Journal of Computer-Aided Molecular Design **17:** 725–738, 2003.

Protein fragment reconstruction using various modeling techniques

Michal Boniecki[a], Piotr Rotkiewicz[a,b], Jeffrey Skolnick[b] & Andrzej Kolinski[a,b,*]
[a]*Laboratory of Theory of Biopolymers, Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland;* [b]*Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington Street, Buffalo, NY 14203, USA*

Received 28 April 2003; accepted in revised form 15 October 2003

Key words: comparative modeling, loop modeling, Monte Carlo sampling, protein structure prediction, reduced protein models

Summary

Recently developed reduced models of proteins with knowledge-based force fields have been applied to a specific case of comparative modeling. From twenty high resolution protein structures of various structural classes, significant fragments of their chains have been removed and treated as unknown. The remaining portions of the structures were treated as fixed – i.e., as templates with an exact alignment. Then, the missed fragments were reconstructed using several modeling tools. These included three reduced types of protein models: the lattice SICHO (Side Chain Only) model, the lattice CABS (Cα + Cβ + Side group) model and an off-lattice model similar to the CABS model and called REFINER. The obtained reduced models were compared with more standard comparative modeling tools such as MODELLER and the SWISS-MODEL server. The reduced model results are qualitatively better for the higher resolution lattice models, clearly suggesting that these are now mature, competitive and complementary (in the range of sparse alignments) to the classical tools of comparative modeling. Comparison between the various reduced models strongly suggests that the essential ingredient for the sucessful and accurate modeling of protein structures is not the representation of conformational space (lattice, off-lattice, all-atom) but, rather, the specificity of the force fields used and, perhaps, the sampling techniques employed. These conclusions are encouraging for the future application of the fast reduced models in comparative modeling on a genomic scale.

Introduction

The gap between the number of known protein sequences and the number of solved three-dimensional structures is rapidly increasing despite the enormous efforts of structural genomics projects [1–3]. Thus, the ability to theoretically predict protein structure is one of the most important challenges of computational molecular biology [4, 5]. For at least 60% of the newly sequenced proteins, it is possible to match their sequences to another sequence (that is similar or related through evolution) for which the structure is already solved. This provides a starting point for comparative modeling where the known structure is used as a template for model building [6, 7]. There are various approaches to this stage of comparative modeling – the actual building of a model with computational tools. The most well known of these is the method proposed by Sali and Blundell [8]. This method is used by MODELLER, a versatile computer program for comparative modeling [7]. MODELLER extracts distance restraints from a template (or templates), builds the consensus model of the aligned part of the query protein, and fills-in the loops using the best fitting fragment from the structural database (in a newer version of MODELLER a loop optimization is employed [9]). Two versions of the MODELLER algorithm were used, a standard one and a modified one with predicted secondary structure superimposed. The second one leads to somewhat better results. Ten examples of loops were generated for each test protein and the best selected for the final analysis. The

*To whom correspondence should be addressed. E mail: kolinski@chem.uw.edu.pl

resulting structures can be subsequently refined using the classical methods of molecular mechanics. The improvement of the models due to refinement is usually marginal, and therefore the refinement step is generally not done. A somewhat different approach is implemented in SWISS-MODEL, a fully automated structure homology-modeling server [10], accessible via the ExPASy server (http://us.expasy.org). The 'framework' for the modeling is built as an average structure obtained from the best superimposition of the detected (via sequence alignments) templates [11]. The loop building technique resembles that of MODELLER. In its final step, SWISS-MODEL performs energy minimization using the GROMOS force field [12].

Recently, we developed a series of reduced protein models [13, 14] (lattice and off-lattice) that proved to be very useful tools for studying the protein folding process [15] and the prediction of protein structure [16–22]. These models have different geometric resolution, but similar force fields. The relatively high resolution and specificity of these reduced models enables their applications in comparative modeling where local details play an important role; i.e., the resolution of the produced models is expected to be high [19]. Indeed, CASP (Critical Assessment of protein Structure Prediction) experiments have proven that the reduced models have now become competitive with approaches that employ all-atom detailed representations and atomic level force fields [18, 23].

Essential for most approaches of comparative (or homology) modeling is to have as good as possible initial sequence alignment (or several alignments) of the query sequence to the sequence of the template [23] (or templates). Various procedures can be used to build the alignments. As a result, comparison of various molecular modeling tools, applied subsequently, becomes somewhat ambiguous. To disentangle these two issues, here we consider a more controlled experiment. Specifically, from a set of proteins (of known structure), fragments of their structures were removed and then rebuilt using various modeling programs. The fragments contained 1–3 randomly selected, consecutive secondary structure elements (α-helices, β-strands) with connecting loops. The average content of regular secondary structure in the modeled fragments was the same as for the entire structures. The length of the removed fragments varied from 10 to 29 residues. Thus, in most cases, the task goes significantly beyond the typical loop-building approach of classical comparative modeling [9], where the missed fragments are usually 2–7 residues long (although sometimes they are longer). Furthermore, in a couple of cases, the removed fragments constitute a part of the protein core. These more difficult cases should help with estimating the limitations of various approaches. All methods employed in this work start from the same (ideal) alignment. Parenthetically, let us note that such an approach has another practical application – i.e., building a complete model from incomplete experimentally solved structures. For various reasons, a significant fraction of the protein structures deposited in the Protein Data Bank (PDB) is incomplete. There are twenty protein structures in the test set, varying in size from 69 to 295 residues. These randomly selected nonhomologous proteins represent various structural classes: all α, β and α/β. It appears that the number of proteins in the test set, while not large, is sufficient for at least a preliminary comparison of various modeling methods. Several modeling tools are compared in this work. Three distinct reduced models developed in our lab employ a very different representation of proteins [18]. The crudest (but the fastest) is the SICHO (SIde CHain Only) model, where only the united atoms representing the center of mass of the side groups and the backbone Cα are explicitly treated [14]. The second model examined is the CABS model, where a residue is represented by up to three united groups (the Cα carbon, the Cβ and the center of the remaining portion of the side chain, where applicable). The third type of reduced models is an off-lattice, continuous space model (REFINER), based on a reduced representation similar to that of the CABS model. The results obtained with these simplified modeling tools are compared with the molecular models generated using MODELLER and SWISS-MODEL. In all cases, a fully automated version of each modeling tool was employed, and no human intervention in the modeling process was allowed. The main purpose of this work is to test the range of applicability of the reduced models of proteins in a well-controlled (yet limited to a single fragment reconstruction), comparative modeling experiment. More complete, and probably more typical, studies of homology modeling will be described in a forthcoming work. The implications of the proposed methodology for the large-scale comparative modeling of genomes are briefly discussed.

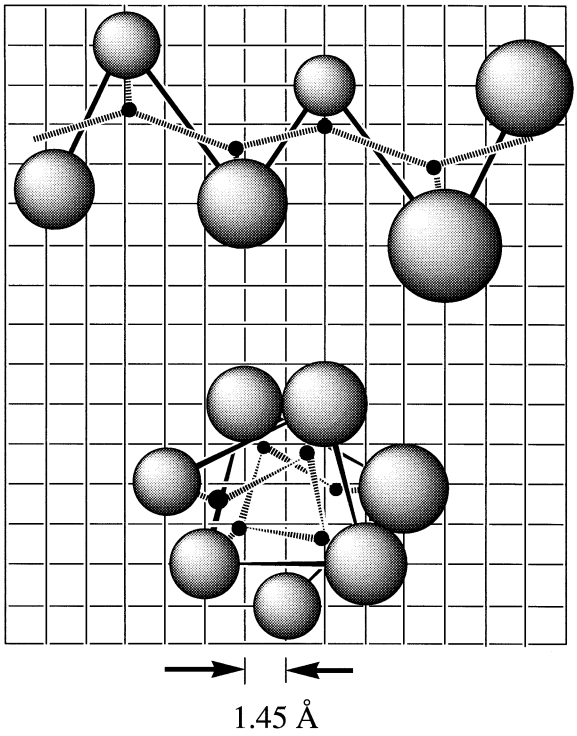

Figure 1. Schematic illustration of the SICHO representation of polypeptides (see the text for details).

Methods

Three different reduced protein models were tested: the SICHO lattice model, the CABS lattice model and the REFINER off-lattice model with up to three united atoms per residue. From the set of twenty test protein structures, relatively large fragments were removed and then rebuilt without any *a priori* information about their native structure passed on to the modeling program. This was done by first projecting the test protein onto a reduced representation (lattice or off-lattice, respectively). Then, the selected fragment of structure is removed, and the corresponding fragment is inserted in a random conformation. During the simulations, the remainder of the structure was kept fixed except for the two residues flanking the modeled fragment. The lowest energy structure was selected from each trajectory as the final structure for further analysis. The results for various reduced models were compared with the models generated via more standard procedures based on an all-atom representation of protein structure.

The SICHO model

This model employs an extremely simplified representation of polypeptide chains. The model chain consists of a string of beads representing the centers of mass of the side chains plus the Cα atoms. The beads are restricted to a simple cubic lattice with a lattice spacing equal to 1.45 Å. Thus, the average accuracy of a projection of a protein structure onto lattice is about 0.8 Å, as measured from the centers of mass of the side chains or from the Cαs (see below). Due to the various sizes of the side chains and their various internal conformations, the virtual bonds span a broad range of lengths that reflect the distribution seen in protein structures (see Figure 1). Cαs are treated implicitly (with their approximate coordinates computed from the positions of the three closest side chains) and provide a framework for definition of interactions that mimic the effect of the main chain hydrogen bonds. Here, in contrast to numerous previous applications of the SICHO model [13–21], only statistical potentials (the same for all proteins) were used. The generic force field contains the following components: sequence independent local conformational bias towards protein-like conformational stiffness, a model of main chain hydrogen bonding (dependent on the mutual positions of the Cα atoms), sequence dependent short-range secondary preferences, orientation dependent pairwise interactions of the side chains, a one-body centrosymmetric burial potential and multibody potentials mimicking the hydrophobic effect and an implicit solvent. The SICHO model (as well as the two remaining models described below) employs the Replica Exchange Monte Carlo [24, 25] (REMC) sampling technique to sample conformational space. We used 20 replicas distributed in a wide range of temperatures. At high temperatures the modeled fragments sampled the random coil regime, while at the lowest temperature the modeled fragment exhibited only small local fluctuations with amplitudes smaller than the accuracy of the method. The lowest energy structure from the lowest temperature replica was used as a final model for comparison with the other approaches. The computational cost of a simulation of a single protein depends on the size of the fragment modeled and varies from 5 min (1 GHz PC) for the shortest fragments represented by the SICHO model up to 1 h for the longest fragments simulated by the REFINER model. The details of the SICHO representation and its knowledge-based force can be found in several previous publications [13, 14, 19, 26], and the potentials

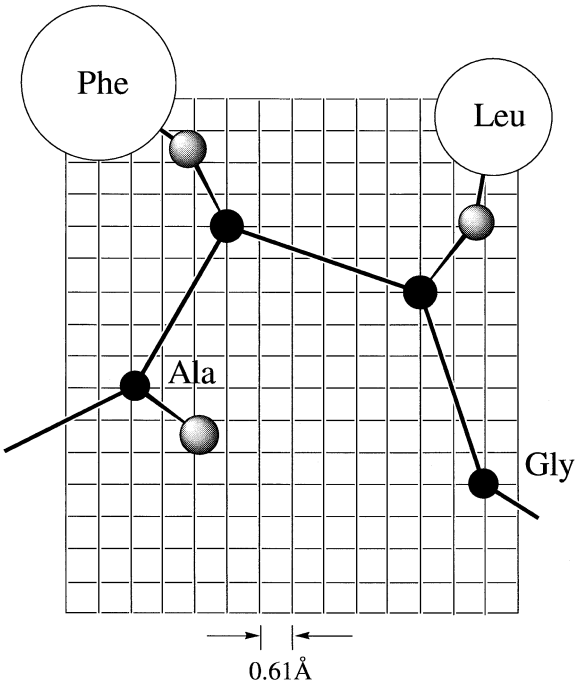

Figure 2. Schematic illustration of the CABS representation of polypeptides (see the text for details).

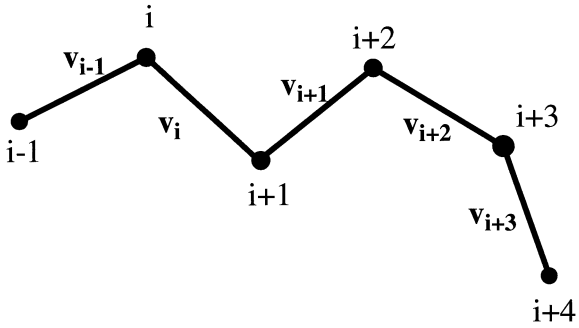

Figure 3. A short fragment of the Cα backbone of the CABS models. A reference for the description of the short range potentials (see the text for more details).

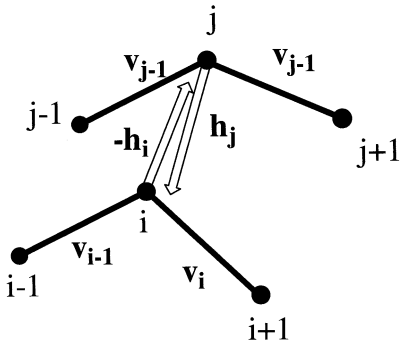

Figure 4. Illustration of the geometry used for the hydrogen bond definition in the CABS model (see the text for more details).

can be obtained by downloading from the web site http://www.biocomp.chem.uw.edu.pl. Finally, it is important to note that the actual accuracy of the SICHO model is lower than the assumed lattice resolution and is in the range of 2–3 Å. This is due to some necessary (related to the lattice representation) geometrical ambiguities of the potentials. The advantage of the SICHO model with respect to the higher resolution models described below is its computational speed resulting from a very flexible and fast local dynamics.

CABS model

In this work, we employed a higher resolution CABS model (a somewhat lower resolution CABS model has been described elsewhere [22]). The acronym CABS stands for the three centers of interactions per residue treated in an explicit way: alpha carbon (CA), beta carbon (B) and center of mass of the remaining side chain atoms (S). The alpha carbon (Cα) trace of the model is restricted to an underlying cubic lattice with a lattice spacing equal to 0.61 Å (see Figure 2). In lattice units, the distance between consecutive Cαs varies from $29^{1/2}$ to $49^{1/2}$. This implies that the Cα-Cα distance is allowed to fluctuate between 3.29 Å and 4.27 Å. Such fluctuations speed up the model's dynamics. The number of possible orientations (lattice vectors) of the virtual Cα-Cα bonds is equal to 800. Consequently, the model essentially avoids any lattice-related artifacts (anisotropy, etc.). The Cβ and side chains are located off-lattice, and their positions are calculated for each residue using the coordinates of three consecutive Cαs as a reference frame. For each amino acid, two rotamers (the most probable according to the statistics of the structural database) are defined – one for expanded conformations and one for more compact conformations (depending on the value of the planar angle of the main chain Cα trace). These two rotamers (and two distinct positions of the Cβs) correspond roughly to the average positions of the side chains (and Cβs) in helical and expanded (β-type) conformations, respectively. The knowledge based force field is designed in a very similar fashion to that of the SICHO model. The details of the geometric definitions and the potentials themselves can be downloaded from our web site (http://www.biocomp.chem.uw.edu.pl). For the readers' convenience, the more complex elements of the model force field are outlined in the Appendix (see also Figures 3 and 4 for reference).

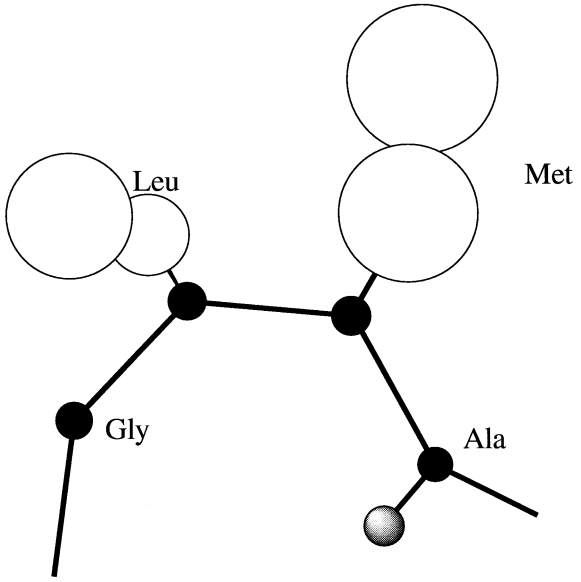

Figure 5. Schematic illustration of the CABS-like off-lattice representation of polypeptides.

REFINER: an off-lattice CABS-like model

This model is very similar in its design to the CABS model, except that it is not restricted to any lattice. Small differences in the force field are related to the continuous space representation. For instance, the side chain pairwise interactions are distance-dependent interpolations of the statistical potentials derived in the form of histograms from the analysis of known protein structures. The only qualitative change in representation (not very significant for the model performance) is the different description of the larger side chains, which are divided onto two united atoms, usually at the point of their maximum conformational flexibility (see Figure 5). Multiple rotameric states of the side chains are allowed in the model and controlled via a set of potentials derived from the statistical analysis of a representative set of crystallographic structures. These changes were intended to increase the spatial resolution of the model. Sampling was again done with the help of the REMC algorithm.

Results and discussion

The results of loop modeling (or, more precisely, fragment remodeling) are compared in Table 1. The first four columns provide the basic data on the 20 protein test systems. Included are proteins of various lengths, with various lengths of removed fragments, and various structural classes. The second block of three columns provides the final cRMSD (coordinate root-mean square deviation from the crystallographic structure after appropriate superimposition) data for the molecular models generated with the SICHO algorithm. The first column gives the cRMSD for the entire protein after the best superimposition with the crystallographic structure. The second column gives the cRMSD values for just the reconstructed fragment after its best superimposition with the corresponding fragment of the crystallographic structure. The third column, the most conservative measure of the model quality, gives cRMSD for the fragment under the condition that only the fixed part of the structure is used to produce superimposition. All values of the cRMSDs are calculated for the Cα atoms for the lowest energy conformations in the pseudo-trajectories (for the lowest temperature replica) obtained in REMC simulations. The results would certainly be somewhat better (i.e., lower cRMSD values) if the trajectory was clustered and a more elaborate fold selection performed [27]. Alternatively, the all-atom models could be reconstructed [28] and used for the selection of the best structure. Here, we opt for the most straightforward comparison of various methods. Clearly, our results for SICHO are consistent with the range of accuracy of this model observed in earlier comparative modeling computational experiments [17]. The next two blocks of Table 1 give the results for the higher resolution CABS lattice model and for the continuous space REFINER. Interestingly, the higher resolution lattice and off-lattice models are essentially of the same quality and accuracy, and the average results are significantly better than for the SICHO model. However, in a couple of cases, the SICHO results were slightly better. Different modeling tools work differently for particular cases.

A more detailed comparison of the CABS and REFINER results shows that the lattice simulations are more consistent. REFINER builds several exceptionally good models, but it also builds a few very bad models. REFINER seems to win in cases of shorter and structurally more regular fragments. The reason is most likely related to the much better sampling efficiency of the lattice model, and may also be related to the somewhat more carefully designed averaged hydrophobic terms of the force field of the lattice model. The results from MODELLER and SWISS-MODEL are qualitatively poorer in almost all cases. Most striking is the comparison of the quality of placing the fragments with respect to the fixed parts of the struc-

Table 1. Comparison of the results of the loop-modeling experiments with various modeling tools.

	System			SICHO			CABS			REFINER			MODELLER-LOOP			MODELLER-LOOP-SS			SWISS-MODEL		
NAME	N	GAP	FRAG	RMSD	FRAG	NOS.	RMSD	FRAG	NOS.	RMSD	FRAG	NOS.	RMSD	FRAG	NOS.	RMSD	FRAG	NOS.	RMSD	FRAG	NOS.
1ad2_	224	19	72–90	2.57	3.84	6.52	1.27	1.73	3.69	0.40	0.85	1.36	5.59	8.61	11.21	5.15	4.61	8.71	1.68	4.99	5.67
1ag4_	103	15	78–92	2.19	3.46	4.00	1.36	2.82	3.07	1.63	3.98	4.22	4.53	3.65	9.50	3.61	3.65	6.85			
1ahk_	129	16	52–67	2.30	3.57	4.28	0.90	1.50	2.08	0.76	1.23	1.99	3.55	4.62	6.92	3.22	4.33	5.08	4.52	7.06	12.05
1ail_	70	18	19–36	3.25	4.19	4.70	2.79	3.98	5.02	1.21	1.73	2.28	4.56	3.40	7.42	3.45	5.10	6.06			
1bfg_	126	16	53–68	2.49	3.18	5.02	1.35	2.16	3.16	2.33	4.59	6.29	3.72	5.22	7.63	3.64	4.32	7.67	7.65	4.62	19.81
1bovA	69	25	23–47	4.74	4.35	6.71	3.52	4.70	5.03	2.50	3.66	4.03	4.83	6.13	11.14	3.96	6.14	8.75			
1cne_	260	14	107–120	2.25	2.47	5.77	0.73	1.45	2.07	1.11	2.21	4.62	4.27	4.36	7.55	3.10	3.26	7.93	3.41	4.28	14.10
1cyo_	88	19	31–49	4.36	4.61	7.75	2.65	4.52	5.26	2.00	2.22	3.84	6.73	5.44	11.46	5.75	3.88	6.28	4.24	3.46	7.89
1dad_	224	29	142–170	7.82	5.23	8.19	1.76	4.04	4.64	0.81	1.66	2.23	3.25	2.88	3.43	2.98	3.35	2.45	0.15	0.35	0.39
1fdr_	245	20	61–80	2.24	4.64	5.34	1.41	3.57	4.40	1.16	2.65	4.53	5.51	5.41	11.34	4.33	4.93	10.63			
1fts_	295	22	219–240	1.84	2.20	2.97	0.98	1.80	2.88	0.83	1.99	2.95	6.51	4.30	6.75	4.65	3.57	6.70	5.10	9.06	17.86
1gifA	115	19	43–61	5.35	7.78	11.86	3.33	5.51	7.36	1.56	3.43	3.74	9.33	8.33	16.61	5.34	3.21	6.44	6.22	6.95	13.24
1hfh_	120	27	39–65	3.92	5.03	6.98	2.25	3.42	4.36	6.19	8.84	12.33	5.88	9.52	12.46	4.91	6.74	11.32	14.95	5.26	27.39
1ife_	91	18	48–65	2.33	2.09	3.61	1.43	1.91	2.42	1.45	1.83	3.07	4.23	4.32	6.78	3.55	3.62	6.23	2.91	5.19	6.30
1jer_	110	17	51–67	3.80	4.71	8.15	2.17	2.50	5.06	1.61	3.28	3.91	5.13	4.29	7.32	6.73	3.54	5.54	2.01	3.44	5.05
1latA	71	20	38–57	5.47	4.77	9.56	3.86	4.13	6.70	5.58	4.44	9.88	5.26	3.44	6.21	5.33	3.46	6.65			
1np4_	184	22	69–90	4.44	8.23	11.43	3.42	5.39	9.22	7.07	5.19	18.52	5.34	9.59	12.79	7.21	8.59	10.56	7.13	5.91	19.31
1plc_	99	10	50–59	1.92	1.84	2.96	1.32	2.38	3.74	1.07	1.68	3.21	4.33	2.43	4.40	3.34	3.21	4.47	2.16	3.57	6.08
1sro_	76	15	20–34	3.57	4.29	6.67	4.62	3.35	5.43	2.37	2.97	5.24	4.68	4.35	5.22	4.23	4.45	6.28	6.90	2.74	14.06
1vhh_	157	18	90–107	2.72	3.71	6.00	1.90	3.59	5.33	2.75	4.52	8.00	4.21	5.84	7.33	5.26	5.34	6.95	2.38	4.40	6.87
AVERAGE				3.48	4.21	6.42	2.15	3.22	4.54	2.22	3.15	5.31	5.07	5.30	8.67	4.48	4.65	7.08	4.78	5.11	12.32
AVERAGE without 2 worst:				3.13	3.79	5.84	1.92	2.97	4.07	1.73	2.72	4.19	4.74	4.83	8.00	4.21	4.11	6.64	3.75	4.24	7.60

Abbreviations:
MODELLER-LOOP – A simple version of MODELLER (LOOP module);
MODELLER-LOOP-SS – MODELLER with a prediction of secondary structure;
RMSD – coordinate RMSD (root-mean-square deviation of the Cα trace from native for entire structure;
FRAG – RMSD for the fragment (only the inserted fragment superimposed);
NOS. – No superimposition on the fragment. Fragment RMSD after best superimposition of the remainder of the structure.
Empty lines for SwissModel means the server refused to build the fragment with fixed remainder of the structure.

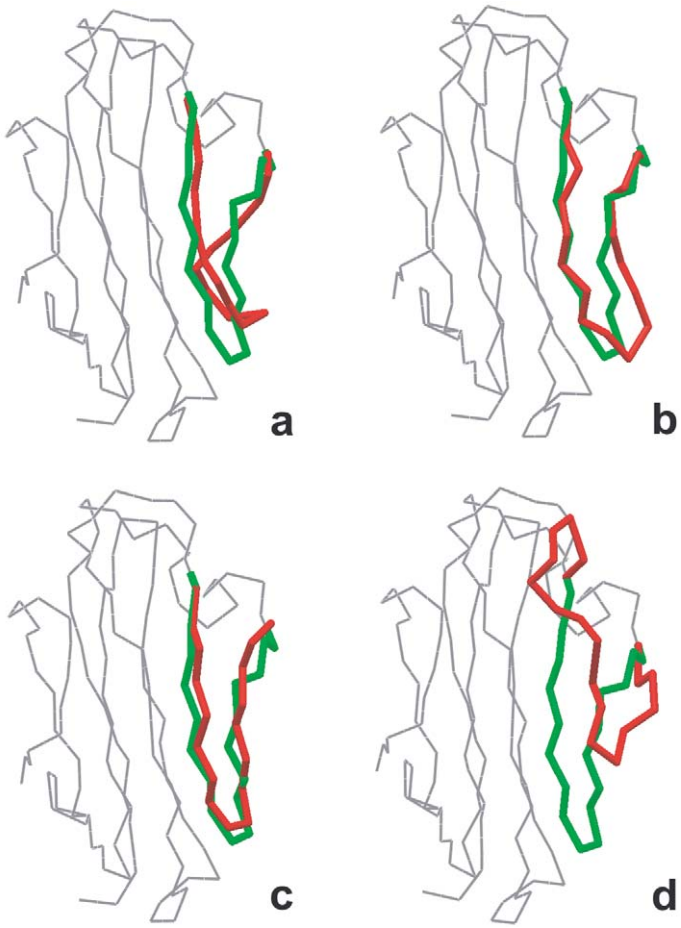

Figure 6. Models obtained by various techniques for 1ahk_. The green tube denotes the native structure of the fragment. The red tube corresponds to the result of modeling. The thin gray line shows the fixed remainder of the structure. (a) the SICHO model, (b) the CABS model, (c) the off-lattice REFINER model and (d) MODELLER.

ture. The best (and the most conservative) measure is the cRMSD of the fragment when the superimposition of the models is done only on the fixed part (the numbers are given in the third columns of each block of data in Table 1). The average difference is huge – 4.5 Å against 8.7 Å for CABS and MODELLER (7.1 for MODELLER with the secondary structure superimposed), respectively – and is correlated with the length of the reconstructed fragment (on average, the longer the fragment, the larger the advantage of the reduced models). MODELLER seems to perform better than SWISS-MODEL since the average cRMSD for the fragments is smaller. SWISS-MODEL frequently builds essentially random conformations of several fragments (with cRMSDs typical for high-temperature random coil conformations seen in the lattice simulations) and, in a few cases, it was impossible to maintain the initial conformation of the fixed portions of the molecules. Of course, this comparison does not intend to devalue the standard techniques of molecular modeling. It is rather to show that the reduced modeling is a complementary approach in situations where rather significant parts of structures need to be modeled in an *ab initio* fashion. It also should be mentioned that other *ab initio* methods could possibly handle the problem of fragment reconstruction [29–31]. However, in order to make a quantitative comparison, the access to source codes and their modifications would be required. This was beyond the scope of this work. Additional comparison of various methods is provided in Figures 6–8, where the final models are shown for three more illustrative cases of protein structures from the test set.

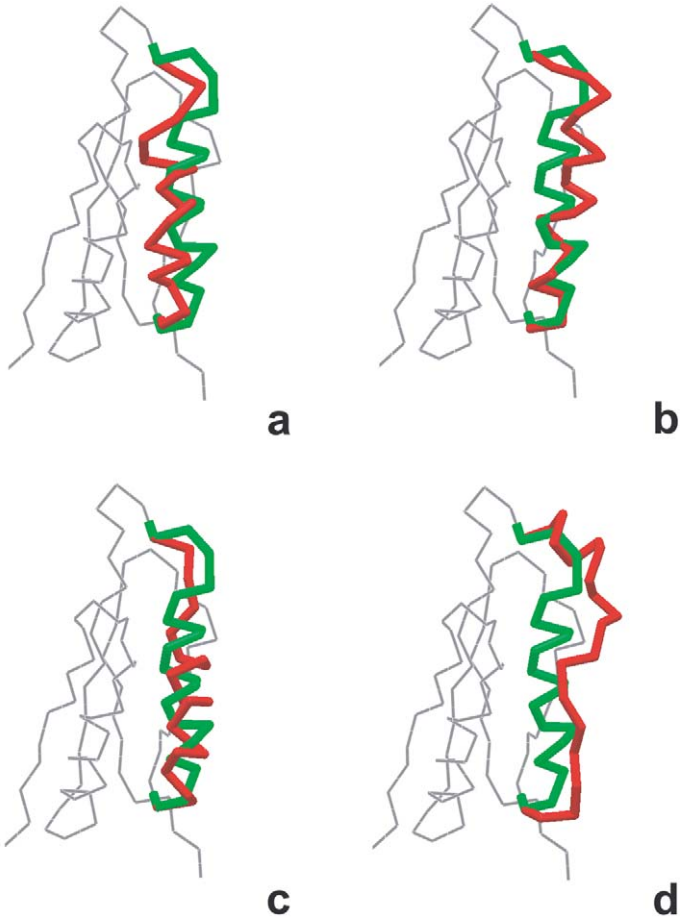

Figure 7. Models obtained by various techniques for 1ife_. The green tube denotes the native structure of the fragment. The red tube corresponds to the results from modeling. The thin gray line shows the fixed remainder of the structure. (a) SICHO model, (b) CABS model, (c) REFINER and (d) MODELLER.

In a few cases, modeling with the reduced models was clearly unsuccessful. The most obvious are the 1hfh and 1np4 results obtained from the off-lattice model. A closer inspection of these cases shows that the failure was due to inability to insert the missing fragment inside of the structure. In these cases the more flexible (and faster sampling) lattice CABS model managed to build better structures. This suggests possible avenues for improvement by enhancing the efficiency of conformational sampling. A related problem is how to identify low quality models (which are rare) in the large-scale modeling exercises. A possible approach may employ a consistency test; if the models from various techniques are close to each other, it is likely that these models are of good quality. Less accurate SICHO models could also be used for reference but not necessarily for building the consensus structures. Another possibility is to use full-atom models for the final evaluation [18, 27]. Tools exist for rapid and accurate reconstruction of atomic details from the reduced model coordinates [28]. Of course, there is a 'brute force' solution to the problem of these poor quality models by allowing relaxation of the entire chain. This is beyond the scope of the present work, and the appropriate analysis will be published in the near future. Genomic scale comparative modeling experiments with the reduced models described here are now underway. The accuracy of the produced models, while still too low for some purposes, approaches the level of accuracy for large-scale approximate drug-design screening, protein–protein docking and related tasks.

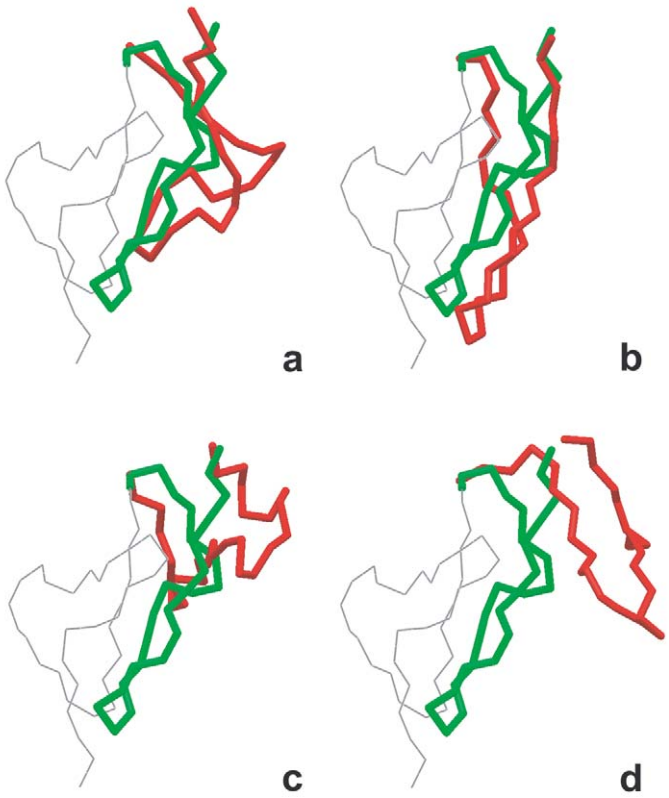

Figure 8. Models obtained by various techniques for 1hfh_. The green tube denotes the native structure of the fragment. The red tube corresponds to the result of modeling. The thin gray line shows the fixed remainder of the structure (only a part is shown for clarity). (a) the SICHO model, (b) CABS model, (c) the off-lattice REFINER and (d) MODELLER.

Conclusions

In this work, we applied three recently developed reduced models of proteins to the test of the quality of the resulting reconstruction of missed protein structure fragments. The results are encouraging; in most cases it was possible to build good quality structures in a fully automated procedure. The newer, higher resolution models (CABS and REFINER) perform systematically better than the older SICHO model. The average difference is about 1 Å cRMSD for the entire structures and 2 Å cRMSD for the reconstructed fragments (measured for the Cα-traces), which is significant. The accuracy of the obtained reduced models seems to be qualitatively better than the accuracy of the models resulting from the automated application of more standard comparative modeling tools. Thus, the reduced model approaches complement the classical techniques in cases where a significant part of a protein structure needs to be built in essentially *ab initio* fashion.

The well-controlled experiment of fragment reconstruction was done for the purpose of easy comparison of the various approaches. With a fixed (and exact) alignment of the main part of the structures, we reconstructed relatively large fragments not necessarily limited to the loop regions. For loop reconstruction, this exercise constitutes a relatively difficult test. It may prove useful for computational completion of experimental structures that have gaps. This is, however, not the most typical type of application of comparative modeling. More challenging tasks need to deal with possible inaccuracies of the entire alignment to the template, and the ability to correct the template structure for distantly homologous cases. Such applications were done during CASP5 [27], and detailed benchmarks will be published in the near future. Elaborate optimization of the force field of the CABS model that is now underway [22] should only increase the predictive strength of this modeling tool. Another interesting observation resulting from this work is that the type of geometrical representation (lattice vs. off-lattice) of

the protein is not very important for the quality of the models. Indeed, the results from the continuous space REFINER, which enables an exact representation of the main chain united atoms and a more accurate (than in the case of the lattice models) representation of the side chains are on average no better than the CABS results. More likely, the structural specificity of the applied force field (and, perhaps, the sampling details) is the deciding factor. In more general comparative modeling applications, the implementation of template (templates) restraints will be very important.

Acknowledgements

This work is partially supported by National Institutes of Health Grant GM-48835 of the Division of General Medical Sciences and by the Oishei Foundation.

Appendix

The force field of the CABS model contains several terms. Generic terms provide protein-like geometrical biases and narrow the conformational space to be searched, partially correcting for reduced representation of the model. The sequence specific potentials are derived via statistical analysis of the structural regularities seen in known structures of globular proteins. Particular potentials are outlined below.

1. Generic (and secondary structure biased) protein-like conformational stiffness

First, we describe the set of sequence independent potentials (or conformational biases) that provides protein-like, conformational stiffness, thereby reducing the conformational space of the model. A detailed description of the generic protein-like conformational biases seems appropriate, since these potentials are very characteristic of our approach to protein modeling. All-atom models do not need them, since the detailed interactions between atoms, to a large extent, provide proper local conformational stiffness. Similarly, quite successful methods based on protein assembly from small fragments excised from the structural database [29, 30] do not need the majority of such terms of the force field. Recently, our approach to protein structure modeling became comparable in accuracy (and predictive strength) with the methods based on fragment assembly [18, 20–22]. Thus, it is very likely that the way the intraprotein interactions are designed in our high-coordination lattice models actually reflects some basic physical features of polypeptides and proteins.

Consider a sequence of five backbone vectors, as shown schematically in Figure 3. Backbone vectors connect the Cα beads: $\mathbf{v_i} = \mathbf{r_{i+1}} - \mathbf{r_i}$, with $\mathbf{r_i}$ being the Cartesian coordinate of the i-th Cα vertex. The following biases are applied to each such fragment.

$$B_1 = 0.5 \times f \times \varepsilon_g \tag{1}$$

for: $(\mathbf{v_{i-1}} \bullet \mathbf{v_{i+3}}) < 0$ and $|\mathbf{r_{i+4}} - \mathbf{r_i}| < 7.0$ Å,

or for: $(\mathbf{v_{i-1}} \bullet \mathbf{v_{i+3}}) < 0$ and $|\mathbf{r_{i+4}} - \mathbf{r_i}| > 11.0$ Å

$$B_2 = -0.5 \times f \times \varepsilon_g - \varepsilon_g \tag{2}$$

for: $|\mathbf{r_{i+4}} - \mathbf{r_i}| < 7.0$ Å,

and 4.0 Å $< |\mathbf{r_{i+3}} - \mathbf{r_i}| < 7.0$ Å,

and 4.0 Å $< |\mathbf{r_{i+4}} - \mathbf{r_{i+1}}| < 7.0$ Å,

and both above fragments are right-handed,

and residues i+1,i+2, i+3 are not assigned as β-type,

and $(\mathbf{v_{i+1}} \bullet \mathbf{v_{i+3}}) < 0$ and $(\mathbf{v_i} \bullet \mathbf{v_{i+3}}) > 0$

$$B_3 = -0.5 \times f \times \varepsilon_g - \varepsilon_g \tag{3}$$

for: $|\mathbf{r_{i+4}} - \mathbf{r_i}| > 11.0$ Å,

and residues i+1,i+2, i+3 are not assigned as α-type,

and $(\mathbf{b_{i+1}} \bullet \mathbf{b_{i+2}}) < 0$

and $(\mathbf{b_{i+2}} \bullet \mathbf{b_{i+3}}) < 0$

In the above equations:

$$\mathbf{b_i} = (\mathbf{v_i} - \mathbf{v_{i+1}})/|(\mathbf{v_i} - \mathbf{v_{i+1}})| \tag{3a}$$

$$f = \min(1, \quad (S/s)^2) \tag{3b}$$

where: S is the radius of gyration of the protein (S can be computed quite accurately for single domain globular proteins from the number of amino acids) and s is the mean square distance of the center of mass of the discussed fragment from the center of mass of the polypeptide chain in the actual conformation. The symbol || denotes vector length.

The conditions written in Equation 1 mean that the chain conformation is penalized for breaking a stretch of compact (usually helical) or expanded conformations (see Figure 3 for a reference). Equation 2 and Equation 3 describe the conditions for the energetically favored helical or β-type conformations of short

fragments, respectively. When a predicted (or experimental) secondary structure assignment is known (in a three-letter code), these loosely defined helical or expanded conformation biases are applied selectively; i.e., the helical bias is never applied to the putative β-type fragments, and vice versa. The scaling of the biases by factor f provides more flexibility for the chain fragments located on the surface of the globule or in a swollen denatured state than in the core of the protein. The prefactor 0.5 in Equation 5 is an arbitrary one, adjusted by trial and error, using the criterion of a protein-like distribution of local distances in the model polypeptide chain. The only parameter that scales the entire scheme is denoted by ε_g, and it is equal to 1.0 at reduced (dimensionless) temperatures of 1.0–2.0.

Local protein-like biases described above lead to a protein-like distribution of local conformations, even in the absence of any sequence-dependent potential. Sometimes fragments of structures 'crumple', forming excessively folded clusters of very short elements of secondary structure. To eliminate this effect, a medium-range generic stiffness term has been introduced in the form of a penalty for too close in sequence U-turns (or a short N-shaped fragment) of the model chain:

$$B_4 = 4.0 \times \varepsilon_g$$

$$\text{for:} \quad (\mathbf{r_{i+5}} - \mathbf{r_i}) \bullet (\mathbf{r_{i+10}} - \mathbf{r_{i+5}}) < 0 \qquad (4)$$

$$\text{and} \quad (\mathbf{r_{i+15}} - \mathbf{r_{10}}) \bullet (\mathbf{r_{i+5}} - \mathbf{r_i}) > 0$$

The minimal length of the stretch (5 residues) between turns and the value of the prefactor in the potentials were adjusted by trial and error. Again, the protein-like distribution of distances between chain U-turns in a folded structure was used as a criterion.

For the fragments with an assumed secondary structure, one more conformational bias has been introduced. In all previous applications of the model, very conservative consensus predictions of the secondary structure were employed. Namely, only the consensus predictions of a helix or a β-strand from high scoring (score>4) outputs of the automatic servers PHD or PSI-PRED were accounted for (see http://www.embl-heidelberg.de/predictprotein/predictprotein.html). In the present application, the secondary structure bias is of marginal importance since most of the protein's structure is frozen during the simulations.

$$B_5 = \delta \times (0.25 \times d \times \varepsilon_g + 0.5 \times \varepsilon_g) \qquad (5)$$

where for helical fragments: $d = \text{abs}(|\mathbf{r_{i+7}} - \mathbf{r_i}| - 10.75$ Å)
with $\delta = 0$ for $d < 0.61$ and $\delta = 1$ elsewhere,
and for β-type fragments: $d = \text{abs}(|\mathbf{r_{i+6}} - \mathbf{r_i}| - 19.1$ Å$) - 1.0$ Å
with $\delta = 0$ for $d < 1.22$ Å and $\delta = 1$ elsewhere,
no bias superimposed onto fragments that contain coil-assigned residues; $\delta = 0$.

In the above definitions, 10.75 Å is the average distance between the i-th and i+7-th residues in helices, while 19.1 Å is the average distance between the i-th and i+6-th residues in β-strands. The different distance along the chain and a different definition of the β function for both types of secondary structure is related to the different spacing between residues along the chain and different dispersion of the distances.

The total energy of the model chain resulting from the generic local conformational stiffness is the sum of all four components over the entire chain (of course, the range of summation differs for various terms due to different lengths of the fragments involved) and could be written symbolically as follows:

$$E_g = \Sigma(B_1) + \Sigma(B_2 + B_3) + \Sigma(B_4) + \Sigma(B_5) \qquad (6)$$

2. Sequence-dependent short-range interactions

These statistical potentials have a form of histograms and were derived from the statistics of a non-redundant structural database (35% sequence identity cut-off) of globular proteins.

- The potentials controlling the distance between the i-th and i+2-nd residues have two bins (rough division between helical and expanded conformations) and depend on the identity of the two flanking amino acids.
- The potentials controlling the distance between the i-th and i+3-rd residues have 24 bins (width of 1 Å) for the distance $|\mathbf{r_{i+3}} - \mathbf{r_i}|^*$ and depend on the identity of the two central amino acids. The range of $|\mathbf{r_{i+3}} - \mathbf{r_i}|^*$ is (−12, 12) and the symbol '*' denotes the chiral character of this potential. Negative (positive) values denote left-handed (right-handed) conformations of the three-bond fragments.
- The potentials controlling the distance between the i-th and i+4-th residues have 16 bins (width of 1 Å) of the $|\mathbf{r_{i+4}} - \mathbf{r_i}|$ and depend on the identity of the two flanking amino acids.

The short-range potentials (except the first one) were derived from three subsets of the database: the entire

database (basic potential), the database of predominantly helical proteins and the database of predominantly β-type proteins. For those fragments strongly assigned as helical or expanded, the basic potentials were modified via 50/50 convex combinations with appropriate database-biased potentials. The scaling factor for the short-range interactions is equal to 0.25–0.5. This magnitude of the potential ensures the proper (protein-like) level of secondary structure in model chains. The details of the derivation of these short range potentials are essentially identical to that described previously for the SICHO model [13, 14, 17]. The only significant difference is that in the SICHO model, the short-range potential controls distances between side-group united atoms, whereas the CABS model operates on the Cα distances. The short-range potentials for the Cα trace are better defined (sharper helical peaks, etc.) than for the virtual chains of side groups. On the other hand, the SICHO short range potential accommodates to some extent local packing preferences (or various rotational isomeric preferences) of the side chains. All potentials can be found on our web site (http://www.biocomp.chem.uw.edu.pl). Therefore, we will not provide here a more detailed elaboration of the technical details.

3. Model of hydrogen bonds

Only main chain hydrogen bonds are accounted for in the model. They are treated in an implicit way. Specifically, it is easy to check that the presence of the main chain hydrogen bonds translates onto a specific geometrical arrangement of the Cα atoms. Thus, as in many previous applications, the model 'hydrogen bond' is defined as highly directional interactions between Cαs. Such an approach leads to some necessary 'renumbering' of the interacting residues. For instance, the canonical hydrogen bond between the i-th and i+4-th residues in helices translates in the Cα-Cα substitute as occurring between the i-th and i+3-rd Cαs. A detailed explanation of the reasons for such renumbering can be found in our previous publications on the SICHO model [14, 17]. Figure 4 provides a geometric reference for the definition of the model hydrogen bond. The vectors **h** are orthogonal to the planes formed by two consecutive Cα-Cα vectors (collinear conformations of the main chain are forbidden), and their length is equal to 4.5 Å. Residues i and j are 'hydrogen bonded' when the following set of geometrical conditions are satisfied:

$$|\mathbf{r_i} - \mathbf{r_j}| < 5.8 \text{ Å} \tag{7a}$$

$$\text{abs}(\mathbf{h_i} \bullet \mathbf{h_j}) > 10 \quad (in \text{Å}^2) \tag{7b}$$

$$\begin{aligned} &(\mathbf{v_{i-1}} \bullet \mathbf{v_{j-1}}) > 0 \quad \text{and} \quad (\mathbf{v_{i+1}} \bullet \mathbf{v_{j+1}}) > 0 \\ \text{or } &(\mathbf{v_{i-1}} \bullet \mathbf{v_{j+1}}) < 0 \quad \text{and} \quad (\mathbf{v_{i+1}} \bullet \mathbf{v_{j-1}}) < 0 \end{aligned} \tag{7c}$$

$$|\mathbf{r_i} - \mathbf{r_j}| - |\mathbf{h_i}| < 1.9 \text{ Å} \tag{7d}$$

The first condition defines a distance cut-off for the interacting atoms. The second condition means that the angle between the vectors is less than 60° or greater than 120°. The third means that the relevant two bond fragments have an approximately parallel (as in helices and parallel β-sheets) or antiparallel mutual orientation (as in antiparallel sheets). The fourth condition means that the vector $\mathbf{h_i}$ points in the vicinity of the jth Cα. A similar condition holds for the second 'hydrogen bond' – i.e., $|\mathbf{r_i} - \mathbf{r_j}| - |\mathbf{h_j}| < 1.9$ Å. Thus, there are always up to two 'hydrogen bonds' per one Cα (similarly, up to two hydrogen bonds per residue in all-atom models, or real proteins). Additionally, the model hydrogen bonds between the i-th and i+4-th residues are forbidden (not counted), since bifurcation of hydrogen bonds in model helices causes distortion. Also, in order to prevent nonphysical crumpling of the β-structures, the parallel β-type of hydrogen bonds for $|i - j| < 20$ are not allowed. Such regularity can be observed in real protein structures. With the above definition, the model hydrogen bonds rarely bifurcate (bifurcation would result in more than two bonds per residue). The DSSP assignment of hydrogen bonds in crystal structures correlates very well with the model definition (ca. 90% of hydrogen bonds coincide). The strength of model hydrogen bonds is computed as follows:

$$E_h = \delta_h \times \varepsilon_h + \delta_\gamma \times \varepsilon_\gamma \times [22.3/(\max(22.3, |\mathbf{r_i} - \mathbf{r_j}|^2) - 0.66], \tag{8}$$

where: $\delta_h = 1$ when conditions given in Equation 7a–d are satisfied and $\delta_h = 0$ otherwise. $\delta_\gamma = 1$ when conditions given in Equation 7a–c are satisfied and $\delta_\gamma = 0$ otherwise. The scaling factors are adjusted arbitrarily and their values are $\varepsilon_h = -1.5$ and $\varepsilon_\gamma = -1.0$. These values provide average protein-like levels of regular secondary structure in collapsed model chains. The second term in Equation 8 provides weak directional attractive interactions between the Cαs separated by a distance slightly larger than the

average length of the model hydrogen bond. The total energy of hydrogen bonds reads as:

$$E_H = \Sigma\Sigma(g_{ij} \times E_h), \quad (9)$$

where g_{ij} is the factor that amplifies the strength of the hydrogen bonds in the predicted (assigned) secondary structure elements and is equal to 1.5 for intrahelical hydrogen bonds and β-sheet assigned residues; otherwise, $g_{ij} = 1.0$. Such differentiation of the strength of hydrogen bonds has a noticeable effect on the assembly of the correct topology in *ab initio* folding. For instance, strongly predicted β-strands are usually the inner strands of a β-sheet. Stronger model hydrogen bonds tend to be saturated (two bonds per residue) first. Consequently, the proper strands tend to be buried inside a sheet, instead being the edge strands with one hydrogen bond per residue. Similarly, intrahelical hydrogen bonds propagate helices, while the weaker bonds at the caps of a helix may break the pattern and allow formation of a turn or loop.

4. Pairwise interactions of the side groups

The statistical potential for the side chain interactions is orientation dependent. Three types of side chain contacts are considered: parallel, intermediate and antiparallel. It seems to be a rather important feature of our force field, which is usually ignored in other statistical potentials. In this approach, the value of the Lys-Glu potential is equal to −1.0, 0.2 and 1.4 for the parallel, intermediate and antiparallel orientations, respectively. Ignoring the angular dependence leads to averaging out of the potential. Consequently, the very strong tendency of the charged residues to adopt mutually parallel orientation of the side groups on the protein surface will be absent in an orientation independent knowledge-based potential. During the derivation of the potentials, two side chains are considered to be 'in contact' when any pair of their heavy atoms are closer to each other than 4.5n Å. The resulting cut-off distances for the side chain center of mass are therefore pair-specific. See the appropriate tables at www.biocomp.chem.uw.edu.pl. Additionally, to account for roughly averaged electrostatic interactions, a tail potential was added for pairs of charged residues

$$E_{Sg} = \Sigma\Sigma(f_{ij} \times \varepsilon_{ij})$$
$$\text{where:} \quad f_{ij} = 1 \quad \text{for} \quad R_{min} < R_{ij} < R_{max} \quad (10)$$
$$\text{and:} \quad f_{ij} = (R_{max}/R_{ij})^2 \quad \text{for} \quad R_{max} < R_{ij}$$

For values $R_{ij} < R_{min}$, there is a constant, high value repulsive energy (soft core excluded volume for the side chains, while hard core assumed for Cα and Cβ united atoms). R_{ij} denotes the distance between the centers of interacting side chains, and R_{min} and R_{max} are the cut-off distances for the square well (or square shoulder) potentials.

5. Burial potentials

Due to the lack of explicit solvent in the model of the force field, the hydrophobic effect is accounted for on the level of packing regularizing potentials. First, there are very weak centrosymmetric interactions that reflect the tendency of the nonpolar residues to be buried inside the protein globule. Then, a packing 'profile' was built for all side chains in a form of statistical potential controlling the number of parallel, intermediate and antiparallel contacts. This is a somewhat different and simpler approach to the burial interactions than that employed in the SICHO model, where the solvent exposure was calculated explicitly in an approximate fashion [17].

The model has two more packing controlling potentials – one includes a predicted target number of contacts in the folded structure, and the second includes the average contact order. In this particular application of loop modeling with a fixed remainder of the molecule, these two components are marginal and could be safely ignored. They were merely mentioned for the sake of completeness.

The sampling of CABS structures was performed using the REMC algorithm [20, 21, 24, 25], with 20 replicas and the lowest replica assigned to the reduced temperature $T = 1.0$.

References

1. Baker, D. and Sali, A., Science, 294 (2001) 93.
2. Chance, M.R., Bresnick, A.R., Burley, S.K., Jiang, J.S., Lima, C.D., Sali, A., Almo, S.C., Bonanno, J.B., Buglino, J.A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M. and Wang, L.K., Protein Sci., 11 (2002) 723.
3. Montelione, G.T., Zheng, D., Huang, Y.J., Gunsalus, K.C. and Szyperski, T., Nat. Struct. Biol., 7 Suppl. (2000) 982.
4. Simons, K.T., Strauss, C. and Baker, D., J. Mol. Biol., 306 (2001) 1191.
5. Skolnick, J., Fetrow, J.S. and Kolinski, A., Nat. Biotech., 18 (2000) 283.
6. Pieper, U., Eswar, N., Stuart, A.C., Ilyin, V.A. and Sali, A., Nucleic Acids Res., 30 (2002) 255.

7. Sali, A., Nat. Struct. Biol., 5 (1998) 1029.
8. Sali, A. and Blundell, T.L., J. Mol. Biol., 234 (1993) 779.
9. Fiser, A., Do, R.K. and Sali, A., Protein Sci., 9 (2000) 1753.
10. Schwede, T., Diemand, A., Guex, N. and Peitsch, M.C., Res. Microbiol., 151 (2000) 107.
11. Bajorath, J., Stenkamp, R. and Aruffo, A., Protein Sci., 2 (1993) 1798.
12. van Gunsteren, W.F. and Weiner, P.K., Computer Simulations of Biomolecular Systems. Theoretical and Experimental Applications. ESCOM Science Publishers B.V., Leiden, 1989.
13. Kolinski, A. and Skolnick, J., Proteins, 32 (1998) 475.
14. Kolinski, A., Rotkiewicz, P., Ilkowski, B. and Skolnick, J., Progress of Theoretical Physics (Kyoto), Suppl. 138 (2000) 292.
15. Kolinski, A., Ilkowski, B. and Skolnick, J., Biophys. J., 77 (1999) 2942.
16. Skolnick, J., Kolinski, A. and Ortiz, A., Proteins, 38 (2000) 3.
17. Kolinski, A., Betancourt, M., Kihara, D., Rotkiewicz, P. and Skolnick, J., Proteins, 44 (2001) 133.
18. Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P. and Boniecki, M., Proteins, Suppl. 5 (2001) 149.
19. Skolnick, J. and Kolinski, A., Adv. Chem. Phys., 1209 (2002) 131.
20. Kihara, D., Lu, H., Kolinski, A. and Skolnick, J., Proc. Natl. Acad. Sci. USA, 98 (2001) 10125.
21. Kihara, D., Zhang, Y., Kolinski, A. and Skolnick, J., Proc. Natl. Acad. Sci. USA, 99 (2002) 5993.
22. Zhang, Y., Kolinski, A. and Skolnick, J., Biophys. J., 85 (2003) 1145.
23. Schonbrun, J., Wedemeyer, W.J. and Baker, D., Curr. Opin. Struct. Biol., 12 (2002) 348.
24. Hukushima, K. and Nemoto, K., J. Phys. Soc. (Jpn.), 65 (1996) 1604.
25. Hansmann, U.H. and Okamoto, Y., Curr. Opin. Struct. Biol., 9 (1999) 177.
26. Kolinski, A., Jaroszewski, L., Rotkiewicz, P. and Skolnick, J., J. Phys. Chem., 102 (1998) 4628.
27. Skolnick, J., Zhang, Y., Arakaki, A. K., Kolinski, A., Boniecki, M., Szylagyi, A. and Kihara, D., Proteins 56 (2003) 469.
28. Feig, M., Rotkiewicz, P., Kolinski, A., Skolnick, J. and Brooks, III, C.L., Proteins, 41 (2000) 86.
29. Samudrala, R. and Levitt, M., Protein Sci., 9 (2000) 1399.
30. Bonneau, R., Ruczinski, I., Tsai, J. and Baker, D., Protein Sci., 11 (2002) 1937.
31. Yang, A.-S. and Wang, L.-Y., Bioinformatics 19 (2003) 1267.