Ab Initio Construction of Polypeptide Fragments: Efficient Generation of Accurate, Representative Ensembles

Mark A. DePristo,* Paul I. W. de Bakker, Simon C. Lovell, and Tom L. Blundell
Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

ABSTRACT **We describe a novel method to generate ensembles of conformations of the main-chain atoms {N, Cα, C, O, Cβ} for a sequence of amino acids within the context of a fixed protein framework. Each conformation satisfies fundamental stereochemical restraints such as idealized geometry, favorable ϕ/ψ angles, and excluded volume. The ensembles include conformations both near and far from the native structure. Algorithms for effective conformational sampling and constant time overlap detection permit the generation of thousands of distinct conformations in minutes. Unlike previous approaches, our method samples dihedral angles from fine-grained ϕ/ψ state sets, which we demonstrate is superior to exhaustive enumeration from coarse ϕ/ψ sets. Applied to a large set of loop structures, our method samples consistently near-native conformations, averaging 0.4, 1.1, and 2.2 Å main-chain root-mean-square deviations for four, eight, and twelve residue long loops, respectively. The ensembles make ideal decoy sets to assess the discriminatory power of a selection method. Using these decoy sets, we conclude that quality of anchor geometry cannot reliably identify near-native conformations, though the selection results are comparable to previous loop prediction methods. In a subsequent study (de Bakker et al.: Proteins 2003;51: 21–40), we demonstrate that the AMBER forcefield with the Generalized Born solvation model identifies near-native conformations significantly better than previous methods. Proteins 2003;51:41–55.**

Key words: conformational sampling; conformational search algorithms; anchor geometry; decoy sets; discrete state sets; loop modeling

INTRODUCTION

Many methods in protein structure determination, analysis, and prediction utilize procedures that generate ensembles of conformations for a sequence of amino acids. Determination of protein structures with X-ray crystallography and nuclear magnetic resonance is greatly aided by automatic methods to generate conformations consistent with experimentally derived restraints.[1–3] Analysis of the energetics and conformational preferences of protein structures often begins with an evaluation of alternate conformations for a given polypeptide.[4] Computational mutagenesis and recent approaches to protein design attempt to account explicitly for protein backbone flexibility and, thus, require efficient methods to generate backbone conformations consistent with the surrounding environment.[5] Ensembles of near-native and non-native conformations are used to assess the discriminatory power of selection mechanisms such as statistical potentials and molecular mechanics forcefields.[6–8] Finally, conformation generation coupled with an effective selection method is a popular approach to protein structure prediction, for both comparative modeling and ab initio methods.[9,10]

To be useful in such a variety of contexts, conformational ensembles should: (1) explicitly represent all heavy main-chain atoms; (2) contain only conformations satisfying stereochemical rules such as reasonable bond lengths, angles, and torsions,[11] excluded volume, and exhibit ϕ/ψ combinations within the allowable regions of the Ramachandran plot[12]; (3) contain a representative sample of the conformational space accessible within the surrounding environment, including conformations near the native structure; (4) be efficiently and reliably generated.

Previous approaches to conformation generation divide roughly into knowledge-based and ab initio methods. Inspired by the observation that models of whole proteins can be composed solely of fragments from previously solved structures,[2] knowledge-based methods extract ensembles from a database of protein structures based on sequence compatibility with the target amino acid sequence and structural compatibility with the framework protein.[13–16] Although in most cases knowledge-based methods can generate conformations very close to the native structure, they frequently fail to produce ensembles with near-native conformations or sufficient conformational diversity, due to limited coverage of conformations in the structure database at longer lengths.[17] Despite its rapidly increasing size, structural databases are never expected to contain representative ensembles for polypeptides longer than eight residues.[17]

Grant sponsor: Marshall Aid Commemoration Commission; Grant sponsor: Cambridge European Trust; Grant sponsor: Isaac Newton Trust; Grant sponsor: NUFFIC Talentenprogramma; Grant sponsor: BBSRC.

*Correspondence to: Mark DePristo, Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK. E-mail: mdepristo@cryst.bioc.cam.ac.uk

Received 28 March 2002; Accepted 7 August 2002

Ab initio approaches to conformational generation encompass a wide range of methods. Discrete conformational search methods explicitly represent the main-chain and/or side-chain atoms with fixed bond lengths and angles, use a simplified energy function, and generate ensembles by exhaustive enumeration or heuristic sampling from discrete sets of ϕ/ψ angles.[18–20] An energy function, such as a molecular mechanics force-field[21,22] or statistical potential,[4,23] is often used to directly sample conformations or restrict conformations to low-energy states. Many methods have been applied to sample directly from the energy function, including molecular dynamics,[24] simulated annealing,[25] Monte Carlo,[26] and local minimization.[27] These energetic methods are intimately dependent on the quality of the energy function and sampling method, often missing or excluding plausible conformations due to poor sampling or incorrect energy calculations.

Loops in Protein Structures

Loop regions in protein structures are the most demanding testing ground for ensemble generation methods not intended for the prediction of whole proteins. Loops occur in a variety of lengths from only a few to as many as 30 residues, though the majority are less than 12 residues long.[25] They exhibit greater structural variability than regular secondary structure elements,[28] and their conformational properties are often likened to those of a random coil.[29] There is only a weak relationship between sequence and conformation for short polypeptide chains in general,[30,31] although characteristic sequence/structure patterns are observed for some classes of loops.[32,33] Their short length, structural variability, and weak sequence/structure relationship make loop regions in protein structures ideal benchmark targets for conformation generation methods.

Decoy Sets

The predictive power of a selection mechanism, such as a molecular mechanics force-field[21,22] or statistical potential,[4,23] is commonly accessed by its ability to differentiate between native-like and non-native "decoy" conformations for the same sequence.[6,8,34] A good decoy set must include a large number of conformations, some near-native and others decoys that are native-like in all respects except the overall folded conformation.[6,8] Further, the set should be generated independently from the evaluated scoring mechanisms, to avoid bias toward any particular selection methodology.

A variety of methods has been used to derive decoy sets. Rojnuckarin and Subramanian[35] used experimentally determined protein structures at differing resolutions to evaluate an all-atom statistical potential. Following the work of Novotny et al;[36] many authors thread an amino acid sequence onto the backbones of proteins of roughly equivalent size but different fold.[34] Decoys have been generated ab initio by enumeration on lattice models,[37] from molecular dynamics trajectories,[38] Monte Carlo sampling,[39] and discrete conformational search.[6,23] Recently, these decoy sets have been collected into a web-accessible database[8] that has seen extensive use.[40]

Efficient Generation of Conformational Ensembles

Our approach to generate conformational ensembles, implemented in a program called RAPPER, is composed of three independent parts: (1) a discrete conformation search algorithm; (2) pre-filter rules such as idealized geometry and propensity-weighted, residue-specific ϕ/ψ states that restrict the conformations examined by the search algorithm; (3) post-filter restraints such as gap closure and excluded volume that eliminate invalid conformations. In this study, we (1) describe the methods and algorithms used to generate conformational ensembles; (2) discuss the ability of these methods to produce ensembles with near-native conformations; (3) compare our fine-grained ϕ/ψ state sets with previously published sets; (4) derive decoy sets for a large set of protein loop structures; (5) evaluate selection using anchor deviation relative to previous efforts to model loops in protein structures. In a related article (de Bakker et al.),[41] we demonstrate that the AMBER forcefield with the Generalized Born solvation model identifies near-native conformations substantially better than previous methods.

METHODS

Top500 Non-Redundant Protein Structure Dataset

The Top500 database of non-redundant protein structures[†] is a hand-curated set of 500 protein structures from the Top240,[42] PDBselect30,[43] and recent high-resolution structures. All database structures were solved by X-ray crystallography to 1.8 Å or better resolution, have few van der Waals clashes[44] and deviations from ideal bond lengths and angles,[11] excluding structures with free-atom refinements, unrefined or absent B-factors, and atomic occupancy less than 1.0. Finally, no two structures share greater than 60% sequence identity. The Top500 and predecessor databases have been used to assess global goodness-of-fit using explicit hydrogens[44] and compile a high-quality library of side-chain rotamers.[42]

Representation of Protein Structure

High-quality protein-like conformations can be generated by varying only the soft dihedral angles ϕ/ψ while fixing bond lengths and angles. In this study, we generate polypeptides with idealized {N, Cα, C, O, Cβ} geometry and the generally invariant ω dihedral fixed to the *trans* state.[‡] Atoms are represented as hard-spheres with van der Waals radii taken from Word et al.,[44] reduced by a factor of 20% to ensure that only energetically infeasible conformations are rejected by the hard-spheres excluded-volume restraint.

Fine-Grained Residue-Specific ϕ/ψ State Sets

Instead of allowing continuous variance in the ϕ/ψ angles, it is common to restrict ϕ/ψ values to a set of

[†]http://kinemage.biochem.duke.edu/databases/top500.php.

[‡]This is a potential problem when the native conformation is in the *cis* state. However, since the frequency of *cis* conformations is ≈ 0.05% for non-proline residues and ≈ 5.00% for proline, excluding *cis* conformations is a minor limitation. Computed from the Top500 database.

discrete states, with as few as four to as many as 55 states.[18,20,45,46] In light of the strong correlation between the size of a state set and its ability to accurately represent native protein structures,[46] we use residue-specific, propensity-weighted state sets with as many as 72^2 states per residue. The state sets are derived from the Top500 database of protein structures, after eliminating residues where any main-chain atom has B-factor > 30.0 Å,^{2} van der Waals overlap > 0.4 Å,[44] or grossly incorrect geometry. An observation histogram is computed for each of the 20 amino acids with 5° resolution in the coil class as defined by the PROCHECK variant of DSSP,[47] convolved with a Gaussian mask with standard deviation 5°, and contoured to include 99.99% of the propensity (see Table I and Figs. 1–3).

Coarse-Grained ϕ/ψ State Sets

Four coarse-grained state sets (see Table I) are examined in this study, two intended for loop modeling and two for whole protein modeling. Deane and Blundell[20] developed an ab initio loop prediction program utilizing an eight-state ϕ/ψ set generated by optimally fitting eight points to a composite ϕ/ψ distribution derived from loop residues in the protein databank.[48] Moult and James[18] derived an 11-state set for loop modeling by selecting ϕ/ψ combinations such that all observed ϕ/ψ angles from protein structures lie within 20° of at least one state. Rooman et al.[45] use a seven-state ϕ/ψ set to model whole protein structures selected by averaging the observed ϕ/ψ combinations in seven (ϕ, ψ, ω) domains, two helical, two extended β, two positive ϕ, and one *cis*. Since we do not generate *cis* conformations, we instead transform the original (−82, 133, 0) *cis* state into a new (−82, 133, 180) *trans* state. Park and Levitt[46] optimized 250 four-state sets against the ability to model accurately whole protein structures,[46] yielding an optimal set that we use here (set "C").

Sampling With a Round-Robin Scheduling Algorithm

We sample conformations from the N to C termini using a discrete search algorithm inspired by fair-share scheduling algorithms that distribute resources evenly among competing processes. The search for a valid conformation begins by creating a fixed length queue of conformations Q, and enqueuing the N-terminal residue. At each step, the leading conformation C is popped off Q and extended by a single residue R. A ϕ/ψ combination is randomly chosen from the ϕ/ψ table for R and positioned into its correct three-dimensional position relative to C. If the extended $C + R$ conformation satisfies all of the RAPPER restraints, then both $C + R$ and C are enqueued into Q and the algorithm iterates with the next conformation at the front of Q. If $C + R$ is invalid, then another unused ϕ/ψ is sampled, until 25 attempts have been made, at which time C is discarded. If $C + R$ bridges the gap from $A \rightarrow B$ and is the correct length, then $C + R$ is a complete conformation satisfying all of the RAPPER restraints. We iteratively execute this search in RAPPER, collecting conformations until 1,000 have been found where no two conformations have RMSD < 0.2 Å.

TABLE I. Summary of the ϕ/ψ State Sets Used in This Study[†]

	ϕ/ψ state sets				
			N non-zero states		
Name	N states	ϕ/ψ states	Gly	Pro	other
RAPPER9	81		79	29	63
RAPPER18	324		290	94	203
RAPPER36	1,296		1,094	334	717
RAPPER72	5,184		4,264	1,207	2,666
Park	4	(−63, −63) (−132, 115) (−42, −41) (−44, 127)	4	4	4
Rooman	7	(−117, 142) (−69, 140) (−89, −1) (78, 20) (−65, −40) (103, −176) (−82, 133)	7	7	7
Deane	8	(−63, −40) (−125, 135) (−70, −47) (−78, 149) (−95, −5) (55, 40) (85, 5) (−119, 120)	8	8	8
Moult	11	(−160, 160) (−120, 150) (−120, 110) (−100, 10) (−90, −30) (−80, 130) (−80, 170) (−80, 70) (−80, 10) (−70, −30) (60, 40)	11	11	11

[†]N states is the number of states in each set, though for the RAPPER sets this number is the number of states, not the number of states with non-zero propensity. The number of non-zero states is the number of states with non-zero propensity and hence accessible to RAPPER sampling, for Gly, Pro, and other (Non-Gly, Non-Pro) residues, respectively. Since the four coarse-grained state sets are not residue-type specific, they have equal numbers of non-zero states for Gly, Pro, and other. Finally, where feasible, the ϕ/ψ states for a set are explicitly listed.

Hard-Spheres Excluded Volume in Constant Time

At the heart of any algorithm, examining van der Waals interactions is a routine that determines whether an atom b overlaps with some atom a in a set of atoms. In many applications, all or most of these atoms are fixed in space; i.e., they do not change their position throughout the lifetime of the program. These atoms are part of a fixed framework set A. The baseline performance of detecting hard-spheres overlaps for a single atom b requires linear time in $|A|$, since b can simply be checked against each $a \in$

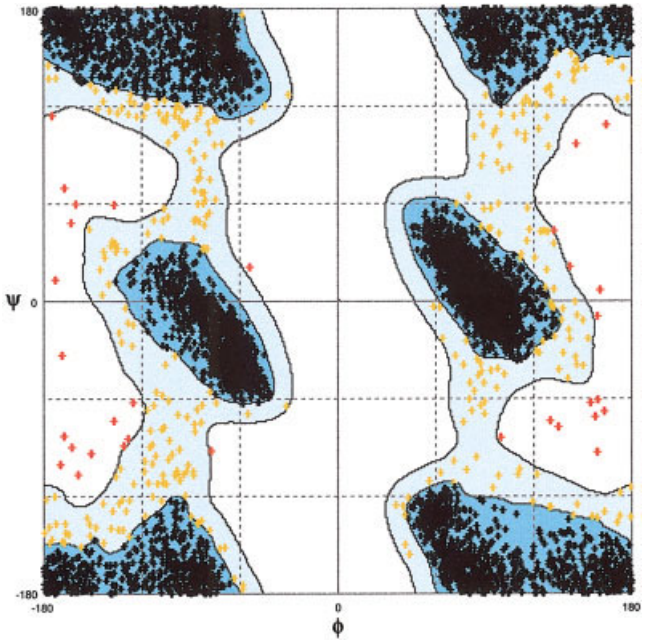

Fig. 1. RAPPER72 glycine ϕ/ψ map. The 7,098 observed Gly ϕ/ψ combinations from the Top500 database are marked with pluses, black for core, orange for disfavored, and red for forbidden data points. The contours are derived from the observed data points using Gaussian smoothing (see Methods), where the darker core contour includes 98.0% of observed data, while the lighter disfavored contour includes the remaining 1.99%. The forbidden points were excluded from the contour calculation after examining the structure and density map where available. Created by RAMPAGE (http://www-cryst.bioc.cam.ac.uk/rampage/).

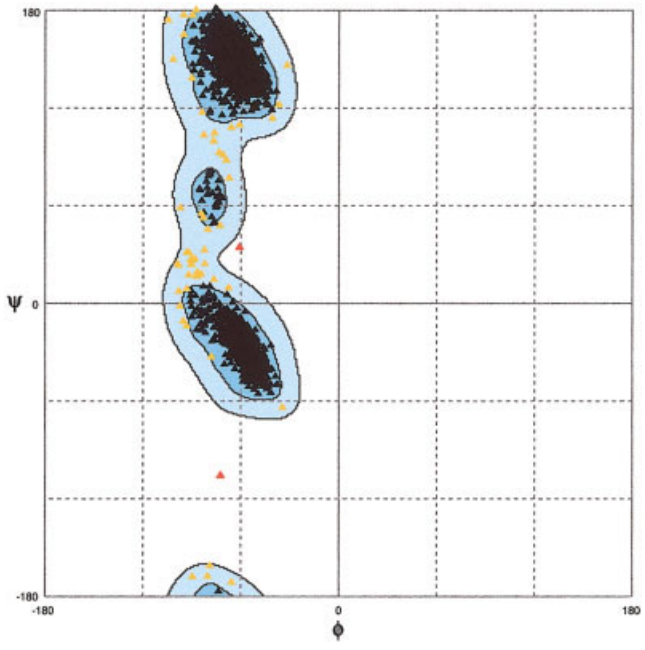

Fig. 2. RAPPER72 proline ϕ/ψ map. The 1,923 observed Pro ϕ/ψ combinations from the Top500 database are marked with triangles, black for core, orange for disfavored, and red for forbidden data points. The contours are derived from the observed data points using Gaussian smoothing (see Methods), where the darker core contour includes 98.0% of observed data, while the lighter disfavored contour includes the remaining 1.99%. The forbidden points were excluded from the contour calculation after examining the structure and density map where available. Created by RAMPAGE (http://www-cryst.bioc.cam.ac.uk/rampage/).

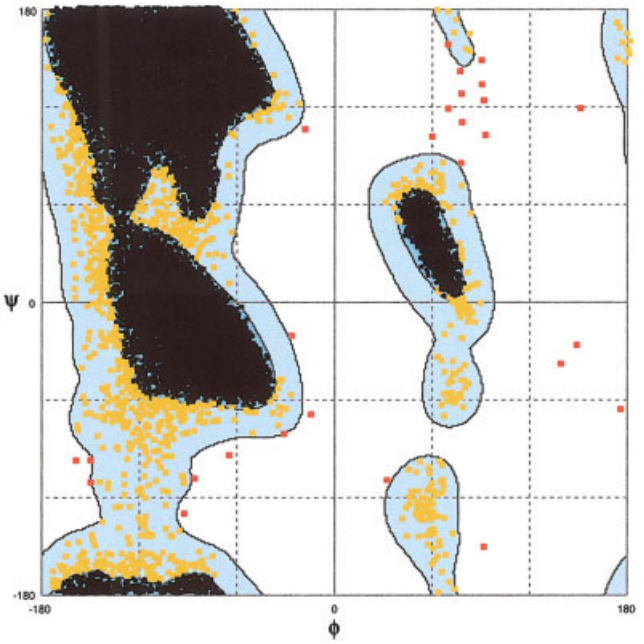

Fig. 3. Composite RAPPER72 ϕ/ψ map for all non-Pro/non-Gly residues. The 65,340 observed non-Pro/non-Gly ϕ/ψ combinations from the Top500 database are marked with squares, black for core, orange for disfavored, and red for forbidden data points. The contours are derived from the observed data points using Gaussian smoothing (see Methods), where the darker core contour includes 98.0% of observed data, while the lighter disfavored contour includes the remaining 1.99%. The forbidden points were excluded from the contour calculation after examining the structure and density map where available. Note that in RAPPER72, each of the 20 amino acids has an individual ϕ/ψ map; this composite map is intended for illustrative purposes only. Created by RAMPAGE (http://www-cryst.bioc.cam.ac.uk/rampage/).

A. This naive algorithm is prohibitively expensive for proteins, which typically contain several thousand atoms.

Since the majority of atoms in our target proteins are fixed during conformational sampling, we approach the overlap detection problem by building a discrete three-dimensional grid around the protein, where each grid vertex contains all possible overlapping atoms lying near the vertex. The algorithm is similar in spirit to methods to compute and update non-bonded neighbor lists, and has been partially developed in previous work on loop modeling.[18] Testing for hard-sphere overlap of *b* against *A* requires only a comparison with the list of atoms in the grid vertex nearest to *b*. The number of atoms in this list is determined by the resolution of the grid and distribution of framework atoms *A* in space, and is independent of the size of *A*. This discrete grid method provides an complete, exact, and efficient constant time solution to the problem of hard-sphere overlap detection against a fixed set of atoms. Without this algorithmic improvement, our conformational sampling method would be orders of magnitude more computationally expensive.

Gap Closure Restraint and C-Terminal Anchor Dihedral Angle Minimization

To ensure that conformations rejoin with the framework protein at the C-terminal anchor, a sequence-separation

dependent gap-closure restraint is applied during conformational sampling. The restraint requires the Cα atom of residue i in an n residue long segment be within a sphere of radius 3.8(n − i + 1) + 1 Å around the Cα atom of the anchor residue $C\alpha_{anchor}$.

When building the C-terminal dummy residue (at position n + 1), the 1 Å margin allows some flexibility in the position of the dummy residue relative to the C-terminal anchor residue. This lenient margin is necessary because of the improbability of sampling a conformation with perfect geometry at the C-terminal anchor residue.

The remaining, potentially large deviations in anchor geometry must nevertheless be eliminated, preferably while maintaining idealized geometry within the conformation. Following previous approaches,[27,49] we employ a dihedral angle minimizer to improve C-terminal anchor geometry, as measured by a harmonic potential between the N and Cα atoms of the dummy and native anchor residues (Equation 1).

$$e_{\text{term}} = \sum_{i \in \{N,C\alpha\}} \|\vec{d}_i - \vec{n}_i\|^2 \quad (1)$$

where $\vec{d}_i$, $\vec{n}_i$ are the coordinates of atom i in the dummy and native anchor residue, respectively. The minimizer iteratively perturbs a randomly chosen ϕ or ψ angle, the conformation by a random angle from a uniform distribution between −5.0 and 5.0, accepting the move if the new ϕ/ψ state has non-zero propensity, the excluded-volume restraint is satisfied, and e_{term} decreases. The minimization continues until 10,000 steps have been made or the average decrease in e_{term} over the last 50 accepted moves is less than −0.01.

Fiser Benchmark Set

We used the test set for loop targets of two residues to 12 residues proposed by Fiser et al.[25] The following structures (PDB codes given) were found to be obsolete and were replaced with the current structure (in parentheses): 2cyr (3cyr), 4fxn (2fox), 3b5c (1cyo), and 4ptp (5ptp). Non-standard amino acids in protein structures were stripped. Although we attempted to fix the structure when there are missing side chain atoms (reason 2, see below) with SCWRL, sometimes the side chain is close to the loop target or there are simply too many missing atoms to unambiguously assign atoms. In many of these cases, the incompleteness of a structure presents a problem when performing energy minimizations in TINKER. Since these shortcomings cannot be fixed easily, we decided to discard such structures altogether. Two structures (2sns and 1ivd) have such poor ϕ/ψ and bond length/angle characteristics that these, in our opinion, cannot justifiably be included in a loop target test set. Other structures were excluded when we found forbidden or many poor backbone dihedral angles in the target loop. The following targets were removed and reasons in parentheses are explained below: 2-mers: 1bam-73 (1,2,3), 1bgc-56 (2,3), lede-148 (1), 1pda-41 (2,3), 1rsy-211 (2), 3cla-108 (2); 3-mers: 1bam-72 (1,2,3), 1bgc-55 (2,3), 1mpp-264 (2), 1pda-40 (2,3), 1rsy-211 (2), 3cla-107 (2); 4-mers: 1bam-92 (2,3), 1bgc-40 (2,3), 1pda-139 (2,3), 2cyp-127 (2), 3cla-27 (1,2); 5-mers: 1bam-132 (2,3), 1ivd-449 (4), 2cyp-139 (2), 2sns-36 (4), 3cla-49 (2); 6-mers: 1mrk-241 (1, right next to C-terminus), 1pda-149 (2,3), 2cyp-159 (2), 3cla-194 (2); 7-mers: 1clc-82 (2), 2sns-134 (4); 8-mers: 1clc-313 (2), 1gof-606 (trouble with TINKER), 1hbq-31 (1,2), 1ivd-413 (4), 1lst-101 (1,2), 1mpp-74 (2), 2sns-17 (4), 3cox-109 (2); 9-mers: 1ivd-244 (4), 1pda-108 (2,3), 2cyp-145 (2); 10-mers: 1ezm-237 (2), 2sns-25 (4), 3cla-96 (2); 11-mers: 1bam-149 (2,3), 1ivd-280 (4), 1mpp-195 (2), 1pda-85 (2,3), 1rsy-230 (2), 1thg-322 (trouble with TINKER), 8acn-323 (1); 12-mers: 1ivd-365 (4), 2cyp-191 (2), 2sns-111 (4), 3cla-176 (2), 3cox-478 (2), Here are the reasons for discarding targets: 1 = poor ϕ/ψ angles or B-factors in loop residues (according to WHAT-CHECK [50] and RAMPAGE; Paul de Bakker and Simon Lovell, unpublished work); 2 = missing side chain atoms; 3 = missing regions (gaps); 4 = overall poor quality.

RESULTS AND DISCUSSION

Anchor Geometry

As shown in Table II, the initial conformations, as expected from the lenient gap-closure restraint, have large anchor deviations (0.82 Å anchor RMSD). After anchor deviation minimization, however, conformations have significantly improved anchor geometry (0.25 Å anchor RMSD). Interestingly, while the initial anchor RMSD increases with loop length, the minimized anchor RMSD remains roughly constant. The initially higher anchor RMSD and greater decrease between initial and final anchor RMSD at longer lengths is expected because of the larger number of degrees of freedom of the polypeptide chain. As shown by the difference between the initial and final global RMSD to native, minimizing the anchor RMSD also moves conformations toward the native structure, an unexpected benefit given that the anchor RMSD and main-chain RMSD are computed over disjoint sets of atoms. This result implies that anchor geometry restricts the overall loop conformation, and consequently minimizing the deviation in anchor geometry forces conformations toward the native structure.

Given the sensitivity of most selection mechanisms to deviations from standard bond lengths and angles, near-perfect anchor geometry is essential for conformational ensembles intended for further analysis with selection mechanisms. Following the suggestion of Zhang et al.,[27] our harmonic potential (Equation 1) includes only {N, Cα} atoms. Although this leads to reasonable bonds lengths at the anchor join, it does not guarantee correct bond angles or good ϕ/ψ combinations with the penultimate residue. Preliminary results with an {N, Cα, C, O} potential are encouraging and appear to improve significantly overall anchor geometry, especially the {$C\alpha_{i-1}$, C_{i-1}, N} and {C_{i-1},N,Cα} bond angles.

Near-Native Conformations Within the Ensembles

The ability to sample near-native conformations in a consistent manner is the most important property of any conformation generation method. Table III summarizes the quality of RAPPER ensembles for two to twelve

TABLE II. Performance of the Dihedral Angle Minimizer at Reducing the C-Terminal Anchor Deviation, Averaged Over All Targets Within a Loop Length for the Benchmark Set[†]

	Dihedral angle minimization					
	C Anchor ⟨RMSD⟩ [Å]			⟨RMSD⟩ to Native [Å]		
Length	Initial	Minimized	Δ	Initial	Minimized	Δ
2	0.61	0.29	−0.32	1.07	0.95	−0.12
3	0.70	0.24	−0.46	1.50	1.30	−0.20
4	0.73	0.20	−0.53	1.85	1.64	−0.21
5	0.76	0.22	−0.54	2.52	2.27	−0.25
6	0.73	0.26	−0.47	3.10	2.91	−0.19
7	0.80	0.25	−0.55	3.98	3.79	−0.19
8	0.84	0.21	−0.63	4.43	4.16	−0.27
9	0.91	0.24	−0.67	5.39	5.00	−0.39
10	0.89	0.28	−0.61	5.79	5.51	−0.28
11	0.92	0.23	−0.69	7.08	6.71	−0.37
12	0.94	0.37	−0.57	7.23	6.96	−0.27
Mean	0.82	0.25	−0.57	—	—	−0.26

[†]The Initial and Minimized columns are average values for conformers before and after minimization, respectively. The C-terminal anchor RMSD is measured over the {N, Cα} atoms of the conformer dummy residue and C-terminal anchor residue of the native protein. The RMSD to native is the global RMSD over all heavy main-chain atoms between the loop residues of the conformer and native protein. The Δ columns are the difference between the preceding minimized and initial columns, explicitly showing the change in anchor RMSD (column 4) and global RMSD (column 7) following minimization. The mean row is the average value for each column over all loop lengths, excluding the RMSD to native columns where the average is not meaningful.

TABLE III. Several Measures of Quality of the RAPPER Ensembles for the Conformation Nearest to the Native Structure and the Entire Generated Ensemble of 1,000 Conformations, for Each Loop Length in the Fiser Loop Target Set

	RAPPER conformational sampling							
			Best generated ⟨RMSD⟩ [Å]				Ensemble ⟨RMSD⟩ [Å] main chain	
			Main chain		Superimposed Cα			
Length	Targets	Runtime ⟨min⟩	Global	Local	Average	% < 1.0	Global	Local
2	34	0.4	0.31	0.11	0.01	100.0	0.95	0.48
3	34	8.9	0.34	0.18	0.02	100.0	1.30	0.80
4	35	57.9	0.43	0.24	0.07	100.0	1.65	1.08
5	35	55.6	0.53	0.32	0.15	100.0	2.27	1.38
6	36	47.9	0.69	0.42	0.27	100.0	3.06	1.71
7	38	75.5	0.78	0.49	0.38	100.0	3.79	1.93
8	32	145.6	1.11	0.70	0.56	93.8	4.16	2.13
9	37	121.8	1.29	0.81	0.72	83.8	5.00	2.50
10	37	176.6	1.67	1.11	1.00	45.9	5.66	2.85
11	33	236.3	1.99	1.27	1.23	36.4	6.71	3.21
12	34	401.8	2.21	1.47	1.46	5.9	6.96	3.35

[†]Targets, total number of targets for each length. Runtime, Average time taken to sample and minimize 1,000 distinct conformations by RAPPER on a 900-MHz AMD Athlon processor. Global and Local, Computed over all main-chain heavy atoms {N, Cα, C, O} without superposition. Average and % < 1.0, Computed over Cα atoms only, after least-squares superposition of the Cα coordinates. % < 1.0, Percentage of targets within the loop length where the superimposed Cα RMSD is less than 1.0 Å for at least one conformation in the ensemble.

residue long loops. The average best-generated conformation, measured by global RMSD from native, indicates that our method samples near-native conformations for almost all targets, averaging 0.43, 1.11, and 2.21 Å for four, eight, and twelve residue long loops, respectively. Since the global RMSD is computed without structural superposition, the best conformations are both internally similar to and placed in the correct spatial orientation as the native structure. The low local RMSD, 0.24, 0.70, and 1.47 Å for four, eight, and twelve residue loops, respectively, shows that the ensembles include conformations with significant internal similarity to the native loop, although misplaced relative to the native conformation.

As shown in Figure 4, our method samples consistently low-RMSD conformations across all targets within a given length, with low variability and few poorly sampled targets, although the variability increases at longer lengths. The worst performing targets, those with high lowest RMSD conformations, are not substantially worse than the average for the loop length, never rising above 2.0 Å for loops up to eight residues long. Further, for no target does the method fail to produce at least some conformations, although they may be somewhat far from native. In applications using conformational ensembles, the average and worst-case performance of the generation method are probably more important than the best-case targets, since

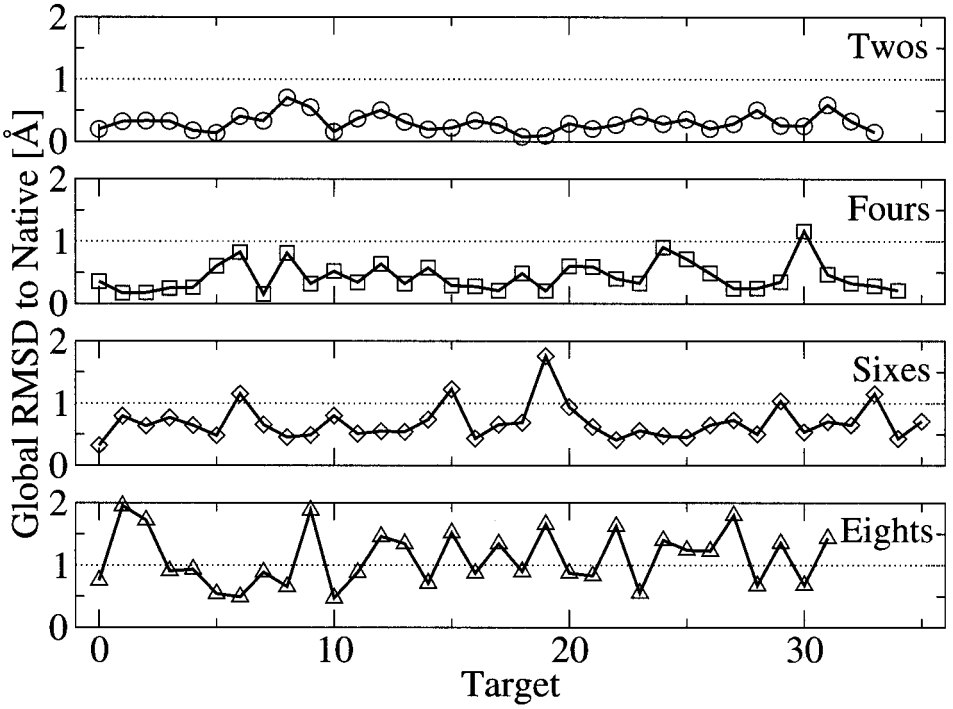

Fig. 4. Plots of the best-generated conformation (lowest main-chain RMSD to the native structure) for each target structure for the two (circle), four (square), six (diamond), and eight (triangle) residue long loops in the benchmark set. The target index (x-axis) is arbitrary, while the global RMSD (y-axis) is ranges between 0–2 Å for each length. Because the best RMSDs are not normally distributed, this representation is necessary to show the variance with the lowest RMSD across each set of loop targets.

these provide expectations for average and upper-bound quality of the sampling for any given structure.

In this study, we use a stringent measure of structural similarity in the non-superimposed RMSD over all main-chain heavy atoms because we believe this similarity measure is sensitive to the subtle differences in conformation needed for further energy function discrimination. A common and yet more forgiving measure of similarity is the RMSD over Cα atoms after superposition. Following Fidelis et al.,[17] who consider two structures with Cα RMSD < 1.0 Å equivalent for selection, we note that RAPPER ensembles contain such a conformation for all targets less than eight residues and the majority of nine residue long targets, though only a minority at ten residues and longer.

In conclusion, our method generates consistently near-native conformations for polypeptides up to 12 residues long. According to the Fidelis et al.[17] standard, RAPPER solves the conformation generation problem for loops with fewer than ten residues.

Comparison of Coarse-Grained and Fine-Grained φ/ψ States Sets

We now compare the effectiveness of several common discrete φ/ψ state sets representing the native conformation with respect to the standard RAPPER sets (Table I). The heuristic nature of the fair-share scheduling algorithm precludes definite statements about the lower-bound accuracy of a state set. Fortunately, the limited size of the coarse-grained sets permits an exhaustive analysis of all allowable φ/ψ combinations. The results for the coarse-grained sets (Table IV and Fig. 5) have been obtained by exhaustive enumeration of all φ/ψ combinations, under idealized geometry, gap-closure, and excluded-volume restraints. Due to its computational cost, the dihedral angle minimizer cannot be used on the large number of conformations examined during exhaustive enumeration. To directly compare the results of the coarse and fine-grained state sets, we consider here the initial, non-minimized conformations produced by the discrete search algorithm with the fine-grained state sets.

The most striking feature of the results for the coarse-grained sets is the number of targets where no conformation exists satisfying the excluded-volume and gap-closure restraints. The number of failures is inversely related to number of available φ/ψ states. The larger number of failures for short loops is due to the inability of the available φ/ψ states to avoid excluded-volume overlap with the framework protein and simultaneously close the gap between anchor residues in the more constrained context of short loops. With more degrees of freedom and, thus, more accessible conformational space away from the framework protein and relatively reduced anchor separations, the number of failures at longer lengths is lesser, though still substantial for all of the state sets except Moult. The generally high failure rate for these state sets is readily explained by the strictness of the restraints utilized in our method, relative to the programs where these sets origi-

TABLE IV. Quality of Conformational Generation With Several ϕ/ψ State Sets Using RAPPER and With a Knowledge-Based Method, as Measured by the Average Global RMSD to the Native Structure of the Best Conformation Generated†

	Conformation generation comparison [RMSD, Å]								
	Coarse-grained sets					Fine-grained sets			
Length	Park	Rooman	Deane	Moult	FREAD	9	18	36	72
2	1.47 (16)	0.71 (9)	0.85 (5)	0.61 (4)	—	0.60	0.38	0.32	0.30
3	1.74 (24)	1.05 (14)	1.14 (9)	0.87 (5)	0.43 (3)	0.85	0.43	0.33	0.32
4	2.27 (26)	1.41 (7)	1.64 (5)	1.25 (1)	0.67 (3)	0.82	0.51	0.47	0.46
5	2.37 (22)	1.63 (9)	1.65 (5)	1.28 (3)	1.11 (6)	0.89	0.56	0.53	0.52
6	3.25 (14)	2.33 (11)	2.23 (4)	1.66 (3)	1.22 (5)	1.01	0.74	0.70	0.70
7	3.58 (21)	2.63 (11)	2.24 (6)	1.87 (2)	1.17 (4)	1.00	0.84	0.83	0.82
8	4.26 (12)	2.59 (4)	2.44 (2)	1.84 (1)	1.51 (2)	1.30	1.20	1.12	1.20
9	3.99 (10)	2.61 (2)	2.06 (2)	1.65 (0)	—	1.57	1.43	1.47	1.44
10	3.61 (10)	2.28 (4)	2.02 (4)	1.53 (1)	—	1.70	1.63	1.65	1.59
11	3.60 (8)	2.30 (4)	2.20 (3)	1.58 (1)	—	2.15	2.18	2.14	2.14
12	3.35 (8)	2.34 (3)	2.25 (0)	1.56 (0)	—	2.19	2.07	2.08	2.06

†The number of failures, in parentheses, is the number of targets where the method could find no conformations satisfying the RAPPER (or FREAD) restraints, but have been omitted for the fine-grained state sets, which never failed. The best conformation for the coarse-grained state sets was computed by exhaustive enumeration. The best conformation for the fine-grained state sets (RAPPER9, RAPPER18, RAPPER36, and RAPPER72) was taken from a sample of 1,000 conformations. The FREAD data (see Methods) are restricted by the FREAD program to targets between three and eight residues in lengths.

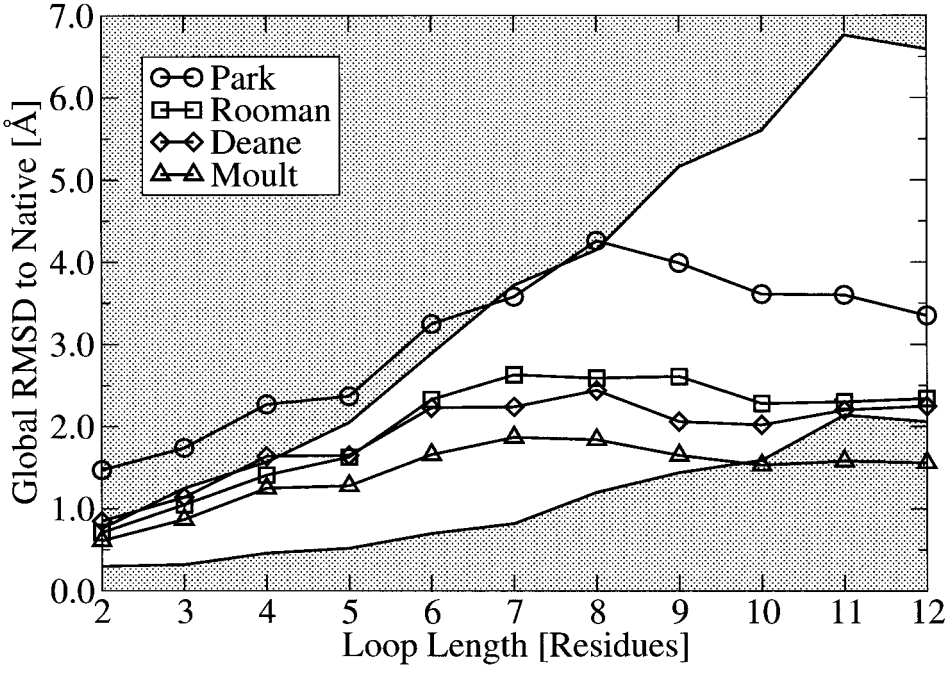

Fig. 5. For each coarse-grained state set the average main-chain RMSD of the lowest possible conformation to native (y-axis) is plotted for each loop length from three to twelve (x-axis). The shaded regions are derived from the best and average performance of 1000 sampled conformations using the RAPPER72 table without minimization. The lower shaded region is the average RMSD of the conformation with lowest RMSD to the native structure, and the upper shaded region is the ensemble average RMSD.

nated. Nevertheless, with the generous 1.0 Å gap-closure margin and without side-chain atoms considered in the hard-spheres overlap, we view the RAPPER restraints as already quite lenient.

In terms of the average RMSD to native of successful targets, we see an inverse correlation between the size of a set and its ability to accurately reproduce the native conformation. With respect to the results of sampling from the RAPPER72 set, we see that the Park set in fact cannot produce conformations better than the average conformation sampled with the RAPPER72 set for loops from three to five residues in length, and remains barely below the ensemble average up to eight residues. The average RMSDs of both the Deane and Rooman sets almost completely coincide and are substantially worse than the lowest sampled RAPPER72 conformation, though well below the RAPPER72 ensemble average for up to ten residue long loops. The Moult set, with only a few more states than

Deane and Rooman, performs significantly better than both these sets, remaining within 1 Å of the lowest RAPPER72 conformation for lengths below ten residues. Even excluding failed targets, the high average RMSDs of the coarse-grained sets clearly indicate that these sets cannot reliably reproduce the internal structure of the target loop and its orientation in the framework protein for loops with fewer than ten residues.

For loops of ten or more residues, the coarse-grained state sets begin to compare more favorably with their fine-grained alternates. The Park set again appears too limited and continues to perform substantially worse than the fine-grained state sets. The Rooman and Deane sets perform better than the RAPPER9 and only marginally worse than the RAPPER18, RAPPER36, and RAPPER72 state sets. The Moult set out-performs all of the fine-grained state sets, stabilizing at 1.5–1.6 Å RMSD from native for ten, eleven, and twelve residue loops. The improving performance of the coarse-grained state sets at longer lengths relative to the fine-grained state sets implies that the penalties for small state set size, dominant at short lengths, become increasingly insignificant as coverage of conformational sampling worsens for the fine-grained sets at longer lengths.

The relative improvement of the coarse-grained state sets, and even the superior performance of the Moult set, does not indicate, however, that the coarse-grained states are equivalent or preferable at longer lengths. It means only that the Rooman and Deane set are capable of producing conformations as close, and Moult closer, to the native than the best conformation in 1,000 sampled conformations from the fine-grained state sets. Finding these minimal RMSD to native conformations required prior knowledge of the native conformation and a flagrant disregard for computational cost in a direct enumeration of hundreds of billions of conformations. It in no way reflects the likelihood that these low RMSD conformations will be found with a realistic sampling method without knowledge of the native conformation.

With their high failure rate and poor best conformations, coarse-grained state sets cannot accurately reproduce native conformations within the context of a fixed framework protein structure for lengths up to ten residues. The fine-grained RAPPER72 set, even with heuristic sampling, can reproduce near-native conformations, outperforming the Park, Rooman, and Deane coarse-grained sets at every length and the Moult set up to ten residues without a single failure.

Performance of Fine-Grained ϕ/ψ States Sets

The conclusion that coarse-grained state sets cannot reliably reproduce near-native conformations leads to the question of how fine-grained a ϕ/ψ state must be to adequately reproduce near-native conformations. Throughout this study, we have used the very fine RAPPER72 state set, with ϕ/ψ states every 5° and over 2,000 ϕ/ψ states with non-zero propensity. We compare the quality of conformation generation using the standard RAPPER72 and the RAPPER9, RAPPER18, and RAPPER36 sets with resolutions of 40°, 20°, 10°, respectively, as measured by the nearest-native conformation of 1,000 conformations produced under the standard RAPPER restraints without dihedral angle minimization (see Table IV).

Most importantly, in no case do any of these state sets fail to produce valid conformations for a target, as occurred frequently with the coarse-grained sets. Thus, although differing in accuracy at reproducing near-native conformations, all of these fine-grained sets can be used with confidence to generate at least some candidate conformations.

In terms of RMSD to native, the general trend is that the higher resolution sets outperform their lower resolution counterparts. Similar to coarse-grained sets, the lower resolution sets perform worse at short lengths, improving relative to the higher resolution sets at longer lengths. The RAPPER9 set is clearly inferior to the other state sets, across all lengths. The differences among the RAPPER18, RAPPER36, and RAPPER72 tables are not substantial, especially for more than six residues, and consequently are equivalent for ab initio conformational sampling under the restraints imposed in this study. State sets at resolutions below 10° were not considered in this study, since no improvement over the RAPPER72 table was anticipated.

The RAPPER72 set with 5° resolution was selected because it most accurately reproduces the native conformation at shorter lengths, though computational costs could be reduced for only a minor accuracy penalty by moving to the RAPPER36 or even RAPPER18 sets.

Sampling Problems

In principle, all conformational sampling methods have a "sampling problem," in that there is no guarantee that the best or even a good conformation will be visited within a reasonable number of sampling steps. The important question is not whether there is a sampling problem—without a doubt there is—but rather how quickly the sampling visits a conformation that is good enough for the application.

In RAPPER, we see an inverse logarithmic relationship between the size of the sampled ensemble and lowest RMSD to native of a conformation within the ensemble (Fig. 6). For both the four and eight residue targets, an ensemble size of 1,000 is well along the tail of the distribution, indicating that larger ensembles would not likely include better conformations. On the other hand, the twelve residue targets might benefit from a larger ensemble of 5,000 or even 10,000 conformations. Certainly for lengths up to ten residues, the accuracy of our conformational sampling method is sufficient for a number of interesting applications in protein structure analysis and prediction.

Nevertheless, we would like to improve the quality of sampling in absolute terms. The rapidly diminishing return on increasing ensemble size, however, indicates that simply expanding the search to include tens of thousands of conformations will not be an effective way to find better conformations. As noted in the comparison of fine-grained state sets, reducing the size of the ϕ/ψ set will

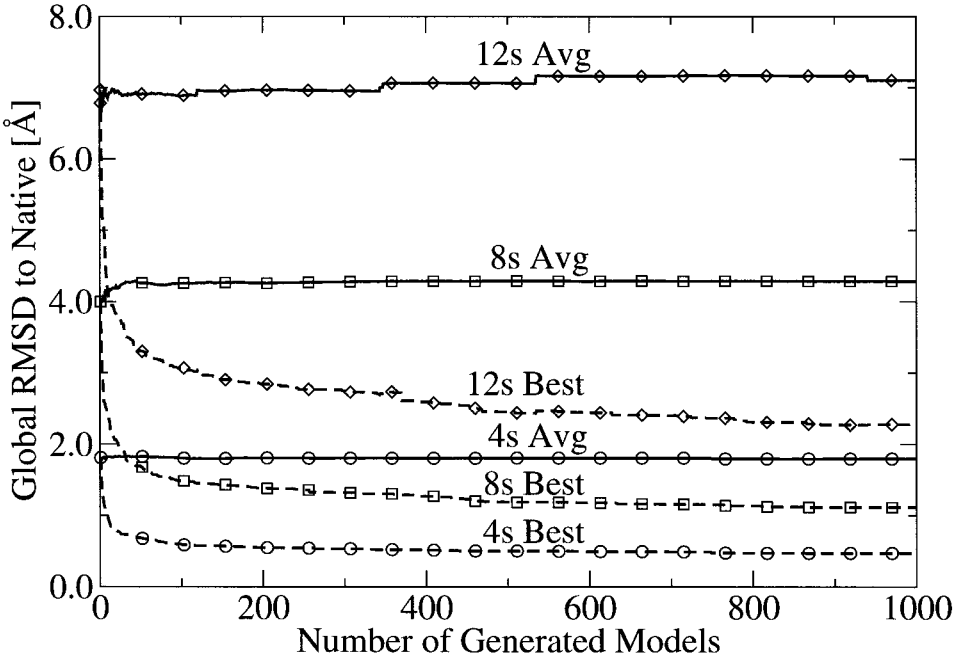

Fig. 6. Global main-chain RMSD to native of the best conformation and ensemble average as a function of the size of the sampled ensemble for four (circle), eight (square), and twelve (diamond) residue long loops. The ensemble average RMSD quickly stabilizes at its final value, while the best RMSD to native follows a inverse logarithmic curve. The four and eight residue loops are well along the tail of this distribution, though the twelve residue loops have yet to converge.

not improve the quality of the best-sampled conformation. One promising option is to add restraints to enforce good packing or avoid unfavorable electrostatic interactions that would restrict the sampling to more native-like conformations.

Comparison With Alternate Conformational Sampling Methods

There are several major competing approaches to conformational generation of short polypeptides, including analytic methods for loop closure,[51,52] knowledge-based fragment matching,[2,20] and molecular dynamics with simulated annealing.[25,53] The major differences among these methods are in (1) accuracy, or ability to sample near-native conformations, (2) computational efficiency of sampling, (3) intrinsic support for external conformational restraints, and (4) general applicability.

Analytic loop closure generates conformations by solving for roots of a high-order polynomial equation describing the geometric restraints on loop atom positions as a function of bond lengths, angles, and the anchor residue atom positions.[52] Although elegant and extremely efficient, analytic loop closure methods suffer from severe restrictions on their applicability and support for restraints. Most significantly, analytic methods can only be applied directly on conformations involving two or fewer residues, though longer lengths can be attempted with supporting sampling routines.[19,52] Further, analytic closure methods cannot directly account for external restraints such as excluded volume.[52] In light of these restrictions, our method is a significant improvement over analytic loop closure methods because of its support for complex conformational restraints such as excluded volume, its high accuracy, computational efficiency, and applicability to conformations of more than two residues.

We can directly compare the conformations generated by FREAD, a recent knowledge-based method, to those obtained by RAPPER. FREAD extracts an ensemble of conformations from a high-quality subset of the experimental-solved protein structures based on sequence and structure compatibility between the target sequence and database protein.[15] Due to original program restrictions and a lack of database coverage, FREAD is limited to targets of between three and eight residues. Conformations were annealed to the target framework structure by optimally superimposing anchor residues, the global RMSD calculated between the candidate conformation and native conformation without further superpositioning, and the best global RMSD conformation averaged over each length (Table IV). Similar to the coarse-grained state sets, FREAD fails to find any conformations with sufficient compatibility to the target structure for approximately 10% of the targets. Despite outperforming the coarse-grained state sets, the best FREAD conformation is worse than the best RAPPER conformation by between 0.2 and 0.4 Å RMSD.

Though knowledge-based methods are an efficient means to extract conformational ensembles that often contain near-native conformations, they suffer from several problems when applied to the general problem of conformational sampling. As is commonly stated, knowledge-based methods are restricted to targets of eight or less eight residues due to limited database coverage.[13,15,17] Further,

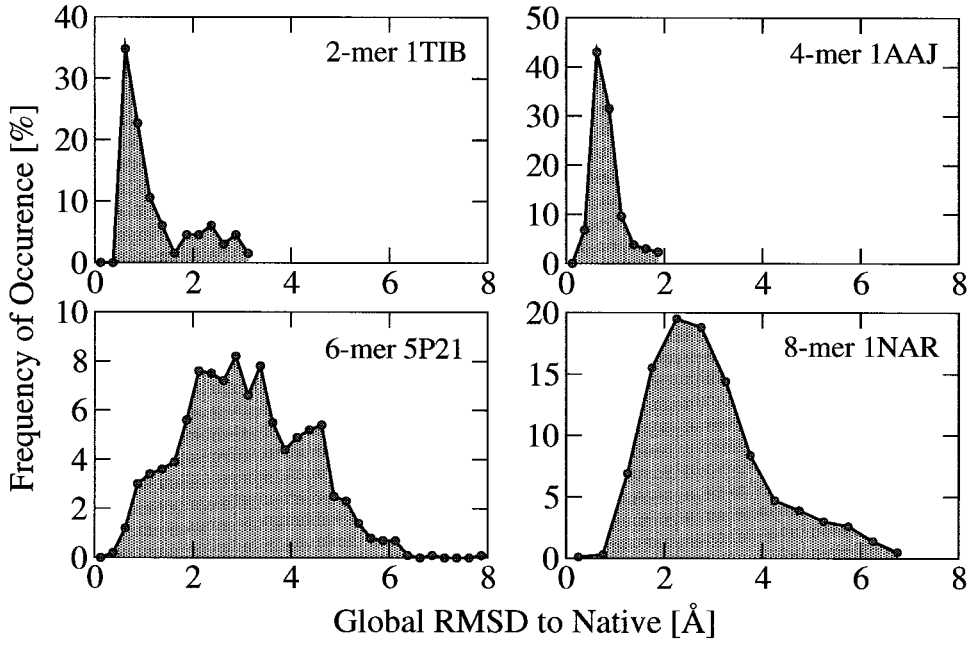

Fig. 7. The distribution of the main-chain RMSDs of all 1,000 conformations within four representative targets for two (1tib), four (1aaj), six (5p21), and eight (1nar) residue long loops. The RMSD to native (x-axis) ranges between 0–8 Å in each graph, while the frequency of conformations at each RMSD to native (y-axis) varies among each graph. The ensemble RMSDs have been collated into bins every 0.25 Å.

there is a significant limit to the size and diversity of the ensembles that can be generated by extracting fragments from the protein database. Since we wish to extract only reasonable conformations, we must restrict ourselves to database conformations with plausible anchor geometry. Unfortunately, there are often few conformations with the minimum quality of anchor geometry that are substantially different from the native structure. Although not a problem for conformation prediction, the coupling of anchor quality and ensemble diversity is problematic in applications requiring both large and diverse ensembles such as for decoy set generation.

Molecular dynamics with simulated annealing (MD/SA) is alternate approach to conformation sampling. Fiser et al.[25] recently applied MD/SA with an all-atom statistical potential to loop modeling with very favorable results, averaging predicted global mainchain RMSDs of 0.8, 1.9, and 3.8 Å for four, eight, and twelve residue loops, respectively. These results are not strictly comparable to those presented here, since MD/SA couples selection of a best conformation with conformation generation, whereas we are solely concerned with quality of conformation generation without simultaneous scoring of these conformations.

The average RMSD of the best conformation sampled by our method is substantially below the average RMSD selected by Fiser et al.[25] (0.4 vs. 0.8, 1.1 vs. 1.9, 2.2 vs. 3.8 Å), though they report that conformations were sampled during the MD/SA calculation with lower RMSD to native than those selected. It is encouraging that with only geometric and excluded-volume restraints, we can reliably and efficiently sample conformations much closer to the native than those selected by the best modeling methods. Coupled with an effective selection method such as a statistical potential or molecular mechanics force-field, our method could in principle predict conformations much closer to native. In a related study (de Bakker et al.[41]), we demonstrate that the AMBER forcefield with the Generalized Born solvation model can do this in practice.

Decoy Set Generation

RAPPER ensembles exhibit many desirable features for use as decoy sets, namely (1) ideal bond lengths and angles; (2) good ϕ/ψ properties; (3) no van der Waals overlaps; and (4) perfect N-terminal and high-quality C-terminal anchor geometry. The minimum RMSD threshold, independent sampling, and large number of conformations ensure diversity within the ensemble, producing smooth distributions of conformations from near to far from native, as shown in Figure 7. Since the conformational restraints used in our method are local, negative filters, the conformations are not biased by any global, positive terms such as electrostatic interactions, solvation free energy, or hydrogen bonding patterns that can vary significantly among selection mechanisms. The large size and diversity of protein targets in the Fiser loop benchmark set[25] ensures that our decoy sets cover a wide range of protein topologies, unlike previous decoy sets.[6,23,34]

Model Selection With Anchor Geometry

Here we evaluate selection using anchor deviation as measured by anchor RMSD, a measure of the goodness-of-fit of a conformation at the C-terminal anchor, commonly used in knowledge-based and some ab initio meth-

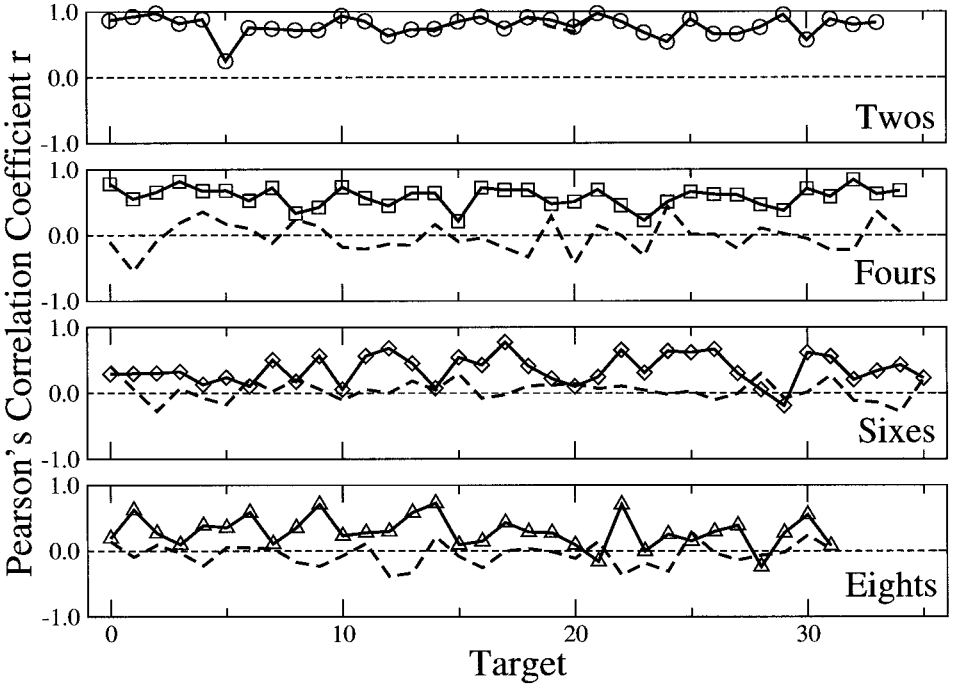

Fig. 8. Plots of the Pearson's correlation coefficient (r) for each target structure in the two (circle), four (square), six (diamond), and eight (triangle) residue long loops in the benchmark set. The correlation coefficient computed between C-terminal anchor RMSD {N, Cα} and global main-chain RMSD. The solid line in each graph is the correlation coefficient over the whole ensemble of 1,000 conformations, while the dashed line is the correlation over a subset of the 50 conformations with the lowest anchor RMSD. Since the two residue long loops do not produce 1,000 distinct conformations, the solid and dashed lines overlap substantially and hence the dashed line is not visible for most targets.

ods.[13,15,17,20,54,55] The anchor RMSD of a conformation is correlated with its RMSD to the native structure (Fig. 8, black line), presumably because native-like conformations can more easily join to the framework anchor regions. The correlation is strongest at short lengths and decreases with increasing length. The correlation remains significant for loops up to eight residues long, but almost disappears by twelves residues (data not shown).

Despite the strong correlation, selecting the fragment with anchor RMSD performs poorly at identifying near-native conformations (Fig. 9). The average mainchain RMSD of the conformations with the lowest anchor RMSD is barely below the ensemble average RMSD and very far from the ideal lower bound of the lowest generated RMSD, indicating that anchor RMSD cannot effectively discriminate among the conformations in the ensemble. The apparent inconsistency between the strong correlation and weak predictive power of scoring with anchor deviation results from the large number of high RMSD conformations with low anchor deviation in the ensemble, confusing the selection. In other words, the correlation between anchor and global RMSD, though significant over all 1,000 conformations, disappears when considering only the 50 conformations with the lowest anchor RMSD (Fig. 8, dashed line). The lack of correlation means that anchor RMSD cannot rank conformations within this subset, and since it includes a large number of non-native conformations, leads to the inability of the anchor RMSD to discriminate effectively between near-native from non-native conformations within the entire ensemble. Nevertheless, the strong correlation within the entire ensemble means that anchor deviation is useful as a filter to eliminate non-native conformations from the ensemble, but should not be used to distinguish near-native from non-native conformations among the remaining low anchor RMSD conformations.

It should be noted that, with a major exception of the recent results reported by Fiser et al.,[25] selection with anchor RMSD performs as well or better than previously reported loop modeling methods.[13,15,16,18,20,55,56] More amazingly, the ensemble average RMSD without selection performs only marginally worse than most of these previous methods (around 0.2–0.4 Å RMSD), demonstrating that naive geometric and steric considerations can match the predictive power of most previously published loop modeling methods.

It must be emphasized that we have only considered experimentally determined protein structures in this study. Though we expect most of the results to hold equally well for comparative models, the prediction of individual loop conformations on inexact models is more complex than on exact models. Nevertheless, previous ab initio and knowledge-based loop modeling methods[25] have also used native loop reconstruction to estimate the quality of loop prediction on comparative models. An analysis of the sensitivity of the described method to errors in the positions of the anchor residues, as well as its application to loops on

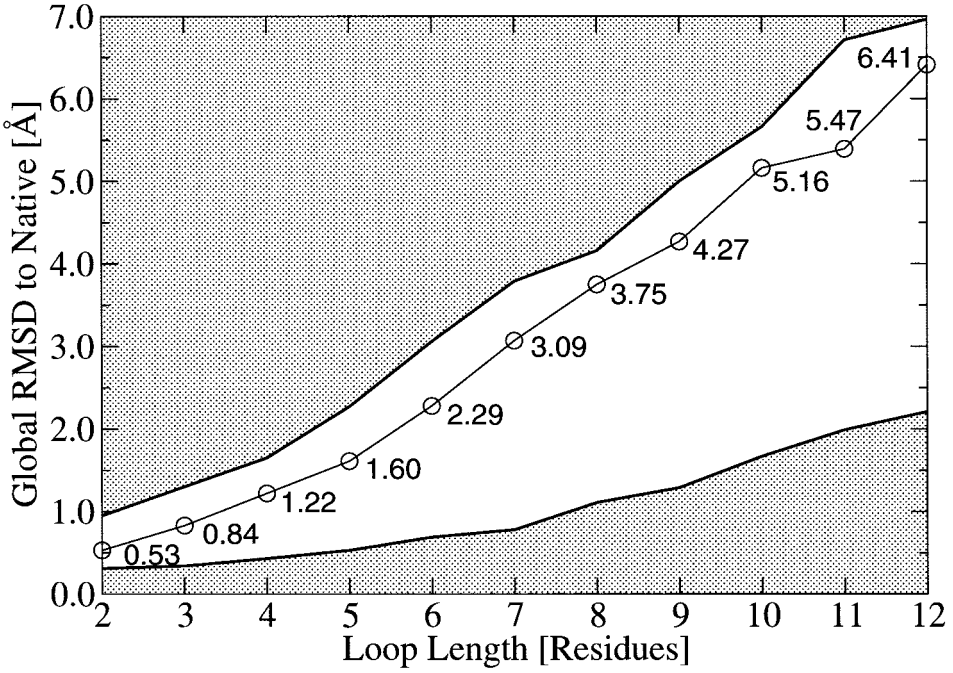

Fig. 9. Average global main-chain RMSD (y-axis) of the conformations with lowest anchor RMSD for each loop length (x-axis) in the benchmark set (circles). The lower shaded region is the average RMSD of the conformation with the lowest RMSD to the native structure. The upper shaded region is the average RMSD of the average RMSD of the entire ensemble with respect to the native structure.

comparative models, is currently underway and will be reported elsewhere.

CONCLUSIONS

In this review, we have discussed a novel method called RAPPER to generate diverse ensembles of polypeptide conformations consistent with a surrounding fixed protein structure. The method samples from fine-grained, residue-specific ϕ/ψ state sets to find conformations that satisfy a number of geometric and knowledge-based filters, including reasonable ϕ/ψ angles, and gap-closure and excluded-volume restraints. Conformational search and minimization are sufficiently fast to generate ensembles containing thousands of structurally distinct conformations in minutes on a workstation-class computer. Using this ab initio conformational generation method on a large benchmark test set of protein loop structures, we have made the following observations:

- The method samples low-RMSD conformations, with average RMSDs of 0.4, 1.1, 2.2 Å for four, eight, and twelve residue loops, respectively, and with average superimposed Cα RMSD < 1.0 Å for every target up to seven residues, and for most targets up to ten residues long.
- Coarse-grained ϕ/ψ state sets cannot reliably reproduce near-native conformations while respecting excluded-volume restraints and, consequently, efficient sampling from fine-grained ϕ/ψ sets is superior to exhaustive enumerations from coarse sets.
- The ensembles make ideal decoy sets because of their large size and diversity, because they are generated independently from any discriminatory function and because each conformation has ideal bond lengths and angles, high-quality C-terminal anchor geometry, and no van der Waals overlaps.
- Despite the strong correlation between anchor RMSD and global RMSD, anchor RMSD is a poor criterion to identify near-native conformations. Nevertheless, the average RMSD of conformations selected by anchor RMSD is comparable to previously published loop modeling results.

Given its generality, robustness, and efficiency, we conclude that our method is a significant improvement over previously published work on conformational sampling. In a related paper (de Bakker et al.[41]), we evaluate the discriminatory power of many additionally selection criteria, including an all-atom statistical potential, the AMBER molecular mechanics forcefield in gas-phase and with the Generalized Born/surface area solvation model, concluding that AMBER/GBSA identifies near-native conformations substantially better than previous methods.

The RAPPER program and all ensembles presented in this work are available at (http://www-cryst.bioc.cam.ac.uk/rapper/).

ACKNOWLEDGMENTS

M.A.D. thanks the Marshall Aid Commemoration Commission for its generous support of his studies in the

United Kingdom. P.D.B. thanks the Cambridge European Trust, Isaac Newton Trust, NUFFIC Talentenprogramma, and BBSRC for financial support. S.C.L. is a Wellcome Trust Fellow of Mathematical Biology.

REFERENCES

1. Holm L, Sander C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. J Mol Biol 1991;218:183–194.
2. Jones T, Thirup S. Using known substructures in protein model building and crystallography. EMBO J 1986;5:819–922.
3. Read RJ, Moult J. Fitting electron density by systematic search. Acta Crystallogr Sect A 1992;48:104–113.
4. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213: 859–883.
5. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282:1462–1467.
6. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258: 367–392.
7. Moult J. Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 1997;7:194–199.
8. Samudrala R, Levitt M. Decoys "R" Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000;9:1399–1401.
9. Sippl MJ, Hendlich M, Lackner P. Assembly of polypeptide and protein back-bone conformations from low energy ensembles of short fragments: development of strategies and construction of models for myoglobin, lysozyme, and thymosin beta 4. Protein Sci 1992;1:625–640.
10. 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.
11. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein-structure refinement. Acta Crystallogr Sect A 1991;47:392–400.
12. Ramachandran G, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;28:283–437.
13. van Vlijmen H, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
14. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
15. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10:599–612.
16. Wojcik J, Mornon JP, 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.
17. 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.
18. Moult J, James MN. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1986;1:146–163.
19. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.
20. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 2000;40:135–144.
21. Cornell W, Cieplak P, Bayly C, Gould I, Merz KJ, Ferguson D, Spellmeyer D, Fox T, Caldwell J, Kollman P. A second generation force field for the simulation of proteins and nucleic acids. J Am Chem Soc 1995;117:5179–5197.
22. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
23. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:893–914.
24. Bruccoleri RE, Karplus M. Conformational sampling using high-temperature molecular-dynamics. Biopolymers 1990;29:1847–1862.
25. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
26. Abagyan R, Totrov M. Biased probability Monte-Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
27. Zhang HY, Lai LH, Wang LY, Han YZ, Tang YQ. A fast and efficient program for modeling protein loops. Biopolymers 1997;41: 61–72.
28. Sibanda BL, Blundell TL, Thornton JM. Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. J Mol Biol 1989;206:759–777.
29. Swindells MB, MacArthur MW, Thornton JM. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol 1995;2:596–603.
30. Kabsch W, Sander C. Identical pentapeptides with different backbones. Nature 1985;317:207.
31. Cohen BI, Presnell SR, Cohen FE. Origins of structural diversity within sequentially identical hexapeptides. Protein Sci 1993;2: 2134–2145.
32. Sibanda BL, Thornton JM. Beta-hairpin families in globular proteins. Nature 1985;316:170–174.
33. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, et al. Conformations of immunoglobulin hypervariable regions. Nature 1989;342:877–883.
34. Holm L, Sander C. Evaluation of protein models by atomic solvation preference. J Mol Biol 1992;225:93–105.
35. Rojnuckarin A, Subramaniam S. Knowledge-based interaction potentials for proteins. Proteins 1999;36:54–67.
36. Novotny J, Bruccoleri R, Karplus M. An analysis of incorrectly folded protein models. Implications for structure predictions. J Mol Biol 1984;177:787–818.
37. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682.
38. Wang Y, Zhang H, Li W, Scott RA. Discriminating compact nonnative structures from the native structure of globular proteins. Proc Natl Acad Sci USA 1995;92:709–13.
39. Monge A, Friesner RA, Honig B. An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. Proc Natl Acad Sci USA 1994;91:5027–5029.
40. Dominy BN, Brooks CL. Identifying native-like protein structures using physics-based potentials. J Comput Chem 2001;31:147–160.
41. 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 and the AMBER forcefield with the generalized Born solvation model. Proteins 2003;51:21–40.
42. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40:389–408.
43. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
44. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 1999;285:1711–1733.
45. Rooman MJ, Kocher JP, Wodak SJ. Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. J Mol Biol 1991;221:961–979.
46. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 1995;249:493–507.
47. Laskowski RA, MacArthur WM, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystall 1993;26:283–291.
48. Berman H, 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.
49. Collura V, Higo J, Garnier J. Modeling of protein loops by simulated annealing. Protein Sci 1993;2:1502–1510.
50. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature 1996;381:272.
51. Go N, Scheraga HA. Ring closure and local conformational deformationals of chain molecules. Macromolecules 1970;3:178–187.
52. Wedemeyer W, Scheraga H. Exact analytical loop closure in proteins using polynomial equations. J Comput Chem 1999;20: 819–844.
53. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
54. Blundell T, Carney D, Gardner S, Hayes F, Howlin B, Hubbard T, Overington J, Singh DA, Sibanda BL, Sutcliffe M. 18th Sir Hans Krebs lecture. Knowledge-based protein modelling and design. Eur J Biochem 1988;172:513–520.
55. Burke DF, Deane CM. Improved protein loop prediction from sequence alone. Protein Eng 2001;14:473–478.
56. Rufino SD, Donate LE, Canard LHJ, Blundell TL. Predicting the conformational class of short and medium size loops connecting regular secondary structures: Application to comparative modelling. J Mol Biol 1997;267:352–367.