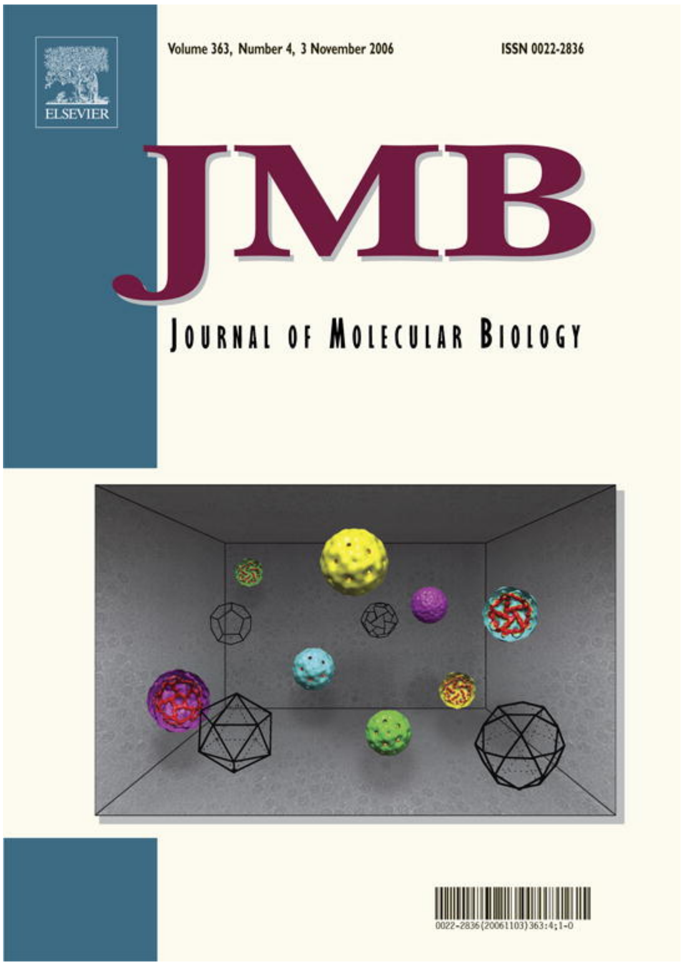
Provided for non-commercial research and educational use only.
Not for reproduction or distribution or commercial use.



ELSEVIER

Volume 363, Number 4, 3 November 2006

ISSN 0022-2836

JMB

JOURNAL OF MOLECULAR BIOLOGY

0022-2836(20061103)363:4;1-0

This article was originally published in a journal published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues that you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

http://www.elsevier.com/locate/permissionusematerial

doi:10.1016/j.jmb.2006.08.035

J. Mol. Biol. (2006) **363**, 835–857

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Minimalist Representations and the Importance of Nearest Neighbor Effects in Protein Folding Simulations

Andrés Colubri[1,2,3], Abhishek K. Jha[1,2,5], Min-yi Shen[4], Andrej Sali[4]
R. Stephen Berry[1,5], Tobin R. Sosnick[2,3]* and Karl F. Freed[1,5]*

[1]*Department of Chemistry*
The University of Chicago
Chicago, IL 60637, USA

[2]*Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL 60637*
USA

[3]*Department of Biochemistry and Molecular Biology*
The University of Chicago
Chicago, IL 60637, USA

[4]*Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry and California Institute for Quantitative Biomedical Research, University of California, San Francisco*
San Francisco, CA 94143, USA

[5]*The James Franck Institute*
The University of Chicago
Chicago, IL 60637, USA

**Corresponding authors*

In order to investigate the level of representation required to simulate folding and predict structure, we test the ability of a variety of reduced representations to identify native states in decoy libraries and to recover the native structure given the advanced knowledge of the very broad native Ramachandran basin assignments. Simplifications include the removal of the entire side-chain or the retention of only the C^{β} atoms. Scoring functions are derived from an all-atom statistical potential that distinguishes between atoms and different residue types. Structures are obtained by minimizing the scoring function with a computationally rapid simulated annealing algorithm. Results are compared for simulations in which backbone conformations are sampled from a Protein Data Bank-based backbone rotamer library generated by either ignoring or including a dependence on the identity and conformation of the neighboring residues. Only when the C^{β} atoms and nearest neighbor effects are included do the lowest energy structures generally fall within 4 Å of the native backbone root-mean square deviation (RMSD), despite the initial configuration being highly expanded with an average RMSD ≥ 10 Å. The side-chains are reinserted into the C^{β} models with minimal steric clash. Therefore, the detailed, all-atom information lost in descending to a C^{β}-level representation is recaptured to a large measure using backbone dihedral angle sampling that includes nearest neighbor effects and an appropriate scoring function.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Ramachandran basin; structure prediction; protein structure; simulated annealing; statistical potential

Author's personal copy

Introduction

First principle models of protein folding generally are preferred over statistical approaches because first principle models provide a theoretical framework to explain the underlying mechanisms, whereas purely statistical approaches only function as computational "black boxes". However, most first principles approaches have the disadvantage either of being too complex to be computationally feasible for proteins with more than a few residues[1] or too simplistic to be useful in situations where a more realistic representation is required.[2] In addition, knowledge based methods have improved substantially over the last several years and today are capable of generating native structures,[3–10] changes in stability upon mutation,[11] disorder propensities,[12] and binding affinities.[13]

An interest in combining first principles methods with statistical information has led us to construct a computational model that accommodates both a knowledge-based approach and a more fundamental methodology. Our present focus is on whether protein folding can be accurately depicted with a backbone representation that either lacks

Abbreviations used: DOPE, discrete optimized protein energy-function; DOPER, DOPE reduced; PDB, Protein Data Bank; SA, simulated annealing; RMSD, root-mean-square deviation; RB, Ramachandran basin; MD, molecular dynamics; LD, Langevin dynamics.

E-mail addresses of the corresponding authors: trsosnic@uchicago.edu; freed@uchicago.edu

0022-2836/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.

any side-chain degrees of freedom or is represented by at most a C^{β} atom. This simplified representation greatly diminishes the conformational search as it only involves the backbone dihedral angles, ϕ and ψ, with no consideration of side-chain rotamers.

To make up for the loss of side-chain information, local interactions are incorporated by sampling using a backbone rotamer library, constructed from the Protein Data Bank (PDB), that tabulates dihedral angles for single residues and also for sequences of dimers or trimers according to their amino acid identities and Ramachandran basin (RB) assignments.[14] Secondly, the tertiary interactions are treated using the scoring function, DOPE-C^{β}, which has residue-dependent parameters for all the atoms in the backbone and for all the C^{β} atoms. Structures are obtained by minimizing the scoring function with a computationally rapid simulated annealing (SA) algorithm using the PDB-based backbone rotamer sampling. These simplifications greatly reduce the search space and have been adopted in various fashions by other studies of protein folding.[15–20]

Rather than conducting an extensive conformational search through the entire backbone conformational space, effectively an *ab initio* structure prediction, our investigation focuses on the much less formidable task of generating structures given the advanced knowledge of the very broad native RB assignments for each residue. In spite of this simplification, we address a variety of issues including how to properly incorporate all-atom information in a reduced description of proteins. In the first place, it is not clear that a suitable scoring function for a reduced model can encode detailed packing propensities arising from all-atom interactions.[21–26] Our scoring function includes only terms involving at most the heavy atoms in the backbone and the C^{β} atoms and is obtained from a previously described all-atom statistical potential, DOPE (discrete optimized protein energy-function)[27,28] Heavy atoms are distinguished by the residue type to which they belong (e.g. $C_{\alpha}^{\text{alanine}}$ interacts differently than a $C_{\alpha}^{\text{valine}}$).

Because our model omits certain fine-grained, atomistic details of the system, we must ensure that the SA algorithm samples realistic regions of conformational space. For example, it is possible that a transition that is acceptable in the reduced representation would be forbidden if all additional variables were explicitly treated because local interactions between neighboring atoms impose strong geometric constraints on the motions of the backbone. As a primary consequence of these local interactions, the RBs are observed as the five predominantly occupied regions in ϕ, ψ maps for each amino acid (Figure 1). Previous *ab initio* folding simulations[29,30] have suggested that using RBs to constrain the conformational search in dihedral space would allow incorporating underlying atomistic information into the motions of a simplified model of a protein.

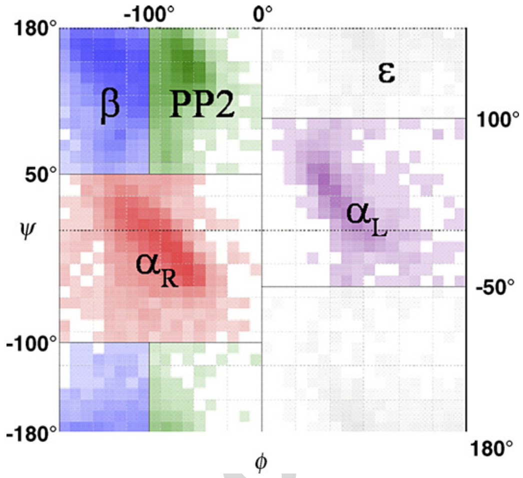


Figure 1. Specification of the five Ramanchandran basins. β (blue), poly-proline II, PPII (green), α_R (red), α_L (magenta) and ε (grey). The color intensity reflects the (ϕ, ψ) occupancy as calculated from all 4701 PDB structures.

Neighboring amino acids have also been shown to exert a substantial influence on the occupancies of the RBs.[31–33] This consideration provides the motivation for constructing a rotamer library of allowed backbone conformations for monomers, dimers and trimers, where amino acid information is coupled with RB assignments to reduce the number of allowed backbone conformations when knowledge is available concerning the basin occupancies. Even more importantly, this library inherently satisfies the constraints arising from the short range, all-atom interactions between nearest neighbor residues, information that is lost in the reduced representations. In a very recent paper, Rose and co-workers take a similar approach with remarkable success, although utilizing a library of pentamer rotamer conformationss with up to sevenfold smaller Ramachandran "mesostates", or sub-basins, along with the specification of secondary structure assignments for each amino acid.[34]

Here we describe the performance of the folding simulations for more than 50 proteins using a reduced representation that includes, in addition to the heavy atoms of the backbone, either the removal of the entire side-chain or the retention of only the C^{β} atom and that either neglects or includes nearest neighbor (NN) effects in the backbone sampling. The structures are generated with the residues being constrained to their native RBs during the entire SA minimization. Although such advanced knowledge precludes this study from being an *ab initio* structure prediction, when a protein is constrained to broad RBs, the protein may adopt a huge number of non-native, highly extended conformations as illustrated in Figures below. Thus, success with this "simpler" problem positions us to address our questions related to what level of representation is required to accurately generate native structures. In addition, the approach is useful for screening the foldability of designed sequences.

Author's personal c

Reasonable success is defined by whether native structures can be obtained with decent accuracy (<4 Å RMSD) and rapidity (<2 h of CPU time per protein). This level of success is achieved when the C^β atom and NN correlations are included. A variation is tested which enhances the C^β–C^β terms to include contributions from the other side-chain atoms, but this enhancement produces no significant improvement. The performance is slightly inferior for the protein model lacking the C^β atom but including NN effects, whereas the representation ignoring NN correlations is noticeably worse. Hence, the information lost in removing an explicit side-chain largely can be recovered with a scoring function that differentiates between heavy atoms on different residue types and when the searching algorithm contains a PDB-based backbone sampling that includes NN correlations. As a further test, side-chains have been inserted into each of the simulated structures, and the resultant structures are scored with the full all-atom statistical potential. The average quality of the predictions is the same whether or not the side groups are introduced, demonstrating that the DOPE-C^β statistical potential also encodes information concerning side group packing.

Author's perso

Results

Reduced representations and scoring functions

To investigate the level of detail required to reproduce native structures, we have utilized a variety of reduced models involving different side-chain representations, scoring functions, and backbone sampling schemes. The model either omits side-chains entirely, or the side-chains are represented by single C^β atoms. The scoring functions are derived from DOPE, an all-atom statistical potential.[27,28] For the model lacking side-chains, only the terms involving the backbone heavy atoms are included in the scoring function termed DOPE-BB. The C^β side-chain scoring function DOPE-C^β includes terms involving the C^β atoms as well.

We also have tested a variant of this scoring function, DOPER, which enhances the C^β–C^β terms to include energetic contributions from the other heavy atoms in the side-chains. This extra energy is the sum of the terms between all heavy atoms on the two side-chains, (averaged over all conformations in the all-atom structures used to generate the DOPE potential). Representative plots of the DOPE and DOPER interaction scores between two different amino acids are contrasted in Figure 2(a) and (b) as a function of the spatial distance separating the respective C^βs. The statistical C^β–C^β DOPER interactions contrast sharply with the corresponding explicit atom–atom C^β–C^β interactions from DOPE, even for the same amino acid pairs. All C^β–C^β interactions of DOPE are more similar to each other

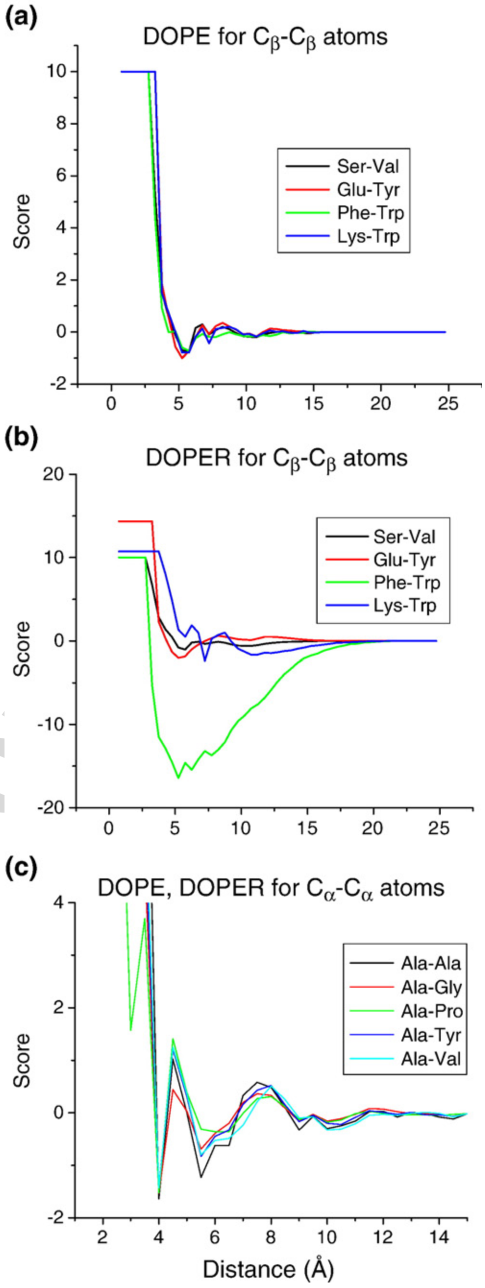


Figure 2. The dependence of the DOPE and DOPER statistical potentials on their inter-atomic distance for different pairs of amino acids. (a) In the all-atom DOPE statistical potential, C^β–C^β interactions are similar even though the chemical compositions of the amino acids are very dissimilar. (b) For the DOPER statistical potential, which has enhanced C^β–C^β terms, the shape of the curves changes substantially, showing that the effective C^β–C^β interaction in the reduced C^β model encodes information concerning the amino acid type and the side group constraints. Small hydrophobic residues, such as serine and valine, yield a DOPER profile that resembles the corresponding DOPE curve. Amino acids with large side-chains display very distinctive profiles reflecting their chemical compositions and complimentarity. (c) The C^α–C^α interactions of DOPE and DOPER are the same, but depend on amino acid type to a degree, and thus contain information about the side groups.

Author's personal copy

than their DOPER counterparts, which exhibit substantial variation among different amino acids. This difference is expected because the DOPER C^{β}–C^{β} interaction includes the sum of interaction of all the heavy atoms in the entire side-chain. As we pass from interactions between pairs of amino acids with small side-chains (serine–valine, glutamic acid–gtyrosine) to pairs of bulkier residues (phenylalanine–tryptophan and lysine–tryptophan), the DOPER profiles increasingly differ from the explicit atom-atom DOPE interaction curves. It is also evident that the hydrophobicity of the amino acids is captured by DOPER, as exemplified by the significantly increased strength between the phenylalanine–tryptophan and lysine–tryptophan effective interactions. The dependence of the DOPE statistical potential on both the atom types and the amino acid identities is illustrated in Figure 2(c) for selected C^{α}–C^{α} interactions. This dependence on amino acid identity, even for the otherwise chemically identical atom type, assists the statistical potential in describing the influence of side-chain packing.

Test with decoy sets

To examine the loss of information incurred by use of the reduced models, we compare the energy computed using the all-atom DOPE statistical potential with those of the three other scoring functions for reduced models of the seven proteins included in the Park-Levitt four-state reduced decoy set,[35] which can be downloaded from the Decoy 'R' Us website†. In the resulting scatter plots (Figure 3), the DOPE-C^{β} potential has the highest correlation coefficient $R{\sim}0.9$ with the all-atom potential, although the DOPE-BB and DOPER potentials yield only slightly lower coefficients, $R{\sim}0.85$.

We briefly describe the compilation of the three libraries of decoys used in this study. The Zhou decoy set includes 96 standard multiple decoy sets of proteins with known X-ray structure.[26] The Baker decoy set includes over 75,000 members for 41 proteins whose structures have been determined with either X-ray or NMR experiments. This decoy set is generated using the Rosetta algorithm and is a subset of the structures used to test an all-atom scoring function.[18] The study by Zhou *et al.* excludes from the original decoy sets those associated with proteins whose structures have been determined by NMR as well as decoy sets of globins and immunoglobin. The decoys sets excluded by Zhou *et al.* comprise our third and final library, which is labeled as Others.

To further examine the utility of the four scoring functions, we test and compare their ability to identify the native structure in three different libraries of decoys (Table 1). The success rate (the percent of native structures that are ranked by their energy scores as number one) is highest for DOPE, the all-atom potential, and decreases for reduced-model scoring functions (Table 1). In addition, the Top5 measure (the percent of native structures that are ranked by their energy score as one of the five structures with lowest energy) displays a similar trend. Finally, we consider the *Z*-score, another measure of the quality of these statistical potentials for dealing with different decoy sets The *Z*-score is defined as:

$$Z = \frac{Energy(Native) - \langle Energy(Decoy)\rangle_{library}}{\sqrt{\langle Energy(Decoy)^{2}_{library}\rangle - \langle Energy(Decoy)\rangle^{2}_{library}}} \qquad (1)$$

where the angular brackets denote the average over the library.[36] *Z*-scores reflect the quality of both the scoring function and the decoy set (worse decoys sets result in better *Z*-scores). The performance of the energy functions is relatively poor for decoy sets that include NMR structures. Moreover, the reduced energy functions do not perform as well as the all-atom potential. Regardless, our overall goal is to develop an algorithm for generating native-like structures and not a statistical potential that can identify the native state from a decoy set using *Z*-scores. From this perspective, we require a suitable potential that can produce an accurate representation of the structure of the native state for any given sequence when used in conjunction with an adequate sampling protocol.

Intra-basin folding simulations

As a first test of the models, we consider simulations that begin with a random assignment of dihedral angles for each amino acid residue within their native RBs, and proceed by minimizing the DOPE-C^{β} score using a SA algorithm (Figure 4). The dihedral angles are constrained to remain within their native RBs during the entire annealing run. New conformations are generated by choosing a trimer, dimer, or monomer from the backbone rotamer library, subject to compatibility with the amino acid sequence and basin assignments. The algorithm first tries choosing dihedral angles from trimers, and defaults to dimers and then to monomers when configurations for the trimer or dimer are absent, respectively, in the rotamer libraries. After a fixed number of steps in which all moves are accepted, the annealing follows a cooling schedule that decreases the temperature until convergence is reached. In the simulations presented here, the number of steps at each temperature is 100.

We consider 50 small globular proteins to test the intra-basin folding algorithm (Table 2). The 50-protein set is generated by combining 41 proteins used by Baker *et al.*[36] with nine additional commonly studied proteins. This test set contains a heterogeneous sample of different protein topologies: α helix bundles, α/β proteins, and β-only structures. A total of 100 separate trajectories are performed for each protein, with every trajectory starting from a

† http://dd.stanford.edu/

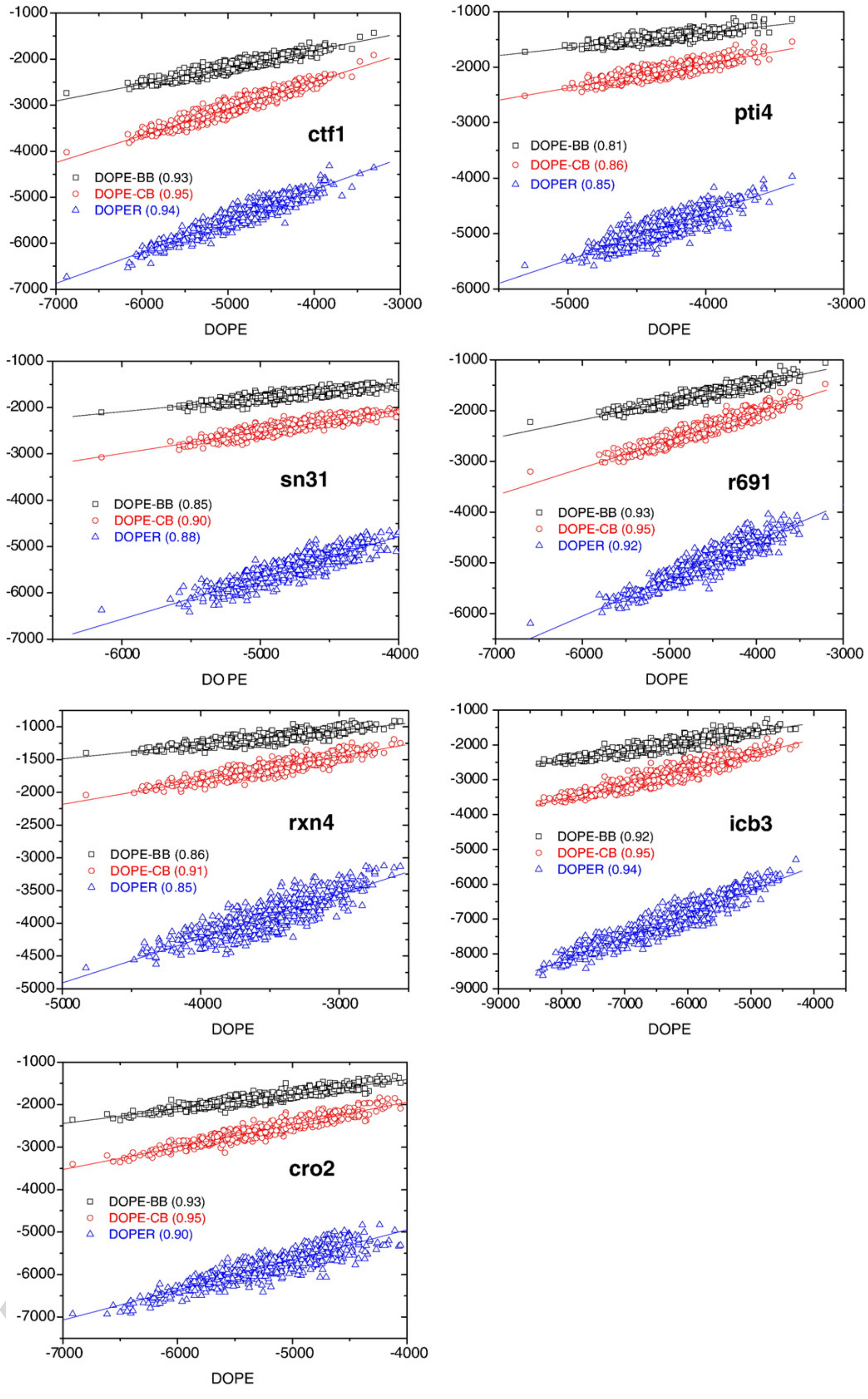


Figure 3. Correlation between all-atom DOPE scores and those of the three other reduced representations. Energies are highly correlated between the all-atom DOPE statistical potential and the DOPE-C$^{\beta}$, DOPE-BB only, and DOPER scoring functions for seven proteins taken from the Park & Levitt four-state reduced decoy set. The high correlation indicates that each of the reduced scoring functions captures a large majority of the information content of the all-atom based statistical potential DOPE from which they are derived. (*R* values in parentheses.)

Table 1. Success rates and Z-scores for different scoring functions and decoy sets

Decoy (no. of proteins)	Scoring function	DOPE (all-atom)	DOPE-C$^{\beta}$	DOPER	DOPE-BB
Zhou (96)	Success(%)[a]	83	65	57	59
	top 5(%)[b]	89	76	72	72
	Z-score[c]	3.8±3.1	2.9±2.7	2.6±2.3	2.6±2.5
Baker (41)	Success(%)	27	20	0	12
	top 5(%)	44	27	15	27.5
	Z-score	1.5±2.2	0.9±2.3	0.7±1.6	0.8±2.2
Others (171)	Success(%)	18	11	9	6
	top 5(%)	37	20	16	14
	Z-score	0.7±2.1	−0.4±2.4	−0.08±2.03	−0.97±2.79
All (308)	Success(%)	39	29	23	23
	top 5(%)	54	39	33	34
	Z-score	1.8±2.9	0.8±2.9	0.9±2.4	0.4±3.1

[a] Success is defined as the native structure having the lowest energy score.
[b] Top5 refers to number of total cases when the native is one of the five structures with lowest energy.
[c] Z-score values are given as the average and the standard deviation for the decoy sets.

different random assignment of dihedral angles within the native basins. For each trajectory, each successive minimum is recorded until the convergence criterion is satisfied. Generally, about 50 such minima from each trajectory are recorded. Using a Pentium IV 2.8 Ghz with 512 Mb of RAM, each run takes an average of 500 s for a protein of 70 residues, when executing 100 annealing steps per temperature. The overall running time is roughly proportional to the number of annealing steps per temperature; hence, a simulation performed with 500 annealing steps per temperature is about five times slower.

Because simulations using reduced models could produce structures for which the side groups clash significantly, tests are made of side-chain packing. After each SA run using the reduced representation, the side groups are introduced employing the SCWRL program version 3.0,[37] without further backbone motions. SCWRL employs a simple energy function based on a backbone-dependent rotamer library and a piece-wise linear repulsive steric energy to remove atom clashes. A final scoring is made using the full heavy atom statistical potential DOPE to ensure that the final structure properly describes the protein packing.

Comparisons of the RMSD for the lowest energy structure obtained before and after the introduction of the side groups, scored with the DOPE-C$^{\beta}$ potential and with the all-atom DOPE potential, respectively, indicates that the inclusion of side-chains generally provides only a modest improvement in the RMSD of the lowest energy structure (Figure 5; Table 2, compare columns 5 and 6). This test has the physical significance of demonstrating that our protein structure algorithm is indeed consistent with good side-chain packing. Because the computational cost is minimal for introducing side-chains with SCWRL and for scoring with the all-atom DOPE potential of a single structure after the SA process (which utilizes a model without side-chain rotamers), this procedure of adding side-chains is applied throughout our study when identifying lowest energy structures and calculating RMSD values.

Intra-basin search is non-trivial

Even when the energy minimization constrains each residue's backbone to remain within its very broad native RBs, a huge number of conformations are still available. Initial configurations, generated from the rotamer library by a random assignment of dihedral angles in the native RBs, are highly unfolded with dimensions comparable to denatured proteins. Moreover, these initial conformations are assembled by piecing together trimers and dimers taken from the rotamer library, so that the conformations even satisfy local chemical and geometrical correlations. However, as demonstrated by Figure 6(a), the lowest RMSD of the initial configurations never falls below 5 Å, with the average around 10 Å or higher. It is also revealing to examine the relationship between the initial configuration and the final result of the minimization. Figure 6(b) presents representative scatter plots of the RMSD of the initial configuration *versus* the RMSD of the final configuration and of the initial against the final configurations' DOPE energy for all annealing runs. The scatter plots indicate that the outcome of the energy annealing is independent of the proximity of the initial state to the native structure.

Figure 4 depicts typical initial 3D structures for 1UBQ. The native structure of this protein features an α-helix between residues 23 and 34, and the Figure shows that this helix is also partially present in the initial structures. This appearance of a helical portion arises because stretches of more than four consecutive residues in the α RB are highly constrained to adopt a standard helical conformation. However, these pictures also display a complete lack of native long-range interactions in these initial structures. Nevertheless, the presence of partial helical structure in the initial conformations suggests that the SA minimization would be expected to fare better for α proteins. Thus, the 50

Author's personal copy

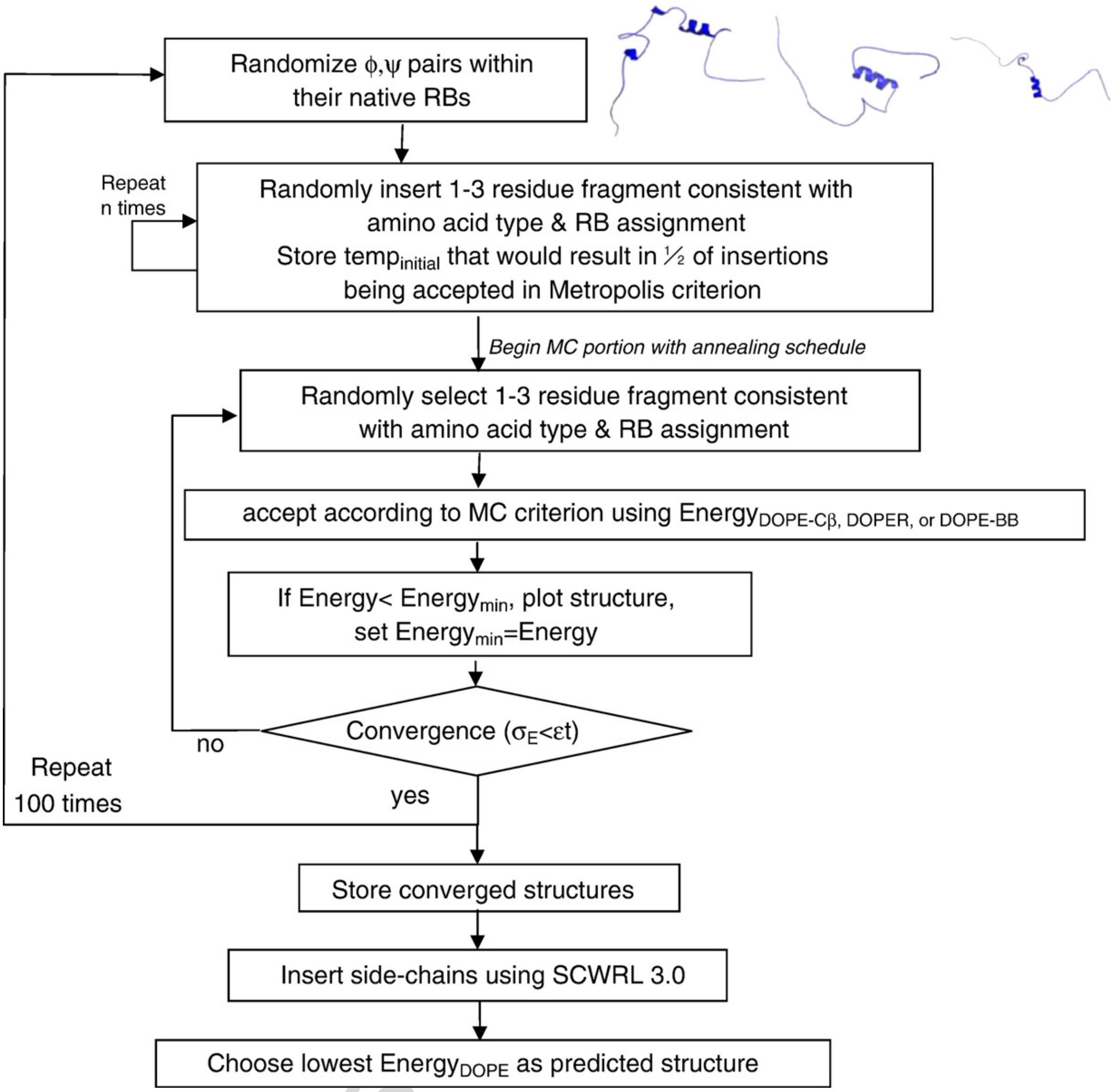

Randomize ϕ,ψ pairs within their native RBs
Repeat n times
Randomly insert 1-3 residue fragment consistent with amino acid type & RB assignment
Store temp$_{initial}$ that would result in ½ of insertions being accepted in Metropolis criterion
Begin MC portion with annealing schedule
Randomly select 1-3 residue fragment consistent with amino acid type & RB assignment
accept according to MC criterion using Energy$_{DOPE-C\beta,\ DOPER,\ or\ DOPE-BB}$
If Energy< Energy$_{min}$, plot structure, set Energy$_{min}$=Energy
Convergence (σ_E<εt)
no
yes
Repeat 100 times
Store converged structures
Insert side-chains using SCWRL 3.0
Choose lowest Energy$_{DOPE}$ as predicted structure

Figure 4. Flow chart of the simulated annealing algorithm. 3D renderings of typical initial structures are shown at the top for 1ubq. While these conformations display no native tertiary interactions, they feature the native α−helix in the correct position. The helix appears because long stretches of α RBs almost uniquely determine alpha helical conformations.

Authors

proteins studied also include β proteins whose initial structures are even more devoid of native, long-range structures.

Many examples demonstrate the highly non-trivial character of the intra-basin search for generating a good approximation to the native structure (Figure 7). For example, among the five lowest energy structures for 1VII, a small α-helix bundle of 36 residues, the lowest energy conformation is almost the correct native structure except for the presence of an incorrect orientation for the C-terminal helix. This observation implies that the specification of the RBs does not uniquely determine the spatial arrangement of the secondary structural elements, even in very simple cases such as this one where the initial structures contain some helical portions.

Generated structures

The resulting all-atom structures are ranked according to their all-atom DOPE energy score, and the five conformations with lowest energies are selected for comparison with the native structure. From this set of five conformations, a structure with less than 4 Å of backbone RMSD is found in 44% of the cases, with no obvious correlation to the size and secondary structure topology (Table 2). The accuracy tends to be slightly better for proteins with only α-helices, perhaps because β structures usually involve more complex topologies and long-range interactions, but more likely, because α-helices are easily formed during the initial assignment of dihedral angles, thereby probably expediting the annealing search. Figure 7 presents

Table 2. Results for SA runs using DOPE-C$^{\beta}$

PDB code	Class, N_{res}	Source	Native lowest energy	Predict RMSD (all-atom)[a]	Predict RMSD (C$^{\beta}$)[b]	Lowest RMSD[c]	RMSD< 2 Å (%)	RMSD< 4 Å (%)	RMSD < 6 Å (%)	RMSD< 8 Å (%)	Min RMSD in top5[d]	Max RMSD in top5	Ave RMSD in top5
1bdc	α,60	NMR	No	7.15	7.15	3.5	0	3	41	91	5.24	7.15	6.05
1bw6	α,56	NMR	No	9.16	10.21	3.19	0	1	26	40	5.42	9.85	7.87
1bxy	αβ,60	X-ray	Yes	2.94	3.25	2.15	0	15	43	49	2.48	4.23	3.07
1ctf	αβ,67	X-ray	Yes	2.24	2.68	1.75	2	39	64	89	2.24	3.72	2.80
1kjs	α,74	NMR	No	4.12	5.24	3.67	0	4	49	78	4.12	6.98	5.55
1msi	β,60	X-ray	Yes	4.34	4.34	4.26	0	0	16	45	4.34	.96	6.23
1mzm	α,71	X-ray	Yes	4.07	6.29	2.88	0	33	51	81	3.49	4.07	3.71
1orc	αβ,56	X-ray	Yes	2.59	3.23	2.13	0	65	92	95	2.13	8.22	3.64
1ubq	αβ,76	X-ray	Yes	4.53	3.73	1.93	2	42	86	97	2.73	4.56	3.74
2pdd	α,43	NMR	No	8.78	**3.61**	3.61	0	2	33	79	4.26	8.78	6.99
1res	α,35	NMR	No	2.88	3.24	0.70	54	79	98	100	0.70	2.88	1.52
1vii	α,36	NMR	No	7.26	7.53	2.55	0	6	13	93	7.06	7.59	7.25
1uxd	α,43	NMR	No	2.86	**2.85**	2.85	0	5	10	78	2.85	6.03	3.66
1uba	α,45	NMR	No	6.72	6.72	3.07	0	22	67	98	6.09	6.72	6.36
1gab	α,47	NMR	No	2.01	**1.97**	1.79	22	97	99	99	1.96	2.40	2.18
1prb	α,53	NMR	No	2.10	**1.98**	1.84	15	96	100	100	2.00	2.25	2.12
1enh	α,54	X-ray	Yes	1.29	2.93	1.24	25	92	100	100	1.29	2.65	1.9
1am3	α,57	X-ray	Yes	2.42	2.50	1.86	4	95	100	100	2.41	3.61	2.70
1r69	α,61	X-ray	Yes	3.49	4.92	2.32	0	28	86	97	2.57	4.76	3.64
1utg	α,62	X-ray	No	4.20	4.63	2.67	0	24	98	100	4.20	5.13	4.52
2ezh	α,65	NMR	No	4.15	4.31	3.38	0	14	77	81	3.76	4.15	3.93
1a32	α,65	X-ray	No	5.50	**5.20**	2.61	0	36	83	95	4.83	5.56	5.28
1nre	α,66	NMR	No	7.12	8.65	1.44	8	50	78	95	2.49	7.12	4.61
1ail	α,67	X-ray	Yes	2.18	8.05	1.97	1	39	41	71	2.09	7.60	3.49
1lfb	α,69	X-ray	No	2.54	3.47	1.76	3	83	97	98	2.54	3.86	3.12
1nkl	α,70	NMR	Yes	2.52	2.52	2.52	0	5	16	70	2.52	9.08	5.47
1pou	α,70	NMR	Yes	12.46	**9.29**	2.99	0	1	3	24	4.82	12.46	9.53
5icb	α,72	X-ray	Yes	3.15	**2.78**	2.56	0	44	57	73	2.78	5.40	3.50
1hyp	α,75	X-ray	Yes	8.11	**4.94**	3.00	0	18	58	83	4.82	8.11	6.80
1cc5	α,76	X-ray	Yes	6.85	7.76	4.79	0	0	3	40	6.85	10.39	8.56
1cei	α,85	X-ray	Yes	6.51	12.65	4.29	0	0	2	13	6.14	11.67	7.81
1ptq	αβ,43	X-ray	Yes	2.07	7.69	1.46	6	17	26	54	1.87	8.32	4.39
3gb1	αβ,56	NMR	Yes	5.93	**5.19**	2.51	0	11	62	92	3.62	6.16	5.34
1aa3	αβ,56	NMR	No	6.44	**6.35**	5.51	0	0	9	94	5.88	6.78	6.41
1pgx	αβ,57	X-ray	Yes	4.27	**3.05**	2.67	0	31	83	95	2.92	5.94	3.98
1tif	αβ,59	X-ray	Yes	2.45	2.45	2.12	0	22	57	88	2.12	3.64	2.77
2ptl	αβ,60	NMR	Yes	7.70	**6.12**	4.48	0	0	3	30	4.48	10.73	7.66
1dol	αβ,62	X-ray	Yes	7.61	9.67	4.00	0	0	5	22	7.61	9.23	8.61
2fow	αβ,66	NMR	Yes	7.51	7.83	6.08	0	0	0	28	7.51	11.10	9.43
1afi	αβ,72	NMR	Yes	9.52	**9.51**	5.57	0	0	1	6	6.97	9.52	8.43
1vcc	αβ,77	X-ray	Yes	6.67	11.63	3.83	0	1	9	29	6.36	11.26	7.77
2fxb	αβ,81	X-ray	Yes	9.88	**8.32**	4.88	0	0	6	20	8.12	11.17	9.94
1e0l	β,37	NMR	Yes	8.49	**4.78**	3.77	0	1	37	48	4.70	8.81	7.10
1vif	β,48	X-ray	Yes	4.30	**3.09**	1.63	2	24	54	80	1.63	6.06	4.14
1bq9	β,53	X-ray	Yes	7.95	**7.8**	2.32	0	23	42	64	3.48	9.38	6.62
5pti	β,55	X-ray	Yes	3.28	**3.14**	3.09	0	32	73	82	3.28	8.30	5.59
1tuc	β,61	X-ray	Yes	3.20	3.20	3.2	0	2	8	23	3.20	7.48	5.06
1csp	β,64	X-ray	Yes	3.98	3.98	3.82	0	2	8	45	3.82	9.08	6.16
1sro	β,66	NMR	Yes	5.33	6.14	4.57	0	0	6	36	5.33	9.87	7.59
1g6p	β,66	NMR	Yes	8.34	**8.12**	7.19	0	0	0	2	8.34	12.38	10.04

Structures are determined using the SA algorithm and the DOPE-C$^{\beta}$ scoring function and with sampling that includes nearest neighbor effects. Unless specified otherwise, final energy rankings are determined for these structures using the all-atom DOPE statistical potential after inserting side-chains using the SCWRL statistical potential. C$^{\alpha}$ RMSDs are relative to the native structure in Å.

[a] C$^{\alpha}$ RMSD for lowest energy structure relative to the native structure in Å (using the DOPE score) among the final structures from 100 trajectories.

[b] C$^{\alpha}$ RMSD for lowest energy structure relative to the native structure in Å (using the DOPE-C$^{\beta}$ score) among the final structures from 100 trajectories.

[c] Lowest RMSD for all the structures among the final structures from 100 trajectories.

[d] Minimum and maximum RMSD between the native and the five lowest energy structures (using DOPE score) among the final structures from 100 trajectories.

Author's personal copy

3D renderings of the predicted (lowest energy) structures for a selection of proteins, along with the superimposed native fold, and the corresponding scatter plots of the DOPE energy and RMSD for the structures generated. Although the 50 target proteins are not explicitly excluded from the rotomer library, only ~4, 6, 9, 12, and 16% of the pair of dihedral angles from the 50 predicted structures (3035 pairs total) are found to lie within 1, 2, 3, 4, and 5° of the native dihedral angles, respectively. The lack of native angles in our structures indicates that the algorithm undergoes a meaningful search

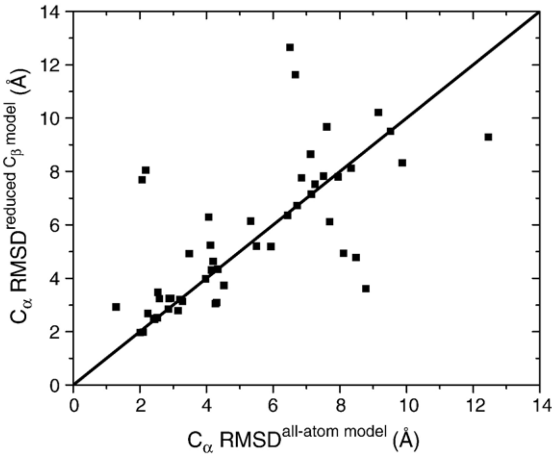


Figure 5. Scoring after the introduction of side-chains produces only marginal improvement in predicted RMSD. For the 50 proteins investigated, multiple structures are obtained using the SA algorithm with the reduced C^β-only representation and the DOPE-C^β scoring function. The RMSD of the lowest energy structures are nearly the same on average, whether they are scored before the introduction of the side groups with the DOPE-C^β statistical potential or after the introduction of side groups (with SCWRL)[37,38] and then scored with the all-atom DOPE statistical potential. The $y=x$ line is inserted for the reader's benefit.

within the basins and that our successes are not due to the insertion of native fragments, which could trivialize our results, for example, if there were only a few trimers in the library to search though. Nevertheless, the recovery of the native angles is desirable, indicating the algorithm and potential can be improved. The algorithm yields comparable accuracy over the range of chain lengths from 20 to 80 residues.

Alternative scoring functions

We have repeated the SA using both the DOPER scoring function, for which the C^β–C^β interaction energy includes the total interaction energy for all the heavy atoms in the two side-chains, and a backbone-only DOPE-BB scoring function. The quality of the predictions is not significantly different when using the DOPE-C^β and DOPER potentials (Table 3); however, the DOPE-BB potential performs worse (Table 4).

We next test whether the inclusion of information about the identity and conformation of the neighboring residues in the backbone sampling is critical to the success of the reduced model. Using the DOPE-C^β scoring function but ignoring neighbor information in the backbone sampling, the results are severely degraded, even below those generated using the DOPE-BB scoring function (Table 5). Hence, the detailed, all-atom information lost in descending to a C^β-level representation can be recaptured to a large measure using backbone sampling that includes NN effects. In summary, in order to obtain reliable structures which often are within 4 Å RMSD of the native structure, the SA treatment requires the use of a statistical potential that minimally employs a C^β side-chain representation and backbone dihedral angle sampling that incorporates nearest neighbor effects.

Longer cooling schedules

The current version of the annealing algorithm uses a fixed number of conformational transitions at each temperature. This fixed number is a critical parameter affecting the outcome of the minimization procedure. We expect that the treatment of larger proteins will require more steps in the folding process because larger proteins have more total conformations available. Figure 8 presents profiles for the DOPE scoring function and the RMSD as a function of the number of steps per temperature. Convergence is attained in all cases, but only when the simulations involve more than 100 of steps per temperature does the annealing process appear to equilibrate towards the proper low energy configurations. The use of longer cooling schedules has the disadvantage of rendering the computations more expensive. On the other hand, quenching the system more rapidly might not be too disadvantageous provided that the algorithm is still capable of finding good, near-native conformations. Additional simulations for 1UBQ indicate that while longer annealing runs do not yield significantly higher accuracy, they merely increase the number of structures that are close to the native (Table 6). This number of near native structures appears to plateau at 200 steps per annealing temperature, which seems to indicate that the optimal number of steps is between 100 and 200, at least for proteins with less than 80 residues. Longer annealing runs do not improve the results for those proteins poorly covered by the trimer library (see below), thus reinforcing the conclusion that having enough trimer configurations is fundamental for the algorithm to provide good predictions.

Correlation between accuracy and richness of the trimer library

Figure 9 displays the presence of a clear correlation between the RMSD for the most native-like simulated model and the number of positions along the input sequence for which trimer conformations are unavailable in the rotamer library, given the sequence and basin specifications for the trimer. Even though such a correlation is expected, it also reflects a non-trivial feature of our approach: simulated low energy structures are very good for a protein when the trimer library is rich enough for that protein's sequence. If we define an amino acid sequence together with its basin assignment to be well covered by the trimer library when there are less than ten missing trimers, 93% (38 out of 41) of the well covered sequences from the proteins considered in this study have predicted structures with less than 4 Å backbone RMSD from the native

Author's personal copy

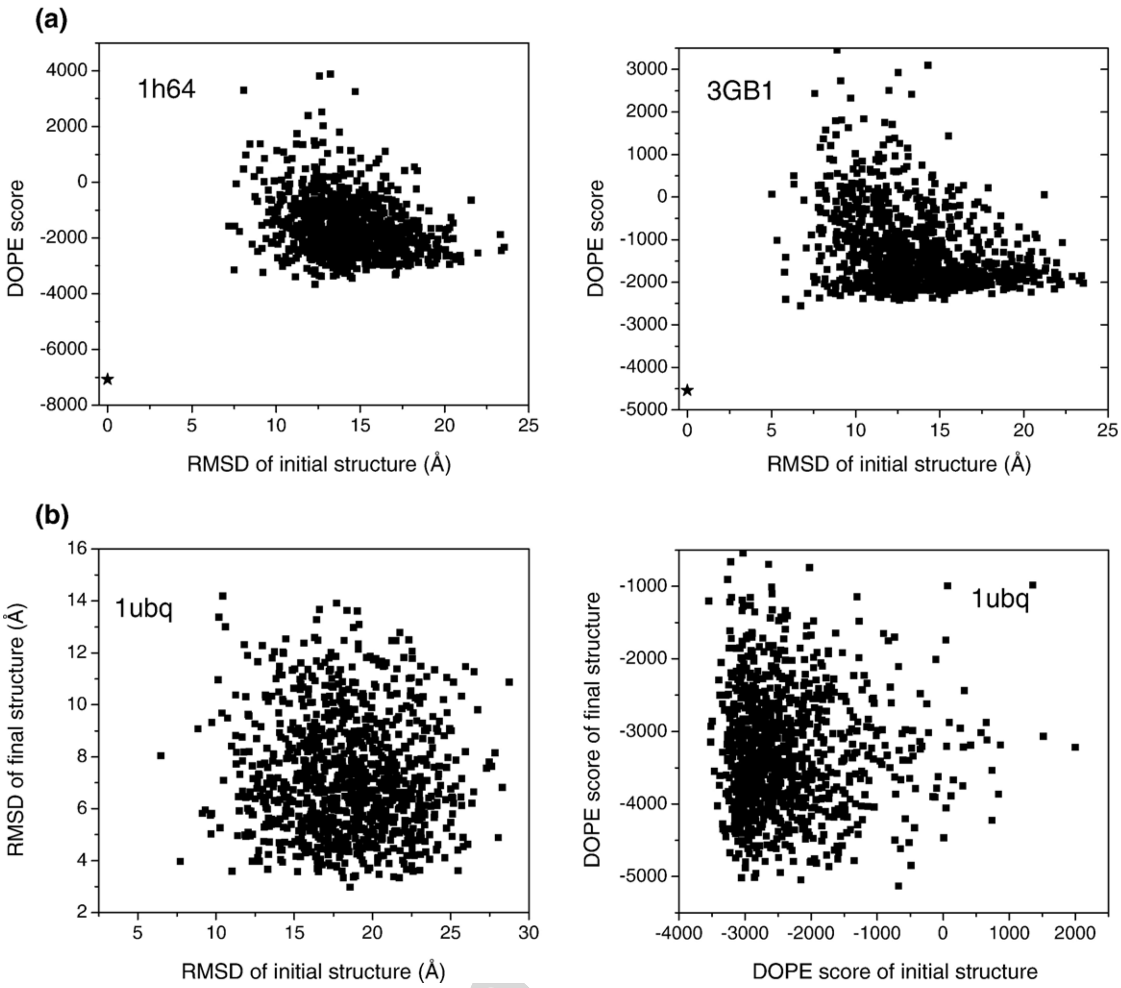


Figure 6. Intra-basin search is non-trivial. (a) Scatter plots for two proteins in which the RMSD values of the initial configurations for the SA runs are plotted against the corresponding DOPE score. The initial configurations bear little resemblance to the native structure, even though the dihedral angles are taken from the backbone library and are restricted to remain within the native RBs. The native structure is also included in the scatter plots as the leftmost star-shaped point with RMSD = 0 Å. (b) Scatter plots are generated from the same set of runs for 1ubq. The initial and final structures for each of these runs are recorded, and the RMSD values to the native structure and the DOPE score are calculated. The scatter plots reveal the lack of a correlation between the initial and final structures for a given annealing run. The initial conformation might be very high in energy and have a poor RMSD value , and still the run might lead to a reasonable structure.

Author's p

structure. This result is particularly significant because it demonstrates that if the algorithm can sample from a number of local structures for each trimer fragment, the simulations produce the correct global fold. Recalling that the SA minimization routine defaults to dimer or monomer conformations when there are no trimer conformations available with the sequence and basin assignments, we can further conclude that the backbone dihedral search becomes ineffective if the (ϕ, ψ) torsional angles are sampled by taking monomers or dimers as the basic units.

Energy function and hydrogen bonds

Leaving aside the fact that a C^{β}-only energy function has intrinsic inaccuracies that can only be rectified with a more detailed model that includes explicit side-chains, our current statistical potential might be improved by adding additional terms to the energy function. First of all, even predictions close to the native structure (~2 Å) have a distinctive feature that readily differentiates them from the native structures. The simulated structures lack the optimal hydrogen bond patterns of the native conformation (Figure 10). Even though the current energy function has terms accounting for the interaction between the main-chain oxygen and nitrogen atoms and thus encodes preferences due to hydrogen bonding, these interactions depend only on distance and not orientation. The hydrogen bonds, however, are very sensitive to the angular orientations between the donor and acceptor groups. Hence, an explicit orientation-dependent hydrogen bond term[38] should improve the description of the hydrogen bonds.

Authors personal copy

Comparison with other methods

After completion of these calculations, a related work by Rose and co-workers was published,[34] motivating a detailed comparison between their algorithm and ours. They build models using native-like pentamer fragments and an energy function composed of hydrogen bonds, steric repulsion and favorable interactions induced by minimizing the radius of gyration. They often obtain a very high degree of success for the six proteins studied, with the most stable structure being within 2 Å of the native structure for three of their targets. The initial dihedral angles in each pentamer are constrained to one of the 36 more confined 60° × 60° mesostates than our five broader RBs (for example, our α-basin covers the area of 7.5 mesostates in their model), and their pentamer library specifies secondary structures, implying that their approach includes considerably more constraints concerning the native structure than our method. Moreover, these additional constraints are more likely to facilitate a better identification of the correct structure for turn fragments.

In order to perform an in-depth comparison using the DOPE-C^{β} potential with NN effects for the six proteins investigated by Rose and co-workers, it is also necessary to investigate the importance of adding a hydrogen bond term because of the central role played by these bonds in the simulations by Rose and co-workers. Thus, to render the comparison more a test of using a reduced basin size with larger fragments, on the one hand, against using a more sophisticated interaction potential, on the other hand, we provide the results of preliminary simulations obtained with HOPE-C^{β}, a generalization of DOPE-C^{β} which includes the amide proton to better represent hydrogen bonding, followed by clustering of the final structures,[39] a step Gong *et al.* find necessary for identifying the native fold. In addition to these simulations, a test is also made for a version of the statistical potential DOPE-C^{β} where the heavy atom backbone interactions depend only on the atom type but not on amino acids, while the C^{β} atoms are still distinguished based on residue types, to provide more comparable situation to the poly-alanine representation by Gong *et al.*

The results of the simulations using DOPE-C^{β} and HOPE-C^{β} are summarized in Tables 7 and 8, respectively, for the half-dozen proteins studied by Gong *et al*. The predicted structures are marginally better when HOPE-C^{β} is used as a scoring function for the simulations. Lower RMSD structures emerge from the simulations, but the scoring function fails to identify these structures as the lowest energy predictions. The predicted structures, however, are closer to the lowest RMSD structures when clustering is used to sort the final structures obtained from each trajectory (Table 9). Finally, the third set of simulations (with heavy backbone atom interactions independent of residue type) yield predicted structures that are marginally poorer, so the details are not presented.

Compared to the work of Gong *et al.*, our HOPE-C^{β} results are slightly better in two of the six proteins, but significantly worse for 1IFB, a 131 residue protein. This intestinal fatty acid binding protein has a solvated core,[40] which we fail to find whereas Rose and co-workers generate a native-like structure for this sequence. It is, thus, critical to have additional constraints, such as mesostates and secondary structure, in identifying the native-like structures for this unusual protein architecture.

Discussion

Intra-basin folding simulations

The intra-basin folding simulations are promising in the sense of demonstrating that, given knowledge of local geometrical constraints in the native state, it is possible to rapidly obtain reasonable predictions of the native structure for a wide range of proteins using a C^{β}-based model and sampling that includes the influence of nearest neighbor interactions. Given the non-trivial nature of the intra-basin conformational search, these results are important by themselves, and not only as one necessary validation of a more complete folding algorithm that does not require the specification of the native RBs. In addition, the RMSDs are comparable for the lowest energy structures before and after the introduction of the side groups with SCWRL,[37] scored with the DOPE-C^{β} and the all-atom DOPE potentials, respectively (Figure 5). This result demonstrates that the algorithm generates physically reasonable protein structures consistent with good side-chain packing.

Overall, the all-atom DOPE energy value is the minimum for the native structure in 68% of the proteins tested in this study, and the decoys close to the native state are those with similarly low energies. The scatter plots in Figure 7 show that the RMSD values and the DOPE scores are highly correlated. The appearance of false positives, impacting negatively on the success rates, is probably due for the most part to the standard deviation and low specificity of the DOPE statistical potential at the current level of resolution of DOPE (~3 Å).

A recent, related work by Rose and co-workers considers protein structure identification using a poly-alanine representation of the protein, an initial dihedral angle specification, an explicit hydrogen bond term, a specification of each amino acid's secondary structure type, and an analysis based on clustering of final structures.[34] Their performance is often, but not always, better than ours, and they obtain a very high degree of success for the six proteins studied, with the most stable structure being within 2 Å of the native structure for three of their six targets (Table 9). Their initial dihedral angles are constrained to one of the 36 $60^{o} \times 60^{o}$ mesostates whereas we utilize five broader RBs (for example, our α-basin covers the area of 7.5 mesostates in their model). Their fragment library uses pentamers and specifies secondary structures, implying that their approach includes considerably more constraints concerning the native structure than our method. The prediction of our model improves when we include hydrogen bond interactions (using the HOPE-C^{β} scoring function) and clustering, in agreement with their findings. Also, the speed of our method (~2 CPU hours per protein) enables us to consider 50 different proteins.

Our method shares common elements with Baker's Rosetta prediction algorithm.[18,41] Their PDB-based backbone sampling uses the conformations from the PDB for trimers or nonamers, biased according to secondary structure prediction. Rosetta also employs a C^{β}-representation. However, their statistical potential only distinguishes the C^{β} heavy atom for different residue types; the backbone interactions depend only on atom type and not on amino acids, as is the case for the DOPE statistical potential used here. Moreover, an additional environmental dependence is incorporated into their statistical potential using Bayesian methods. In addition, their docking of structural elements is based upon observed statistics in the PDB.

The present work differs from studies by Colubri & Fernandez[29] in that these works are aimed at

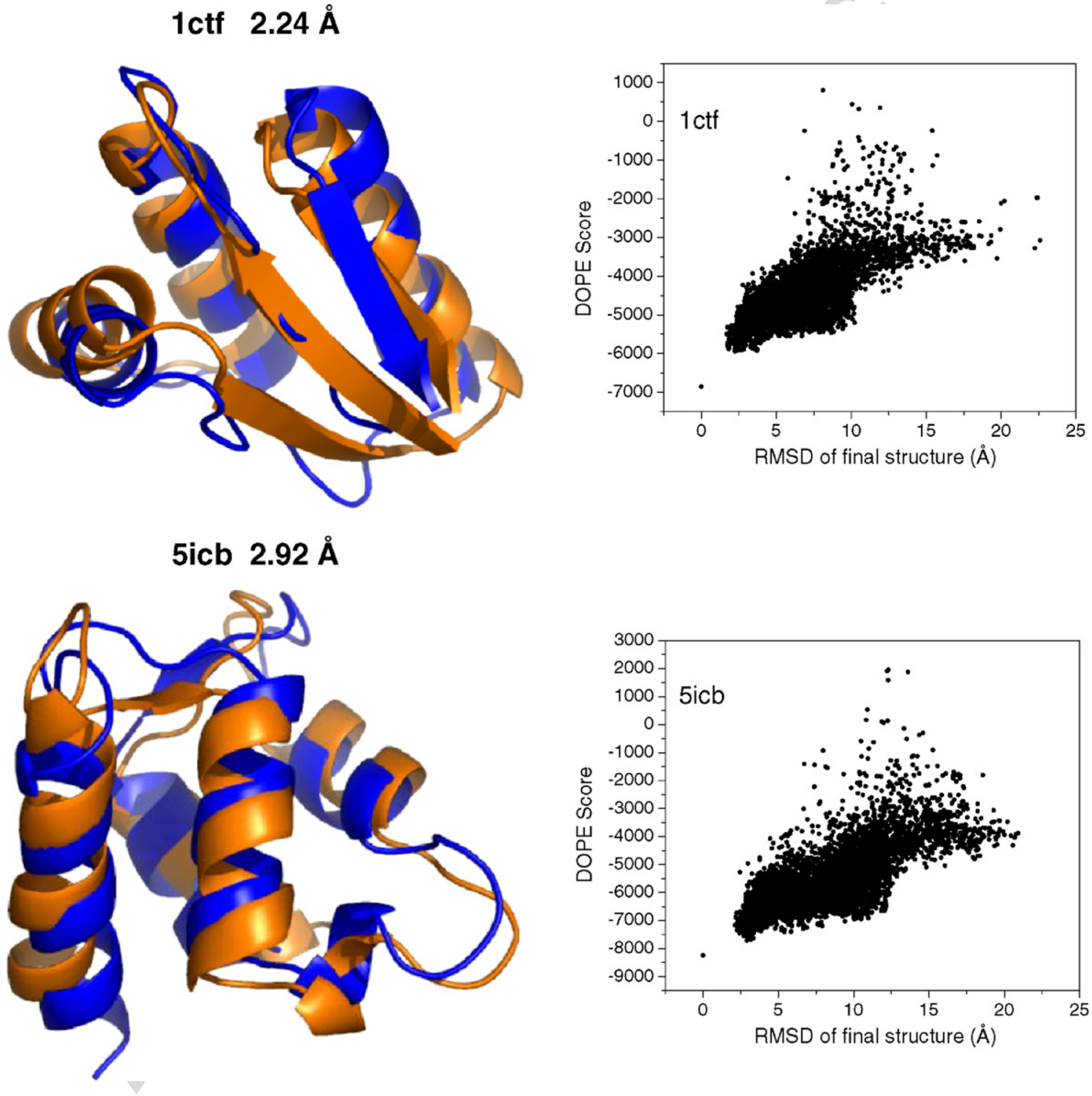


Figure 7. Lowest energy structures. 3D X-ray-traced renderings and corresponding scatter plots of the RMSDs and DOPE scores from the simulations using DOPE-C^{β} for six of the 56 proteins simulated in this study. The predicted (lowest energy, blue) structures are displayed for six representative proteins spanning the range of low to high RMSD values from the native structure (orange). In the scatter plots, the *X* and *Y* coordinates of each point correspond, respectively, to the RMSD between the simulated and the native structures and to the DOPE score for the computed structure. The native structure is also included in the scatter plots as the leftmost star-shaped point with RMSD=0 Å. In 68% of the energy RMSD scatter plots, the native structure corresponds to the conformation with lowest energy. The renderings have been generated with PyMol [http://www.pymol.org/ (DeLano, W.L. (2002). The PyMOL Molecular Graphics System, DeLano Scientific, San Carlos, CA, USA].

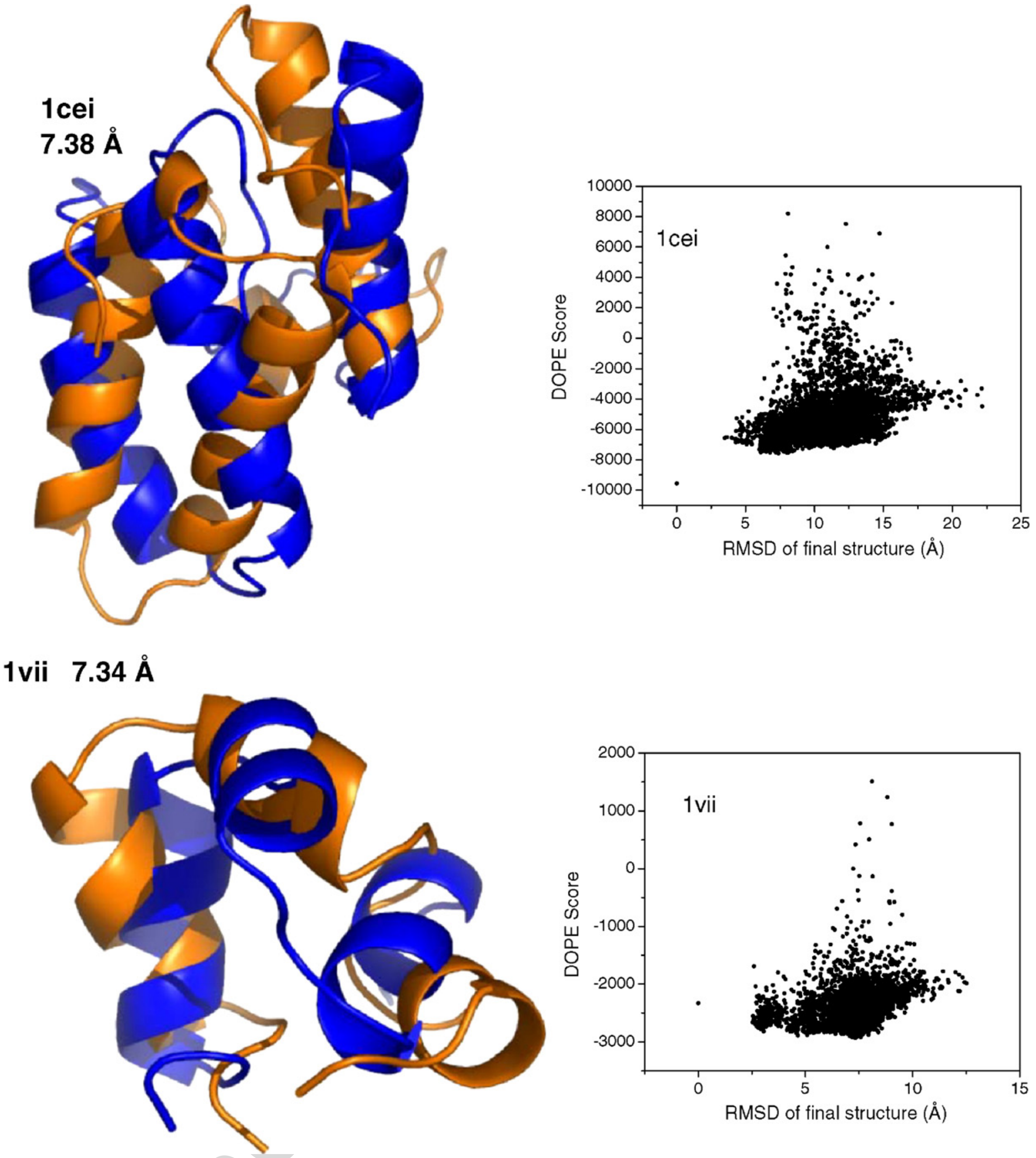


Figure 7 (*legend on previous page*)

generating coarse-grained folding pathways that capture features of real pathways, such as their diversity and transition states structures. Algorithmically, the major similarity is that ϕ,ψ moves are constrained by RBs. In their study, the ϕ,ψ values are obtained by uniformly sampling from one of the four Ramachandran quadrants. The present basin definitions are more accurate (e.g. separating β and PPII conformers and including turns in the helical basin, rather than with the β and PPII conformers). Also, we sample dihedral angles from a library obtained from PDB structures according to residue type and with the inclusion of nearest neighbor effects. In addition, the scoring function is different. The earlier work uses a semi-empirical energy function that stresses the importance of three-body interactions,[42] rather than a residue-dependent statistical potential in the present treatment. Previously, a kinetic Monte Carlo routine was used, while a SA is now employed to minimize the statistical potential. Using the earlier algorithm, Colubri predicts native structures with some success by incorporating independent secondary structure knowledge.[30]

Free energy surfaces

The study by Gong *et al.* stresses that the use of native-like pentamers, hydrogen bonding, steric

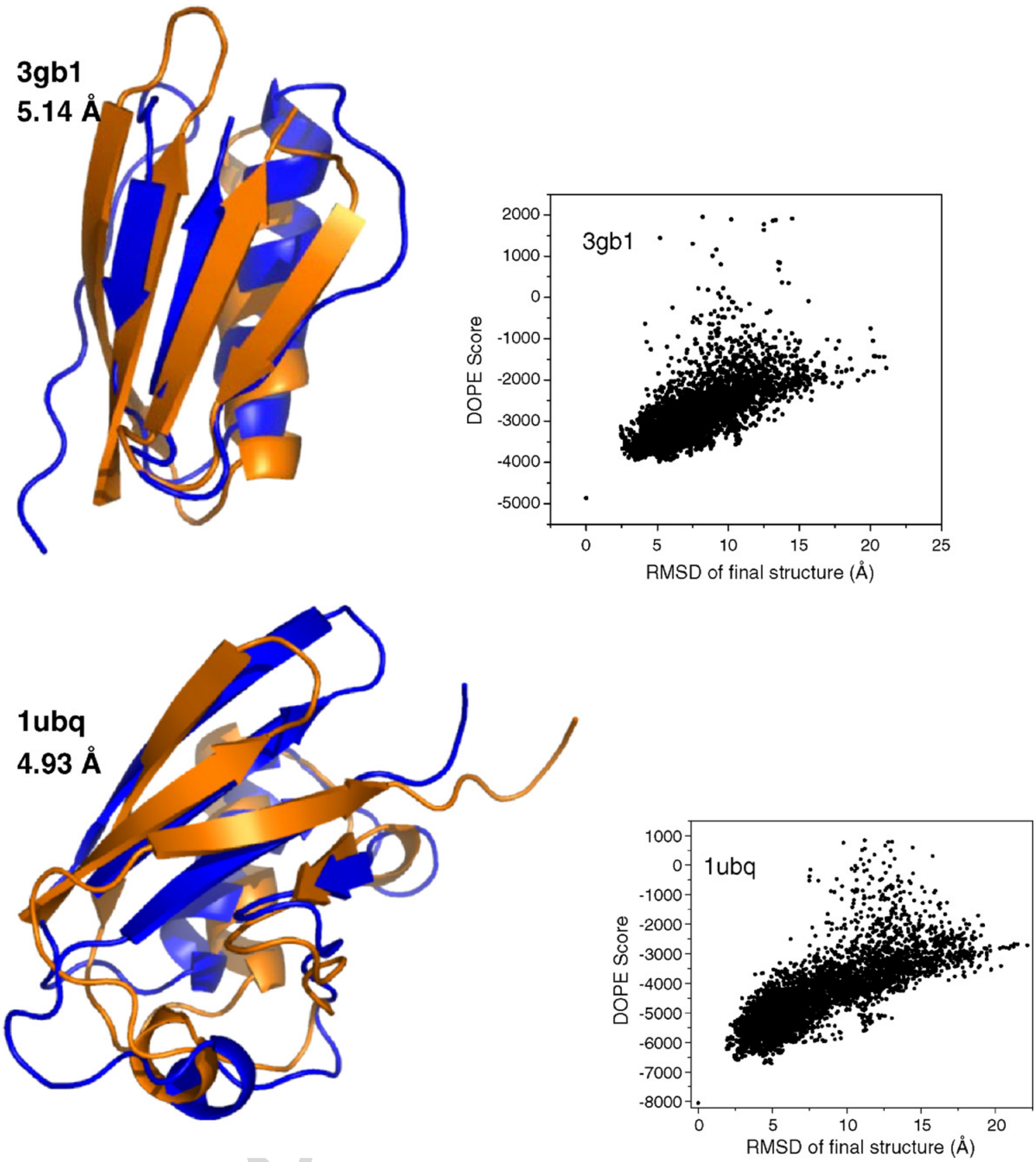


Figure 7 (*legend on page 846*)

repulsion and compaction, generates only a "remarkably small number of topologically distinct clusters".[34] A recent study by Takada *et al.*[20] similarly finds that the local biases, encoded by the use of fragments obtained from the PDB, are strong enough that even non-specific collapse often produces native-like conformations and funnel-like landscapes for many small proteins. Takada *et al.* conclude that these properties rationalize the success of fragment insertion methods in *ab initio* structure predictions. Their reasoning also rationalizes why the inclusion of NN effects, which is inherent to fragment insertion, has a greater impact on our predictions than the inclusion of the C^{β} terms in the statistical potential.

However, several studies, including ours, Takada *et al.*[7] and Gong *et al.*,[34] obtain plots of RMSD *versus* energy containing many low energy, non-native minima (RMSD>5 Å). Hence, the accompanying free energy surface is quite rugged when a single folding trajectory is considered, even if the RMSD for multiple trajectories correlates well with the energy. Native local biases and collapse often are insufficient to uniquely define the native fold. In order to identify the native folds, clustering or other methods are required. In contrast, real proteins readily fold to the native state in a two-state manner,[43,44] and hence, do not traverse the complicated landscapes observed in many simulations. This difference suggests that improvements in energy functions or better move sets are required for a more realistic description of the folding process.

Table 3. Results for SA runs using DOPER[1]

PDB code	Class, N_{res}	Source	Native lowest energy	Predict RMSD (all-atom)[a]	Predict RMSD (C^{β})[b]	Lowest RMSD[c]	RMSD < 2 Å (%)	RMSD < 4 Å (%)	RMSD < 6 Å (%)	RMSD < 8 Å (%)	Min RMSD in top5[d]	Max RMSD in top5	Ave RMSD in top5
1bdc	α,60	NMR	No	7.44	7.44	5.96	0	0	2	67	6.37	8.00	7.07
1bw6	α,56	NMR	Yes	6.62	6.62	3.06	0	7	29	39	3.41	6.62	4.57
1bxy	αβ,60	X-ray	Yes	2.97	**2.69**	1.68	1	33	46	62	2.67	4.27	3.11
1ctf	αβ,67	X-ray	Yes	3.05	3.05	2.36	0	33	67	92	2.36	3.77	3.20
1kjs	α,74	NMR	Yes	3.59	7.34	3.59	0	1	28	69	3.59	7.34	5.89
1msi	β,60	X-ray	Yes	4.99	4.99	4.08	0	0	6	44	4.08	10.82	7.18
1mzm	α,71	X-ray	Yes	5.57	10.08	2.93	0	24	51	73	3.21	8.83	5.95
1orc	αβ,56	X-ray	Yes	7.35	8.15	2.68	0	14	72	88	3.65	8.15	5.60
1ubq	αβ,76	X-ray	Yes	3.41	**2.97**	2.92	0	29	79	95	3.02	4.26	3.46
2pdd	α,43	NMR	No	8.39	**7.67**	3.47	0	2	35	66	5.50	8.58	7.68
1res	α,35	NMR	No	1.13	1.13	1.13	77	99	100	100	1.13	1.56	1.29
1vii	α,36	NMR	No	3.32	7.95	3.01	0	6	19	95	3.32	7.17	5.84
1uxd	α,43	NMR	Yes	2.95	2.95	2.19	0	13	19	69	2.66	3.92	3.11
1uba	α,45	NMR	Yes	5.06	5.06	4.57	0	0	78	96	5.06	5.59	5.42
1gab	α,47	NMR	No	2.08	**1.72**	1.72	16	84	99	100	1.72	2.08	1.95
1prb	α,53	NMR	No	2.81	2.81	2.67	0	24	61	91	2.81	4.01	3.22
1enh	α,54	X-ray	Yes	3.24	**2.49**	2.42	0	96	100	100	2.45	3.98	3.37
1am3	α,57	X-ray	Yes	3.40	**2.93**	2.71	0	50	91	98	2.93	6.79	4.11
1r69	α,61	X-ray	Yes	3.27	4.59	2.27	0	12	92	99	3.27	5.08	4.30
1utg	α,62	X-ray	Yes	5.40	5.57	4.54	0	0	75	99	5.38	5.67	5.48
2ezh	α,65	NMR	Yes	5.18	**4.78**	4.42	0	0	45	79	5.18	7.60	6.48
1a32	α,65	X-ray	Yes	6.72	**5.53**	3.37	0	3	35	72	5.53	6.72	6.20
1nre	α,66	NMR	Yes	7.59	**3.67**	2.64	0	9	19	41	3.67	7.59	5.88
1ail	α,67	X-ray	Yes	6.99	**4.65**	4.26	0	0	68	84	4.61	9.86	6.29
1lfb	α,69	X-ray	Yes	4.27	**3.98**	2.84	0	27	53	75	2.84	4.27	3.47
1nkl	α,70	NMR	Yes	4.11	9.01	3.05	0	5	16	59	4.11	9.24	7.94
1pou	α,70	NMR	Yes	9.01	**7.19**	3.37	0	3	11	30	4.89	10.72	8.05
5icb	α,72	X-ray	Yes	4.21	**3.43**	2.72	0	30	50	60	2.72	8.50	4.59
1hyp	α,75	X-ray	Yes	4.94	4.94	3.93	0	1	26	73	3.93	7.30	5.74
1cc5	α,76	X-ray	Yes	9.59	**7.56**	5.03	0	0	3	36	7.34	9.79	8.71
1cei	α,85	X-ray	Yes	11.49	**6.53**	5.06	0	0	5	23	6.02	11.49	7.89
1ptq	αβ,43	X-ray	Yes	8.83	**2.70**	1.96	1	39	58	78	1.96	8.83	5.59
3gb1	αβ,56	NMR	Yes	4.21	4.95	2.40	0	28	71	90	4.21	7.16	5.29
1aa3	αβ,56	NMR	Yes	6.23	7.06	5.00	0	0	25	87	5.92	6.47	6.16
1pgx	αβ,57	X-ray	Yes	2.64	**2.79**	2.05	0	45	86	95	2.64	4.59	3.38
1tif	αβ,59	X-ray	Yes	6.06	**2.80**	2.66	0	23	51	85	2.80	6.06	4.11
2ptl	αβ,60	NMR	Yes	6.98	7.63	4.41	0	0	6	55	6.98	9.17	7.91
1dol	αβ,62	X-ray	Yes	7.27	9.33	4.08	0	0	11	54	5.61	9.33	7.88
2fow	αβ,66	NMR	Yes	5.11	7.46	4.54	0	0	4	32	5.07	9.91	7.50
1afi	αβ,72	NMR	Yes	10.17	10.17	6.43	0	0	0	14	8.05	10.39	9.48
1vcc	αβ,77	X-ray	Yes	10.36	10.36	6.30	0	0	0	26	7.97	11.31	9.41
2fxb	αβ,81	X-ray	Yes	8.22	8.22	4.34	0	0	12	32	4.65	11.32	7.95
1e0l	β,37	NMR	Yes	7.90	**4.83**	4.44	0	0	39	95	4.67	7.90	5.99
1vif	β,48	X-ray	Yes	2.26	3.39	2.26	0	17	65	93	2.26	8.32	4.67
1bq9	β,53	X-ray	Yes	4.79	**3.95**	2.50	0	27	50	83	3.62	4.79	4.02
5pti	β,55	X-ray	Yes	6.73	**4.88**	2.76	0	15	60	86	3.78	7.14	5.17
1tuc	β,61	X-ray	Yes	4.71	**4.57**	3.05	0	1	17	42	4.71	9.14	8.01
1csp	β,64	X-ray	Yes	7.93	7.93	3.85	0	1	14	51	5.57	10.65	7.53
1sro	β,66	NMR	Yes	4.54	6.17	4.51	0	0	10	46	4.54	9.72	7.41
1g6p	β,66	NMR	Yes	10.00	**8.96**	6.62	0	0	0	9	7.94	10.79	9.95

Structures are determined using the SA algorithm using the enhanced C^{β}-only DOPER scoring function with sampling that includes nearest neighbor effects. Final energy rankings are determined for these structures after inserting side-chains using SCWRL, along with the all-atom DOPE statistical potential. C^{α} RMSD structures are relative to the native structure in Å.

[a] C^{α} RMSD for lowest energy structure relative to the native structure in Å (using DOPE score) among the final structures from 100 trajectories.

[b] C^{α} RMSD for lowest energy structure relative to the native structure in Å (using DOPE-C^{β} score) among the final structures from 100 trajectories.

[c] Lowest RMSD for all the structures among the final structures from 100 trajectories.

[d] Minimum and maximum RMSD between the native and the five lowest energy structures (using DOPE score) among the final structures from 100 trajectories.

Future improvements

The previous sections indicate three areas where the algorithm might be substantially improved: a richer rotamer library, a better cooling schedule, and a more detailed energy function. The rotamer library could be enriched by two different approaches. The first approach would be to include more structures in our training set, while the second would be to construct a library of trimer conformations using dimer information already available in the library. Another interesting possibility is to generate

Author's personal copy

Table 4. Results for SA runs using DOPE-BB

PDB code	Class, N_{res}	Source	Native lowest energy	Predict RMSD[a]	Predict RMSD[b]	Lowest RMSD[c]	RMSD < 2 Å (%)	RMSD < 4 Å (%)	RMSD < 6 Å (%)	RMSD < 8 Å (%)	Min RMSD in top5[d]	Max RMSD in top5	Ave RMSD in top5
1bdc	α,60	NMR	No	6.96	7.50	4.18	0	0	21	94	5.16	7.50	6.68
1bw6	α,56	NMR	No	5.51	5.51	4.63	0	0	40	66	5.51	10.06	8.06
1bxy	αβ,60	X-ray	Yes	2.83	2.83	1.90	1	24	41	49	1.90	2.98	2.56
1ctf	αβ,67	X-ray	Yes	5.00	5.00	4.35	0	0	37	69	4.41	5.75	5.10
1kjs	α,74	NMR	No	5.48	5.48	3.7	0	2	33	68	5.48	6.97	6.34
1msi	β,60	X-ray	Yes	8.30	8.30	8.17	0	0	0	0	8.30	10.73	9.64
1mzm	α,71	X-ray	Yes	8.65	7.06	5.62	0	0	1	46	7.06	8.65	7.43
1orc	αβ,56	X-ray	Yes	8.54	8.54	7.91	0	0	0	4	8.29	8.85	8.50
1ubq	αβ,76	X-ray	Yes	3.74	3.46	2.67	0	22	77	94	3.20	3.74	3.51
2pdd	α,43	NMR	No	5.03	5.03	3.93	0	3	42	70	5.03	8.83	7.30

Structures are determined using the SA algorithm, the DOPE-BB scoring function and nearest neighbor effect in sampling. Unless specified otherwise, final energy rankings are determined for these structures after inserting side-chains using SCWRL, using the all-atom DOPE statistical potential. C^{α} RMSD structures are relative to the native structure in Å.

[a] C^{α} RMSD for lowest energy structure relative to the native structure in Å (using DOPE score) among the final structures from 100 trajectories.

[b] C^{α} RMSD for lowest energy structure relative to the native structure in Å (using DOPE-C^{β} score) among the final structures from 100 trajectories.

[c] Lowest RMSD for all the structures among the final structures from 100 trajectories.

[d] Minimum and maximum RMSD between the native and the five lowest energy structures (using DOPE score) among the final structures from 100 trajectories.

"synthetic" rotamers by means of MD or LD simulations of short peptides.[45]

The SA minimization could be improved by implementing an adaptive cooling schedule, where the chain length dictates the number of steps at constant annealing temperature, to ensure reaching thermal equilibrium. As for improving the energy function, we are currently considering adding an orientation-dependent backbone hydrogen bond term.

All these issues are interconnected. For example, even with the addition of an explicit hydrogen bond term, if the search is not sufficiently precise to attain the resolution needed to form the correct hydrogen bonding pattern, the final structure might not necessarily be improved. A more precise search might require a gradient minimization to relax the dihedral angles in smaller increments than that possible using the PDB-based library. This later enhancement of the algorithm could use splines[46] or another interpolation method to construct a differentiable function from DOPE-C^{β}, which currently is represented by linearly interpolation from grid values and, thus, is unsuitable for gradient minimization.

Future perspectives

Our long-term goal is to develop a full *ab initio* algorithm for structure prediction. This paper

Author's personal copy

Table 5. Results for SA runs using DOPE-C^{β} with without NN effects

PDB code	Class, N_{res}	Source	Native lowest energy	Predict RMSD[a]	Predict RMSD[b]	Lowest RMSD[c]	RMSD < 2 Å (%)	RMSD < 4 Å (%)	RMSD < 6 Å (%)	RMSD < 8 Å (%)	Min RMSD in top5[d]	Max RMSD in top5	Ave RMSD in top5
1bdc	α,60	NMR	No	7.49	7.41	4.52	0	0	7	84	7.01	7.57	735
1bw6	α,56	NMR	No	6.21	5.55	4.08	0	0	39	67	6.21	7.97	6.94
1bxy	αβ,60	X-ray	Yes	8.75	5.22	3.31	0	7	27	43	3.77	8.75	5.80
1ctf	αβ,67	X-ray	Yes	5.92	5.79	3.90	0	2	53	85	4.98	7.31	6.13
1kjs	α,74	NMR	No	10.55	4.73	4.38	0	0	46	68	5.57	10.55	6.78
1msi	β,60	X-ray	Yes	12.18	9.53	6.61	0	0	0	5	7.99	12.18	9.37
1mzm	α,71	X-ray	Yes	5.85	8.27	5.85	0	0	8	25	5.85	5.85	5.85
1orc	αβ,56	X-ray	Yes	7.91	7.91	7.13	0	0	0	28	7.91	7.91	7.91
1ubq	αβ,76	X-ray	Yes	4.41	3.78	3.78	0	6	41	46	4.41	4.54	4.43
2pdd	α,43	NMR	No	4.58	8.81	4.58	0	0	6	53	4.58	4.58	4.58

Structures are determined using the SA algorithm and the DOPE-C^{β} scoring function but with sampling that does not include nearest neighbor effects. Unless specified otherwise, final energy rankings are determined for these structures after inserting side-chains using SCWRL, using the all-atom DOPE statistical potential. C^{α} RMSD structures are relative to the native structure in Å.

[a] C^{α} RMSD for lowest energy structure relative to the native structure in Å (using DOPE score) among the final structures from 100 trajectories.

[b] C^{α} RMSD for lowest energy structure relative to the native structure in Å (using DOPE-C^{β} score) among the final structures from 100 trajectories.

[c] Lowest RMSD for all the structures among the final structures from 100 trajectories.

[d] Minimum and maximum RMSD between the native and the five lowest energy structures (using DOPE score) among the final structures from 100 trajectories.

describes an approach that can determine folded protein structures with an acceptable accuracy if the native RBs are given, a necessary but not sufficient hurdle for any folding algorithm. Thus, the ultimate goal requires the development and implementation of an efficient method to search through different basins and to couple this search with the intra-basin minimization module tested here. In this regard, the intra-basin minimization algorithm could serve as the “structure generation” module for a more physically grounded algorithm, where the folding kinetics are simulated at the level of the RB transitions.[29,30]

Another important application of the intra-basin minimization routines involves using as input estimates of the ϕ,ψ angles from easily generated NMR measurements (e.g. *J*-couplings and chemical shifts)[47] rather than native basin assignments. Then, the minimization algorithm could help in finding the three-dimensional protein structure using these estimates as constraints.

We are also interested in using this model as a test bed for comparing statistically based approaches with first principle ones. For instance, the reduced scoring function could be recalculated using a physically based energy function, such as one containing explicit electrostatic and non-polar contributions. A comparison between the purely knowledge based and the more physically based approaches might reveal insights about the weaknesses and strengths of each of them.

ersonal copy

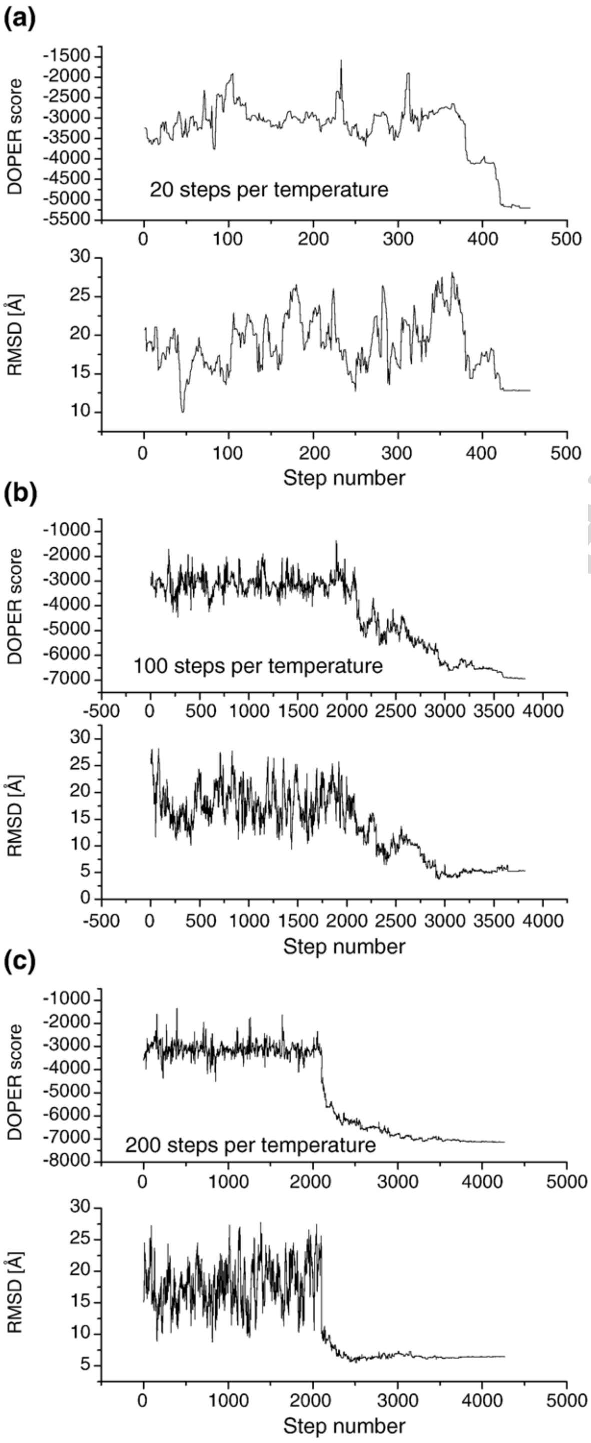


Figure 8. Different annealing schedules. The evolution of the DOPER scoring function (top) and the native RMSD (bottom) along the intra-basin annealing pathway for 1ubq. In (a), 20 steps are performed at each annealing temperature; in (b), 100 steps; and in (c), 200 steps. A low energy conformation is obtained in all cases, and convergence is attained. However, a substantial decrease in the RMSD occurs only when using 100 and 200 steps. The simulations presented here are generated using 20 steps per temperature. Hence, the fact that results in good agreement with the native structure are generated demonstrates that using only 20 steps can also lead to structures with small RMSD values, but it appears that the number of failures, as the one displayed in (a), also increases. The advantage of using 20 steps over using 100 or a higher number of steps is that the simulations take fivefold less computer time. The first ~2000 steps are used to determine the initial “temperature” t for the subsequent annealing portion of the simulation.

Methods

Reduced protein representation and Ramachandran basins

The only atoms explicitly included in the simulated annealing simulations are the heavy atoms of the main-chain, the nitrogen (N), α carbon (C^{α}), carbonyl carbon (C) and oxygen (O) atoms, together with the β carbons (C^{β}) of the side-chains for all amino acids except glycine. The backbone planar angles and bond lengths are held fixed at their mean values, so that the only variables considered are the main-chain dihedral angles ϕ, ψ and ω. Since ω occurs in the *trans* conformation for the great majority of the protein structures, ω is also chosen as fixed at 180° during the simulations. Initial structures retain ω in the *cis* conformation ($\omega = 0$) for residues with such geometries in the native conformation.

Our model permits each amino acid to reside in one of five RBs, called the β, poly-proline II (PPII), right handed α (α_R), left handed α (α_L), and ε basins (Figure 1). The specification of these basins for every residue in the protein codifies a substantial amount of information

Table 6. Minimization results as a function of the number of annealing steps

Number of steps[a]	$\text{RMSD}_{\text{min,all}}$ (Å)	$N_{\leq 4.0}$[b]
20	3.0	36
100	2.9	68
200	2.7	96
300	2.5	90

These results correspond to 1000 minimization runs for the protein 1UBQ using the DOPER scoring function.

[a] Number of annealing steps per constant temperature value.

[b] Number of final structures with a backbone RMSD to the native of less or equal than 4.0 Å.

regarding local chain conformations. For example, stretches of more than four consecutive residues in the α_R basin are usually found in α−helices. Likewise, stretches of residues in the β basin indicate the location of beta strands. Turns and bends are also associated with specific RB patterns.

Training set

The scoring function and rotamer library are parameterized by constructing a training set consisting of high quality resolved X-ray and NMR structures with low sequence similarity. The stand alone version of the Pisces culling server is used for this purpose.[48] The training set includes X-ray structures with 2.2 Å maximum resolution and *R*-factor of 0.3, appended with NMR structures, both with less than 30% mutual sequence identity. This selection of culling parameters yields 4701 unique structures. Although the 50 target structures are not explicitly excluded from the library, only two of the 3035 native dihedral angle pairs are found in the predicted structures. This result indicates that the algorithm undergoes a meaningful search within the basins, and our successes are not due to the insertion the native fragments, which could trivialize our successes.

DOPE, statistical potential

The discrete optimized protein energy function[27,28] is a distance-dependent statistical potential that is similar to DFIRE,[26] except for the reference state. DOPE is based on an improved reference state that corresponds to non-interacting atoms in a homogeneous sphere with the radius dependent on the sample native structure. The DOPE potential thus accounts for the finite and spherical shape of the native structures. DOPE and five other scoring functions were tested by the detection of the native state among six multiple target decoy sets,[25] the correlation between the score and the model error, and the identification of the most accurate non-native structure in the decoy set. For all decoy sets, DOPE was the best performing function in terms of all criteria. Moreover, in another comparative study of 23 different physics-based energy functions, statistical potentials, and machine learning-based scoring functions, the DOPE statistical potential was one of the best at identifying non-native decoys.[28] Our DOPE-C^β statistical potential includes interactions involving only the backbone heavy atoms and the C^β. The interaction parameters for these atoms are the same as in DOPE, while those involving the side-chain atoms other than the C^β are set to zero. Similarly, the DOPE-BB potential includes interactions involving the backbone heavy atoms only.

DOPER, side chain average of DOPE

We also test DOPER, a modified version of DOPE-C^β in which the C^β interactions are modified to include the average effects of the inter-residue interactions between all heavy atoms in the side-chains. To determine the energy score for the interactions involving the C^βs, the DOPE score is averaged over the side-chain–side-chain and side-chain–main-chain interactions for all side-chain conformations that present the same C^β–C^β interatomic separation. Equation (2) summarizes the averaging procedure for the C^β–C^β interactions:

$$E_{a_1,a_2}(r) = \left\langle \sum \bar{u}_{ij}(r_{ij}) \right\rangle_{SC(a_1,a_2,r)} \qquad (2)$$

where $E_{a1,a2}(r)$ represents the average interaction energy between two C^βs that are separated by the distance *r*, one on amino acid a_1 and the other on amino acid a_2, and $\bar{u}_{ij}$ (r_{ij}) stands for the DOPE energy score between atoms *i* and *j* when separated by the distance $r = r_{ij}$. Given the side-chain pair (a_1,a_2), the set $P(a_1,a_2)$ is formed by all the atom pairs in which the first atom belongs to the side-chain of residue a_1 and the second to the side-chain of

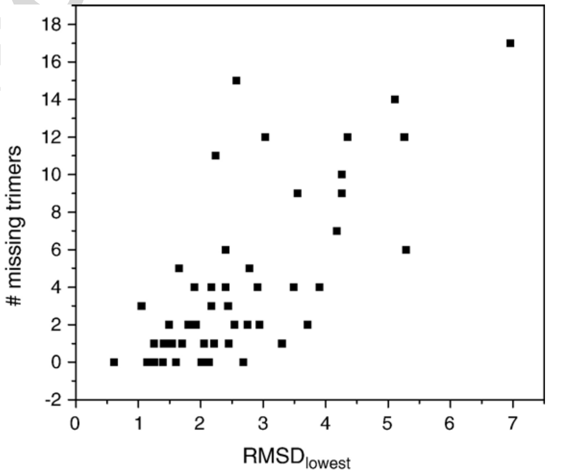


Figure 9. Improvement in RMSD with trimer coverage. The RMSD for the most native-like structure among all conformations is depicted for each protein versus the number of amino acids whose basin trimer assignments have no compatible ϕ,ψ coordinates within the rotamer library. The correlation coefficient is 0.74, indicating that the dominant source of failures of our folding algorithm likely arises from the lack in the rotamer library of a sufficient set of compatible trimer conformations for the given protein. Moreover, when the rotamer library contains enough compatible trimers to span the proteins sequence (e.g. less than ten trimers with no entries in the rotamer library), the program generates quite reasonable predictions in most cases: for 93% of the proteins with less than ten trimers lacking entries in the library, the best model among the five lowest DOPE score structures is less than 4 Å RMSD from the native conformation. As expected, the proteins from the control set always have at least one compatible trimer, the X-ray structures have enough trimers in most cases, while the NMR structures repeatedly fail to be covered well by the rotamer library.

Author's personal copy

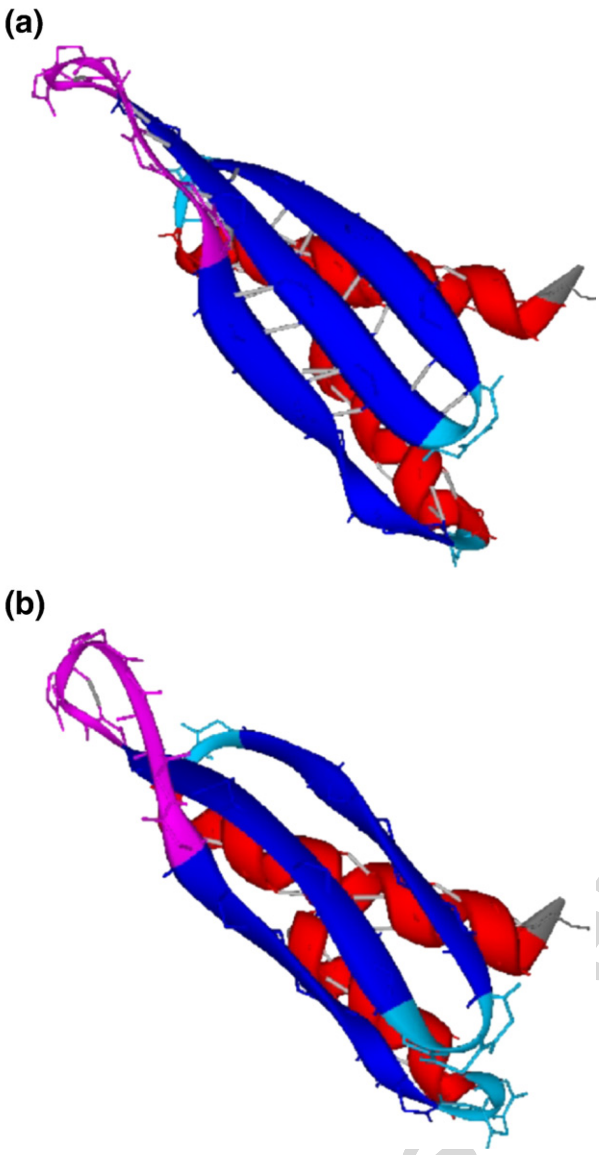


Figure 10. Hydrogen bonds. (a) The native structure of 1di2 with the backbone hydrogen bonds. (b) The backbone hydrogen bonds for a simulated model with 2.5 Å RMSD from the native. It is apparent from this comparison that even though the simulated structure has the right topology, it lacks most of the hydrogen bonding pattern present in the native structure. This might be due to the absence of an explicit orientationally dependent hydrogen bond term in the scoring function used for the simulations. The 3D renderings have been generated with YAPView [http://protlib.uchicago.edu/dloads.html].

residue a_2. The average is calculated over all occurrences of the pair of amino acids a_1 and a_2, whose C^{β}s are separated by the distance r, within a single protein in the training set, a collection denoted by the set $SC(a_1,a_2,r)$. By using PDB structures, we weight the contribution from each side-chain conformation according to the frequency with which the conformation appears in the native structures of the training set. Similar equations are used to calculate the C^{β}–N, C^{β}– C^{α}, C^{β}–C and C^{β}–O effective interaction terms.

HOPE-C^{β} statistical potential with hydrogen bond term

We also include preliminary tests for an extension of the statistical potential DOPE-C^{β}, which has an additional term corresponding to interaction of the amide hydrogen atoms with all the other atoms in the protein. These parameters are obtained in a similar fashion to that for any other pair of atoms described earlier for DOPE. The program REDUCE[49] is used to add amide H to the PDB structures in our training set. This version of the statistical potential (HOPE) is only used in comparison of our model with the recent work of Rose and co-workers.[34]

Sampling method

The model has been implemented computationally using a low level C++ library, called the Protein Library (PL), which provides the basic routines for handling protein structures and motions. The program that integrates all the components of the model and that performs the constrained SA minimization of the statistical potential is named OOPS (OOPS is an Open Protein Simulator). The word "open" implies that the code is licensed as GPL and that the program is architecturally open to different methodologies and strategies for simulating protein folding. The PL and OOPS can be downloaded from the webpage‡.

The sampling routine consists of a discrete SA algorithm in which the (ϕ, ψ, ω) dihedral angles are changed in groups of one, two, or three consecutive residues at a time. The move set is constrained to the set of (ϕ, ψ, ω) dihedral angles contained in a rotamer library for the particular backbone fragments. This rotamer library is constructed (as described below) using all the occurrences of particular configurations of one, two, or three consecutive residues that are found in the native structures of our training set. Throughout the entire annealing run, all residues are constrained to remain within their pre-assigned RBs.

Backbone rotamer library

We have constructed a library of monomer, dimer, and trimer backbone (ϕ, ψ) rotamers. The dimer and trimer libraries correlate the amino acid sequence for pairs and triples, respectively, of sequential residues with the specific backbone RB occupancies, thereby incorporating local correlations that arise from all-atom nearest neighbor interactions. Hence, for every monomer, dimer, and trimer fragment contained in the training set, the corresponding amino acids, RBs and dihedral angles are computed and stored in a database that indexes the rotamer dihedral angles by the amino acid sequence and their basin occupancies. The library provides the set of available dihedral angles for a monomer, dimer, or trimer peptide fragment given a specification of the residue amino acids' identity and basin occupancies.

Clustering

For the comparison with Rose and co-workers, structures are clustered using the program Cluster

‡ http://protlib.uchicago.edu/index.html

Author's personal copy

Table 7. Results for SA runs using DOPE-C$^{\beta}$ for proteins studied by Gong *et al.*

PDB code	Class, N_{res}	Source	Native lowest energy	Predict RMSD[a]	Predict RMSD[b]	Lowest RMSD[c]	RMSD < 2 Å (%)	RMSD < 4 Å (%)	RMSD < 6 Å (%)	RMSD < 8 Å (%)	Min RMSD in top5[d]	Max RMSD in top5	Ave RMSD in top5
2gb1	αβ,56	NMR	Yes	8.93	4.61	2.87	0	6	82	94	3.16	8.93	5.05
1ubq	αβ,76	X-ray	Yes	3.18	3.16	1.93	1	39	93	98	3.18	4.58	3.58
1c90a	β,66	X-ray	Yes	2.96	3.72	2.56	0	13	49	84	2.56	5.91	3.80
1ifb	β,131	X-ray	Yes	14.79	16.16	10.52	0	0	0	0	13.97	15.73	14.91
1vii	α,36	NMR	No	7.60	7.52	2.68	0	4	15	95	5.33	7.66	7.12
1r69	α,63	X-ray	Yes	3.43	4.74	2.38	0	39	83	92	2.76	3.43	3.10

Structures are determined using the SA algorithm and the DOPE-C$^{\beta}$ scoring function but with sampling that does not include nearest neighbor effects. Unless specified otherwise, final energy rankings are determined for these structures after inserting side-chains with SCWRL, using the all-atom DOPE statistical potential. C$^{\alpha}$ RMSD structures are relative to the native structure in Å.

[a] C$^{\alpha}$ RMSD for lowest energy structure relative to the native structure in Å (using DOPE score) among the final structures from 100 trajectories. Values in parenthesis are from Gong *et al.*[34]

[b] C$^{\alpha}$ RMSD for lowest energy structure relative to the native structure in Å (using DOPE-C$^{\beta}$ score) among the final structures from 100 trajectories.

[c] Lowest RMSD for all the structures among the final structures from 100 trajectories.

[d] Minimum and maximum RMSD between the native and the five lowest energy structures (using DOPE score) among the final structures from 100 trajectories.

3.0[39] Final structures from each trajectory are chosen and are clustered based on their C$^{\alpha}$ atom RMSD to each other. The lowest average energy (based on the all-atom DOPE scoring function) cluster is selected from a group of clusters that have a correlation coefficient of ≥95% between the structures comprising the clusters.

Simulated annealing minimization

The SA routine is based on the algorithm described by Aarts & Korst.[50] It begins with a random assignment of dihedral angles within the native basin and converges to the global minimum by gradually diminishing the annealing temperature. This annealing temperature controls the fraction of transitions that increase the energy of the system, and the initial annealing temperature is chosen so that half of the transitions at the beginning of the minimization result in increasing energy. For each annealing temperature, a fixed number of elementary conformational transitions are computed in order to achieve thermal equilibrium. The annealing temperature is decreased according to a Cauchy cooling schedule, until convergence is reached. The temperature update formula is:

$$t = \frac{t'}{1 + \dfrac{t'\log\delta}{3\sigma_{t'}}} \tag{3}$$

where t' and t are the old and new annealing temperatures, respectively, $\sigma_{t'}$ is the standard deviation for the energy distribution at temperature t', and δ is a tunable parameter for the cooling schedule. Lower case t is used for the annealing temperature to distinguish from the physical temperature T. The convergence criterion is based on the magnitude of the energy fluctuations for each annealing temperature, and annealing stops when the inequality:

$$\sigma_t \le \varepsilon t \tag{4}$$

is satisfied, where ε is the convergence tolerance, also a tunable parameter.

An elementary move in this algorithm begins by randomly choosing a residue along the chain. The amino acid and basin of the three consecutive residues, starting at the N terminus of the selected residue, are recorded, and a

Table 8. Results for SA runs using HOPE-C$^{\beta}$ for proteins studied by Gong *et al.*

PDB code	Class, N_{res}	Source	Native lowest energy	Predict RMSD[a]	Predict RMSD[b]	Lowest RMSD[c]	RMSD < 2 Å (%)	RMSD < 4 Å (%)	RMSD < 6 Å (%)	RMSD < 8 Å (%)	Min RMSD in top5[d]	Max RMSD in top5	Ave RMSD in top5
2gb1	αβ,56	NMR	Yes	4.04	3.24	3.24	0.00	10.00	86.00	93.00	3.24	5.14	4.22
1ubq	αβ,76	X-ray	Yes	1.57	3.27	1.57	1.00	46.00	87.00	96.00	1.57	5.25	3.27
1c90a	β,66	X-ray	Yes	3.43	3.68	3.16	0.00	11.00	47.00	85.00	3.43	4.71	3.96
1ifb	β,131	X-ray	Yes	10.24	13.80	10.24	0.00	0.00	0.00	0.00	10.24	14.52	13.15
1vii	α,36	NMR	No	7.38	7.38	2.58	0.00	4.58	10.75	80.73	7.17	7.39	7.28
1r69	α,63	X-ray	Yes	5.42	5.22	2.16	0.00	28.00	83.00	97.00	2.51	5.33	3.56

Structures are determined using the SA algorithm, the HOPE-C$^{\beta}$ scoring function, and sampling that does include nearest neighbor effects. Unless specified otherwise, final energy rankings are determined for these structures after inserting side-chains with SCWRL, using the all-atom DOPE statistical potential. C$^{\alpha}$ RMSD structures are relative to the native structure in Å.

[a] C$^{\alpha}$ RMSD for lowest energy structure relative to the native structure in Å (using HOPE score) among the final structures from 100 trajectories.

[b] C$^{\alpha}$ RMSD for lowest energy structure relative to the native structure in Å (using HOPE-C$^{\beta}$ score) among the final structures from 100 trajectories.

[c] Lowest RMSD for all the structures among the final structures from 100 trajectories.

[d] Minimum and maximum RMSD between the native and the five lowest energy structures (using HOPE score) among the final structures from 100 trajectories.

Authors personal copy

Table 9. Comparison of the results using DOPE-C^{β} with that of Gong *et al*.

PDB code	Class,N_{res}	Source	Native lowest energy	Gong Predict RMSD[a]	HOPE Predict RMSD[b]	DOPE Predict RMSD[c]
2gb1	αβ,56	NMR	Yes	1.11	3.24	4.19
1ubq	αβ,76	X-ray	Yes	1.81	1.61	3.18
1c90a	β,66	X-ray	Yes	1.38	3.74	2.96
1ifb	β,131	X-ray	Yes	3.05	11.58	11.41
1vii	α,36	NMR	No	3.78	7.19	7.54
1r69	α,63	X-ray	Yes	4.99	2.89	3.43

[a] Result of Rose and co-workers.[34]
[b] Predicted RMSD after clustering 100 final structures obtained from HOPE-C^{β} (Table 7).
[c] Predicted RMSD after clustering 100 final structures obtained from DOPE-C^{β} (Table 6).

compatible specification for the dihedral angles of this trimer is retrieved from the rotamer library. If more than one set of dihedral angles is available for this trimer with the given amino acid and basin specifications, one of the trimer conformations is randomly chosen from the available set by weighting all possible trimer conformations uniformly. If the library fails to contain examples with the specified sequence for the trimer and its basins, the algorithm considers the dimer starting at the selected residue, and repeats the same procedure. If the required dimer conformations are also absent, monomer information is used. The energy of the new conformation is evaluated, and the change is accepted according with probability:

$$P = \min\left\{1, e^{-\Delta E/t}\right\} \tag{5}$$

where ΔE is the energy difference between the new and old conformations.

On line material

The homepage of the Protein Library and OOPS is:

http://protlib.uchicago.edu/index.html

All the raw simulation data can be downloaded from:

http://protlib.uchicago.edu/decoys.html

Acknowledgements

We thank Professors A. Fernández, S. Koide, and R. Scott and members of our group for comments and discussions. This work is supported by grants from the NIH (to T.R.S. GM55694, to A.S. GM54762) and NSF (to K.F.F. CHE-031226 and CHE-0416017). The computations have been performed using the AVIDD cluster at Indiana and Purdue Universities, which is supported by grant no. 0116050 from the NSF, resources of the Illinois Bio-Grid at DePaul University, College of Computer Science, Telecommunications, and Information Science, and the cluster of Caroline M. Taylor, Michigan Technological University. This material is also based upon work supported by the National Science Foundation under the following NSF programs: Partnerships for Advanced Computational Infrastructure, Distributed Terascale Facility (DTF) and Terascale Extensions: Enhancements to the Extensible Terascale Facility. A.C. was supported by postdoctoral fellowship from The Program in Mathematics and Molecular Biology. A.J. acknowledges the support of Burroughs Wellcome Fund Interfaces 1001774. We are also grateful to the Sandler Family Supporting Foundation (to A.S.). Joe DeBartolo kindly ran the clustering calculations.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.08.035

References

1. Chowdhury, S., Lee, M. C., Xiong, G. & Duan, Y. (2003). *Ab initio* folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* **327**, 711–717.
2. Yue, K., Fiebig, K. M., Thomas, P. D., Chan, H. S., Shakhnovich, E. I. & Dill, K. A. (1995). A test of lattice protein folding algorithms. *Proc. Natl Acad. Sci. USA*, **92**, 325–329.
3. Arakaki, A. K., Zhang, Y. & Skolnick, J. (2004). Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, **20**, 1087–1096.
4. Ginalski, K., Grishin, N. V., Godzik, A. & Rychlewski, L. (2005). Practical lessons from protein structure prediction. *Nucl. Acids Res.* **33**, 1874–1891.
5. Fang, Q. & Shortle, D. (2003). Prediction of protein structure by emphasizing local side-chain/backbone interactions in ensembles of turn fragments. *Proteins: Struct. Funct. Genet.* **53** (suppl. 6), 486–490.
6. Misura, K. M. & Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. *Proteins: Struct. Funct. Genet.* **59**, 15–29.
7. Bradley, P., Misura, K. M. & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, **309**, 1868–1871.
8. Hardin, C., Pogorelov, T. V. & Luthey-Schulten, Z. (2002). *Ab initio* protein structure prediction. *Curr. Opin. Struct. Biol.* **12**, 176–181.
9. Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2003). Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins: Struct. Funct. Genet.* **53** (suppl. 6), 334–339.

Author's personal copy

Author's personal copy

10. Kretsinger, R. H., Ison, R. E. & Hovmoller, S. (2004). Prediction of protein structure. *Methods Enzymol.* **383**, 1–27.
11. Capriotti, E., Fariselli, P. & Casadio, R. (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, **20** (suppl. 1), I63–I68.
12. Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N. & Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**, 1989–2000.
13. Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **295**, 337–356.
14. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–97.
15. Jones, T. A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO. J.* **5**, 819–822.
16. Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507–533.
17. Bowie, J. U. & Eisenberg, D. (1994). An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl Acad. Sci. USA*, **91**, 4436–4440.
18. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.
19. Jones, D. T. (1997). Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins: Struct. Funct. Genet.* (suppl. 1), 185–191.
20. Chikenji, G., Fujitsuka, Y. & Takada, S. (2006). Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study. *Proc. Natl Acad. Sci. USA*, **103**, 3141–3146.
21. Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.
22. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K. *et al.* (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167–180.
23. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86–89.
24. Tobi, D. & Elber, R. (2000). Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins: Struct. Funct. Genet.* **41**, 40–46.
25. Melo, F., Sanchez, R. & Sali, A. (2002). Statistical potentials for fold assessment. *Protein Sci.* **11**, 430–448.
26. Zhang, C., Liu, S., Zhou, H. & Zhou, Y. (2004). An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.* **13**, 400–411.
27. Shen, M. Y. & Sali, A. (2006). Statistical Potential for Assessment and Prediction of Protein Structures. *Protein Sci.* in press.
28. Eramian, D., Shen, M. Y., Devos, D., Melo, F., Sali, A. & Marti-Renom, M. A. (2006). A composite score for predicting errors in protein structure models. *Protein Sci.* **15**, 1653–1666.
29. Colubri, A. & Fernandez, A. (2002). Pathway diversity and concertedness in protein folding: an ab-initio approach. *J. Biomol. Struct. Dyn.* **19**, 739–764.
30. Colubri, A. (2004). Prediction of protein structure by simulating coarse-grained folding pathways: a preliminary report. *J. Biomol. Struct. Dynam.* **21**, 625–638.
31. Shortle, D. (2002). Composites of local structure propensities: evidence for local encoding of long-range structure. *Protein Sci.* **11**, 18–26.
32. Jha, A. K., Colubri, A., Zaman, M. H., Koide, S., Sosnick, T. R. & Freed, K. F. (2005). Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry*, **44**, 9691–9702.
33. Jha, A. K., Colubri, A., Freed, K. F. & Sosnick, T. R. (2005). Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl Acad. Sci. USA*, **102**, 130099–130104.
34. Gong, H., Fleming, P. J. & Rose, G. D. (2005). Building native protein conformation from highly approximate backbone torsion angles. *Proc. Natl Acad. Sci. USA*, **102**, 16227–16232.
35. Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367–392.
36. Tsai, J., Bonneau, R., Morozov, A. V., Kuhlman, B., Rohl, C. A. & Baker, D. (2003). An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins: Struct. Funct. Genet.* **53**, 76–87.
37. Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L., Jr (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001–2014.
38. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326**, 1239–1259.
39. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. (2004). Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
40. Sacchettini, J. C., Gordon, J. I. & Banaszak, L. J. (1989). Refined apoprotein structure of rat intestinal fatty acid binding protein produced in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **86**, 7736–7740.
41. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* **3** , 171–176.
42. Fernández, A., Colubri, A. & Berry, R. S. (2002). Three-body correlations in protein folding: the origin of cooperativity. *Physica A*, **307**, 235–259.
43. Sosnick, T. R., Mayne, L., Hiller, R. & Englander, S. W. (1994). The barriers in protein folding. *Nature Struct. Biol.* **1**, 149–156.
44. Krantz, B. A., Mayne, L., Rumbley, J., Englander, S. W. & Sosnick, T. R. (2002). Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J. Mol. Biol.* **324**, 359–371.
45. Peter, C., Rueping, M., Worner, H. J., Jaun, B., Seebach, D. & van Gunsteren, W. F. (2003). Molecular dynamics simulations of small peptides: can one derive conformational preferences from ROESY spectra? *Chemistry*, **9**, 5838–5849.

46. de Boor, C. (2001). *A Practical Guide to Splines. Applied Mathematical Sciences*, vol. 27, Springer, New York.
47. Neal, S., Nip, A. M., Zhang, H. & Wishart, D. S. (2003). Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR*, **26**, 215–240.
48. Wang, G. & Dunbrack, R. L., Jr (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
49. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747.
50. Aarts, E. H. L. & Korst, J. H. M. (1989). *Simulated Annealing and Boltzmann Machines: A stochastic approach to combinatorial optimization and neural computing.* Wiley, New York.

Edited by M. Levitt

(Received 9 March 2006; received in revised form 18 July 2006; accepted 16 August 2006)
Available online 18 August 2006

Author's personal copy