	4.15	0.78
T0185	426–433	8	2.75	3.40	1.44
T0185	443–457	15	0.26	0.99	0.14
T0186	10–14	5	1.29	2.61	0.50
T0186	27–34	8	2.86	6.19	4.00
T0186	52–63	12	4.39	9.40	0.79
T0186	83–110	28	2.38	7.12	0.78
T0186	116–132	17	2.63	11.08	2.28
T0186	150–159	10	2.63	5.49	2.64
T0186	177–181	5	2.52	3.93	1.97
T0186	188–193	6	0.85	2.36	1.51
T0186	197–211	15	4.42	7.75	1.37
T0186	230–238	9	3.46	8.29	3.26
T0186	244–250	7	2.12	3.33	1.35
T0186	256–294	39	5.20	17.27	1.71
T0186	301–309	9	1.12	1.73	1.44
T0186	327–331	5	0.94	1.35	0.35
T0186	344–351	8	1.02	5.11	2.06
T0186	354–359	6	1.59	2.46	1.96
T0188	6–16	11	2.87	6.06	1.34
T0188	46–56	11	2.88	4.87	1.11
T0189	1–3	3	0.30	0.65	0.25
T0189	16–33	18	4.67	12.38	1.21
T0189	63–68	6	1.00	1.63	0.56
T0189	74–81	8	1.36	2.94	1.13
T0189	88–95	8	2.85	3.80	3.26
T0189	101–107	7	2.63	5.57	4.70
T0189	112–124	13	0.79	1.78	1.90
T0189	141–151	11	1.35	3.52	0.75
T0189	174–185	12	3.15	3.97	2.00
T0189	193–202	10	3.72	6.98	1.65
T0189	219–223	5	0.59	1.19	0.72
T0189	229–235	7	1.73	3.00	2.08
T0189	241–250	10	2.33	4.83	2.06
T0189	273–279	7	1.57	2.07	0.54
T0189	299–307	9	2.03	5.71	1.92
T0189	317–319	3	1.33	11.06	1.42
T0190	1–5	5	0.66	1.34	0.21
T0190	29–34	6	0.95	1.83	0.39

TABLE IX. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0190	51–62	12	2.90	3.59	1.30
T0190	90–96	7	2.68	3.55	0.89
T0191	1–105	105	6.64	10.39	0.21
T0191	143–147	5	1.79	5.28	1.06
T0191	164–175	12	2.70	8.64	3.29
T0191	180–190	11	3.27	6.89	1.32
T0191	196–208	13	4.62	8.97	1.18
T0191	215–219	5	1.99	2.69	0.87
T0191	224–234	11	3.52	6.95	1.94
T0191	254–268	15	3.31	14.61	5.43
T0192	1–3	3	0.76	2.52	0.32
T0192	27–36	10	2.45	4.25	1.42
T0192	41–45	5	0.69	1.21	2.04
T0192	47–51	5	2.04	4.07	1.68
T0192	58–70	13	4.14	13.77	2.04
T0192	78–89	12	0.70	2.14	0.26
T0192	143–153	11	4.04	7.25	1.49
T0192	159–171	13	1.45	21.24	0.18
T0193	1–13	13	3.63	5.54	0.24
T0193	22–28	7	1.20	2.14	2.46
T0193	54–60	7	1.06	6.91	2.85
T0193	64–81	18	4.63	7.00	4.35
T0193	98–105	8	2.73	4.40	0.74
T0193	114–125	12	4.49	6.73	1.83
T0193	132–141	10	2.28	4.26	1.61
T0193	149–166	18	1.04	3.40	0.56
T0193	170–178	9	2.14	6.91	3.80
T0193	189–195	7	3.33	8.54	5.64
T0193	199–211	13	3.09	13.25	1.52
T0195	1–12	12	4.92	11.85	0.24
T0195	35–47	13	3.61	3.82	0.71
T0195	58–73	16	3.07	8.02	2.49
T0195	77–79	3	0.85	4.54	2.97
T0195	91–124	34	6.58	20.45	1.84
T0195	142–154	13	4.60	6.77	1.25
T0195	173–180	8	0.96	1.43	0.59
T0195	188–215	28	5.90	11.07	4.59
T0195	217–232	16	4.00	6.43	3.43
T0195	242–253	12	2.23	3.10	1.50
T0195	259–266	8	1.87	2.68	1.39
T0195	291–299	9	3.10	5.62	0.14

[a]Not adjusted for missing density in experimental PDB files. Superposition and RMSD calculations use only atoms for which density is reported in the experimental PDB file.
[b]Root-mean-square deviation of residues in the SVR following optimal superposition of the SVR residues.
[c]Root-mean-square deviation of residues in the SVR following optimal superposition of the three stem residues N- and C-terminally adjacent to the SVR.
[d]Root-mean-square deviation of the three stem residues N- and C-terminally adjacent to the SVR following optimal superposition of these stem residues.

mations for protein segments up to about five or six residues are adequately sampled in known protein structures,[47] and fragment assembly is unlikely to significantly improve the accuracy of predictions for segments below this size. Because accurate backbone conformations can be selected from known structures, however, the benefits of the rotamer approximation for optimizing atomic interactions likely do contribute to the accuracy of the method for such short segments. Conversely, for long SVRs, sampled conformations may not be sufficiently accurate that optimization of detailed atomic interactions can improve the predictions, but fragment assembly is likely to be critical for effective sampling of backbone conformations.

For short loops, the mean prediction accuracies obtained by the Rosetta method are comparable with those obtained by other loop modeling approaches. Among the best results reported are those of Fiser et al.[17] who obtain RMSD-G values of 0.79, 1.89, and 4.24 Å for 4, 8, and 12 residue loops, respectively. Other recent successful methods have reported mean RMSD-G values of 0.85 and 1.45 for five and eight residue loops[24] and 1.00 and 3.09 Å for four- and eight-residue loops.[13] The mean prediction accuracies

obtained here, 0.59, 1.45, and 3.62 Å for 4, 8, and 12 residue loops, are at least comparable with these methods. Given that real loop modeling does not happen in environments of perfect accuracy, it is unclear what significance, if any, the differences in performance of various methods in the segment reconstruction test have for actual loop modeling. Although the mean prediction accuracies of the best methods are reasonably comparable, the most accurate method for any particular loop region varies, as illustrated in Figure 1. In this small sample set, the de novo prediction method of Fiser et al.[17] and the consensus hybrid approach of Deane and Blundell[13] are the most likely to yield the best prediction, whereas the database method of Van Vlijmen and Karplus[11] yields the best prediction in two cases. The Rosetta method gives good predictions on average but does not result in the top ranked prediction in any of these examples. The fact that the Rosetta-based method does not use native side-chain conformation information in segment reconstructions may contribute in part to this ranking.

Although a variety of methods can predict short loop conformations with reasonable accuracy, reliable prediction of the conformation of long SVRs is an unsolved problem. Because the conformational space accessible to a polypeptide chain increases exponentially with increasing chain length, the difficulty of the structure prediction problem increases dramatically as chain length increases and, consequently, the accuracy with which protein segments are predicted decreases. A hypothesis guiding this work is that the fragment buildup strategy used in the Rosetta method could combine the strengths of database methods with conformational search methods. By assembling shorter fragments to generate conformations for longer regions, the conformational database can be extrapolated, allowing longer protein segments to be modeled with greater accuracy. The predictions obtained for 13- to 35-residue segments, although insufficient to give statistically significant estimates of mean accuracies, illustrate that the method is indeed extendable to long SVRs. In 5 of the 10 cases examined, predictions >2.5 Å RMSD-G were obtained for segments ranging from 13 to 34 residues. In addition, examples from CASP 5 comparative modeling targets, although anecdotal, are quite promising. In several cases where long SVRs were modeled in the context of reasonably accurate alignments, regions modeled as SVRs have accuracies comparable with regions modeled by alignment to a homologue of known structure (Fig. 6).

Given these promising results, how can additional improvements in the method be obtained? For longer segments, conformational sampling becomes a limiting factor in the accuracy of predictions. The native conformation is frequently significantly lower in energy than the lowest-energy conformation sampled (Fig. 5), indicating that significant improvement in the accuracy of long segment predictions could be obtained by additional sampling. For short segments, the potential is not sufficiently accurate to identify the native conformation in general (Table I). Although improvements in the potential clearly would be required to improve the accuracy of the short segment predictions, a bigger practical limitation on the accuracy of short segments is the alignment and environment accuracy. Perhaps the most fruitful target for improvements to the method is in the selection of optimal predictions from the population of sampled conformations. The current discrimination scheme relies solely on ranking conformations according to the potential used for optimization. Clustering has been previously shown to improve discrimination in both de novo structure prediction[48] and loop modeling[13,24] by identifying conformations corresponding to wide energy basins. Addition of clustering to the discrimination scheme is likely to yield an improvement in the current method as well.

CONCLUSION

Comparative modeling provides 3D models for proteins based on sequence similarity to a protein of known structure, and improving the accuracy and completeness of such models requires methods capable of modeling structural divergences between homologous proteins. Because the differences between related structures are responsible for differences in functional specificity, the ability to accurately model SVRs in homologous sequences is required to fully exploit comparative models for functional insight. Although both optimization and database search methods are able to provide accurate models for short loop regions in proteins, accurate structural modeling of longer SVRs in proteins is an unsolved problem. Providing accurate models of longer insertions and template perturbations, however, is perhaps the most biologically relevant application of comparative modeling because such structural changes add novel functions and specificities to protein scaffolds. Here we use the fragment buildup strategy of the de novo prediction algorithm Rosetta in an attempt to overcome some of the sampling limitations that restrict the accuracy of modeling methods by extrapolating the structure database to cover longer protein segments. The resulting method performs as well as existing loop modeling methods on short loops, and initial results for longer segments illustrate the promise of the method for predicting structures of long SVRs as well.

ACKNOWLEDGMENTS

CAR was supported by the Interdisciplinary Training in Genomic Sciences program. DC is a fellow of the Program in Mathematics and Molecular Biology at the Florida State University, with funding from the Burroughs Wellcome Fund Interfaces Program.

NOTE ADDED IN PROOF

The Rosetta potential and methods for local sampling and rapid fragment screening used in this study are described in detail in a forthcoming volume of Methods in Enzymology (Rohl CA, Strauss CEM, Misura KMS, Baker D. Meth Enzym 2004;383:66–93).

REFERENCES

1. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.

2. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–566.
3. Simons KT, Ruczinski I, Kooperberg C, Fox B, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;35:82–95.
4. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
5. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4 Proteins 2001;Suppl 5:22–38.
6. Sippl MJ, Lackner P, Dominguez FS, Prlíc A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001;Suppl 5:55–67.
7. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J 1986;5:819–822.
8. Martin ACR, Thornton JM. Structural families in loops of homologous proteins: automatic classification, modeling and application to antibodies. J Mol Biol 1996;263:800–815.
9. Oliva B, Bates PA, Querol E, Avilés FX, Sternberg MJE. An automated classification of the structure of protein loops. J Mol Biol 1997;266:814–830.
10. Rufino SD, Donate LE, Canard LHJ, Blundell TL. Predicting the conformation class of short and medium size loops connecting regular secondary structures: application to comparative modeling. J Mol Biol 1997;267:352–367.
11. Van Vlijmen WWT, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;257:975–1001.
12. Wojcik J, Mornon J-P, Chomilier J. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J Mol Biol 1999;289:1469–1490.
13. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10:599–612.
14. Burke DF, Deane CM. Improved protein loop prediction from sequence alone. Protein Eng 2001;7:473–478.
15. Bruccoleri RE, Karplus M. Conformational sampling using high-temperature molecular dynamics. Biopolymers1990;29:1847–1862.
16. Hornak V, Simmerling C. Generation of accurate protein loop conformations through low-barrier molecular dynamics. Proteins 2003;51:577–590.
17. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2001;9:1753–1773.
18. Rapp CS, Friesner RA. Prediction of loop geometries using a generalized Born model of solvation effects. Proteins 1999;35:173–183.
19. Moult J, James MN. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1986;1:146–163.
20. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments in proteins by systematic search. Biopolymers 1987;26:137–168.
21. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 2000;40:135–144.
22. Galaktionov S, Nikiforovich GV, Marshall GR. Ab initio modeling of small medium and large loops in proteins. Biopolymers 2001;60:153–168.
23. DePristo MA, de Bakker, PIW, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 2003;51:41–55.
24. Shenkin PS, Yarmush DL, Fine RM, Wang HJ, Levinthal C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 1987;26:2053–2085.
25. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
26. Go N, Scheraga HA. Ring closure and local conformation deformations of chain molecules. Macromolecules 1970;3:178–187.
27. Wedemeyer W, Scheraga HA. Exact analytical loop closure in proteins using polynomial equations. J Comp Chem 1999;20:819–844.
28. de Bakker PIW, DePristo MA, Burke DF, Blundell TL. Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential ant the AMBER force field with the Generalized Born solvation model. Proteins 2003;51:21–40.
29. Mas MT, Smith KC, Yarmush DL, Aisaka K, Fine RM. Modeling the anti-cea antibody combining site by homology and conformational search. Proteins 1992;14:483–498.
30. Martin AC, Cheetham JC, Rees AR. Modeling antibody hypervariable loops: combined algorithm. Proc Natl Acad Sci USA. 1989;86:9268–9272.
31. Sudarsanam S, DuBose RF, March CJ, Srinivasan S. Modeling protein loops using a Φ I+1, Ψ i dimer database. J Mol Biol1995;206:759–777.
32. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in CASP4: Progress in ab initio protein structure prediction. Proteins 2001;Suppl 5:119–126.
33. Bradley P, Chivian D, Meiler J, Misura KMS, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CEM, Baker D. Rosetta predictions in CASP5: successes, failures and prospects for complete automation. Proteins 2003. Forthcoming.
34. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci USA 1987;84:6611–6615.
35. Press WH, Teukolski SA, Vetterling WT, Flannery BP. Numerical recipes in Fortran 77: the art of scientific computing, 2nd ed. Cambridge: Cambridge University Press; 2001.
36. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
37. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97:10383–10388.
38. Bowers PM, Strauss CEM, Baker D. De novo protein structure determination using sparse NMR data. J Biomol NMR 2000;18:311–318.
39. Kabsch W, Sander C. Dictionary of protein secondary: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
40. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326:1239–1259.
41. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
42. 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.
43. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–602.
44. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CEM, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the ROBETTA server. Proteins 2003. Forthcoming.
45. http://PredictionCenter.llnl.gov/CASP4.
46. Zemla A. LGA program: a method for finding 3-D similarities in protein structures. 2000; accessed at http://PredictionCenterllnl-gov/local/lga
47. Fidelis K, Stern PS, Bacon D, Moult J. Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng 1994;7:953–960.
48. Shortle D, Simons KT, Baker D. Clustering of low energy conformations near the native structures of small proteins. Proc Natl Acad Sci 1998;95:11158–11162.
49. Koradi R, Billeter M, Wüthrich K. MOMOL: a program for display and analysis of macromolecular structures. J Mol Graphics 1996;14:51–55.
50. Kraulis P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24:946–950.