Acta Biochimica Polonica

Vol. 51 No. 2/2004

349–371

QUARTERLY

Review

Protein modeling and structure prediction with a reduced representation★

Andrzej Kolinski✉

Faculty of Chemistry, Warsaw University, Warszawa, Poland

Received: 19 April, 2004; accepted: 27 May, 2004

Key words: protein folding, lattice proteins, structure prediction, comparative modeling, Monte Carlo simulations

Protein modeling could be done on various levels of structural details, from simplified lattice or continuous representations, through high resolution reduced models, employing the united atom representation, to all-atom models of the molecular mechanics. Here I describe a new high resolution reduced model, its force field and applications in the structural proteomics. The model uses a lattice representation with 800 possible orientations of the virtual alpha carbon-alpha carbon bonds. The sampling scheme of the conformational space employs the Replica Exchange Monte Carlo method. Knowledge-based potentials of the force field include: generic protein-like conformational biases, statistical potentials for the short-range conformational propensities, a model of the main chain hydrogen bonds and context-dependent statistical potentials describing the side group interactions. The model is more accurate than the previously designed lattice models and in many applications it is complementary and competitive in respect to the all-atom techniques. The test applications include: the *ab initio* structure prediction, multitemplate comparative modeling and structure prediction based on sparse experimental data. Especially, the new approach to comparative modeling could be a valuable tool of the structural proteomics. It is shown that the new approach goes beyond the range of applicability of the traditional methods of the protein comparative modeling.

★Presented as invited lecture at the 29th Congress of the Federation of European Biochemical Societies, Warsaw, Poland, 26 June–1 July 2004.

✉Correspondence: Faculty of Chemistry, Warsaw University, L. Pasteura 1, 02-093 Warszawa, Poland; phone: (48 22) 822 0211, ext. 320; fax: (48 22) 822 5996; e-mail: Kolinski@chem.uw.edu.pl

Abbreviations: CABS, Cα–Cβ–Side group protein model; Cα, alpha carbon; Cβ, beta carbon, cRMSD, coordinate root mean square deviation, DSSP, secondary structure assignment method (by Kabsch and Sander), NMR, nuclear magnetic resonance; PDB, protein data bank; REMC, replica exchange Monte Carlo; UNRES, united residue model (Scheraga and co-workers)

Under proper conditions of solvent and temperature a majority of globular proteins fold to a unique three-dimensional structure (Anfinsen, 1973; Anfinsen & Scheraga, 1975). Usually, the process is reversible and takes milliseconds to minutes, depending on the protein size and structural complexity (Brooks *et al.*, 1998). In spite of the rapid progress in the computing technology (computing speed increases approximately two times every 1.5 year) a brute force approach to computational modeling of protein folding, that treats all degrees of freedom (and the surrounding solvent) in an explicit fashion, remains impractical (Hansmann & Okamoto, 1999). This is due to the enormous size of the protein conformational space (Jernigan, 1992; Wolynes *et al.*, 1995; Hardin *et al.*, 2002), complex interactions and topological obstacles involved (Alm *et al.*, 2002). Only local relaxation processes and folding of small polypeptides is accessible by the traditional all-atom molecular mechanics (Ding *et al.*, 2002). Knowledge of a protein three-dimensional structure is a key to understanding the protein biological function, to the rational drug design, protein engineering, etc. (Skolnick *et al.*, 2000; Simons *et al.*, 2001; Chance *et al.*, 2002). Experimentally, protein structures could be determined *via* the X-ray crystallography or protein NMR (Montelione *et al.*, 2000; Chance *et al.*, 2002). Other experimental techniques are presently of a lesser practical importance. Due to the time-consuming techniques, large cost, and other factors we know now about 30 thousands of protein structures. This is only a small fraction (about 0.001) of the number of known protein sequences, resulting from the systematic sequencing of numerous genomes (Klose, 1989; Arnold & Hilton, 2003) of living organisms of various complexity, from viruses to humans (Lander *et al.*, 2001; Harrison & Gerstein, 2002; Cherkasov & Jones, 2004). The above explains a need for new molecular modeling tools that can facilitate study of the protein dynamics (Kolinski *et al.*, 2003a) and thermodynamics and extend the possibility of the *in silico* protein structure prediction (Mirny *et al.*, 2000; Baker & Sali, 2001; Vajda *et al.*, 2002; Zacharias, 2003; Cherkasov & Jones, 2004; Kihara & Skolnick, 2004).

Protein modeling could be made more tractable by reducing the number of explicitly treated degrees of freedom and by simplification of computations of the intramolecular and intermolecular interactions (Miyazawa & Jernigan, 1985; Shakhnovich, 1997; Kolinski *et al.*, 2001). A number of reduced models of proteins were proposed in the past (Levitt, 1976; Sun, 1993; Monge *et al.*, 1995; Kolinski *et al.*, 2000; Betancourt, 2003). Some of them employed a continuous space representation of the protein conformational space, other confined the protein to a discrete grid, or lattices (Hinds & Levitt, 1992; Godzik *et al.*, 1993a; Hinds & Levitt, 1994; Kolinski & Skolnick, 1996; Sun *et al.*, 1999; Kolinski *et al.*, 2000; Kolinski & Skolnick, 2004). A typical example of the first approach is the classical model proposed almost 30 years ago by Levitt and Warshel (1975). This approach assuming a reduced Cα representation of the main chain and a single united atom representing the side groups was followed (with various modifications) by others (Levitt, 1976; Hagler & Honig, 1978). Some continuous space reduced models assumed an all atom representation of the main chain backbone and a single united atom for the side group (Sun, 1993). Probably, one of the most advanced (or the most advanced) continuous reduced model has been proposed by Scheraga, Liwo and coworkers (Lee *et al.*, 1999; 2001; Pilardy *et al.*, 2001; Liwo *et al.*, 2002). This UNRES (UNited RESidues) model incorporates more realistic intraprotein interactions, including cooperative hydrogen bonds and flexible ellipsoidal side chains. Predictive power of these simplified continuous models varies from a possibility to find an overall correct topology (with generally wrong structural details) within a number of low energy structures to a low and

moderate resolution correct structures of small and simple proteins.

Interestingly, the early lattice models of real proteins (we defray here from discussion of protein-like simple lattice models and Go-type models (Go *et al.*, 1980)) exhibited a similar level of the structure prediction ability to that of the simple off-lattice, continuous models (Covell & Jernigan, 1990; Crippen, 1991; Covell, 1992). Designed in early 90' higher resolution models were capable to predict moderate and low resolution models of small globular proteins (Kolinski *et al.*, 1993; Kolinski & Skolnick, 1994a; 1994b; 1996; 1997b). The interaction schemes of these models (Godzik *et al.*, 1995) based solely on the knowledge-based potentials derived from a statistical analysis of various structural regularities (Godzik *et al.*, 1993) seen in the known three-dimensional structures of globular proteins (Kolinski & Skolnick, 1998b; Miyazawa & Jernigan, 1999; 1999a; 1999b). It is worth to mention that a properly designed Monte Carlo dynamics scheme for lattice models mimics surprisingly well Molecular Dynamics (or Brownian Dynamics) of the otherwise equivalent continuous reduced models (Rey & Skolnick, 1993; 1994). However, the lattice models facilitate significantly faster simulations (Kolinski & Skolnick, 2004). Lattice moves could be designed in such a way that the local conformational transitions occur between local minima of the model's energy landscape – the local energy barriers could be easily surmounted. Moreover, due to the finite number of the local conformations various energy terms for the lattice models could be calculated "in front", once for all simulations, and revoked by simple look-up procedures during the proper simulations. For instance, (just to mention a simplest possibility) all possible coordinates of the β-carbons could be stored as a function of identities (indices) of two successive virtual Cα–Cα bonds, provided the Cα positions are restricted to a lattice. As a result, simulations of the lattice models are usually a couple of orders of magnitude faster than simulations employing equivalent continuous models. Thus longer relaxation processes or/and bigger systems are computationally accessible (Kolinski & Skolnick, 1997a; 1997b; 1998; Kolinski *et al.*, 2001; Kolinski *et al.*, 2003a).

In this work I describe a version of recently developed high resolution lattice model and its applications. The model assumes a lattice-confined Cα representation of the main chain backbone, with 800 possible orientations of the Cα–Cα vectors. The lattice spacing of the underlying simple cubic lattice (sc) is assumed to be eual to 0.61 Å. Consequently, the α-carbon trace of a PDB (Protein Data Bank (Bernstein *et al.*, 1977)) structure of a globular protein could be projected onto this lattice with the average accuracy range of 0.35 Å. This is better than accuracy of the high resolution crystallographic structures. Actual accuracy of the model is lower due to inaccuracies in the model interactions. The model assumes four united atoms (interaction centers) per residue: α-carbon, center of the virtual Cα–Cα bond (serving as a center of interactions for the peptide bonds), Cβ and the center of mass of the side-group (where applicable). While the coordinates of the α-carbons are restricted to the underlying sc lattice, the coordinates of the remaining united atoms are off-lattice and are defined by the Cα-coordinates and the amino acid identities. The force-field of the CABS model (an acronym for CA-B and Side group) consists of several potentials that mimic averaged interactions in globular proteins. The solvent is treated in an implicit fashion. The new model can be used in studies of the protein dynamics and thermodynamics (Kolinski & Skolnick, 2004), including *in silico* protein folding leading to the *ab initio* prediction of protein structures (Skolnick *et al.*, 2003) and protein–protein interactions. In this paper I describe a test application to the loop-modeling, a typical task of the comparative modeling (Sali, 1995) of protein structures in the range of significant sequence similarity,

where a high fidelity of a modeling protocol is strongly required (Schonbrun *et al.*, 2002).

DESIGN OF THE MODEL AND ITS FORCE FIELD

Representation of the protein conformational space of the present model is relatively simple and could be easily reproduced basing on the description given in the next section and in an earlier published work (Boniecki *et al.*, 2003). Significantly more complex is the design of the force field, which needs to correct for the reduced representation and an implicit treatment of the solvent. The derivation of all components of the new interaction scheme is outlined in detail. Due to large size of the files containing numerical data for the histogram-type potentials only some examples are provided here. The full set of the force field parameters could be viewed and downloaded from our homepage: www.biocomp.chem.uw.edu.pl

Protein representation

The framework for the protein representation and the conformational updating during the Monte Carlo simulations consist of a pseudochain of the α-carbons confined to the underlying sc lattice with the lattice spacing equal to 0.61 Å (see Fig. 1). This particular value of the lattice spacing has been selected taking into consideration the assumed resolution of the model (the average accuracy of the Cα trace representation is equal to about 0.35 Å) and some other factors related to the range of interactions, the protein geometry and the computational efficiency and feasibility. Nevertheless, the choice of the lattice spacing is to some extent arbitrary – a related earlier model (Li *et al.*, 2003; Skolnick *et al.*, 2003) assumed the lattice spacing equal to 0.87 Å. The virtual bonds connecting the Cαs have a form of vectors with integer coordinates $\mathbf{v} = [\pm i, \pm j, \pm k]$. The length of these vectors $|\mathbf{v}|$ is restricted to the following range: $29 \leq |\mathbf{v}|^2 \leq 49$ (in the lattice units). This implies that the number of possible Cα–Cα vectors is equal to 800 and the length of the vectors varies between 3.28 Å and 4.27 Å, with the average value very close to the Cα–Cα distance seen in the real proteins, which is equal to 3.78 Å. The fluctuating bond length of the model ensures efficient chain dynamics and prevents effects of various lattice artifacts. For instance, the lattice anisotropy, characteristic for the low resolution lattice protein models (Godzik *et al.*, 1993a; Reva *et al.*, 1996), is undetectable in the present representation.

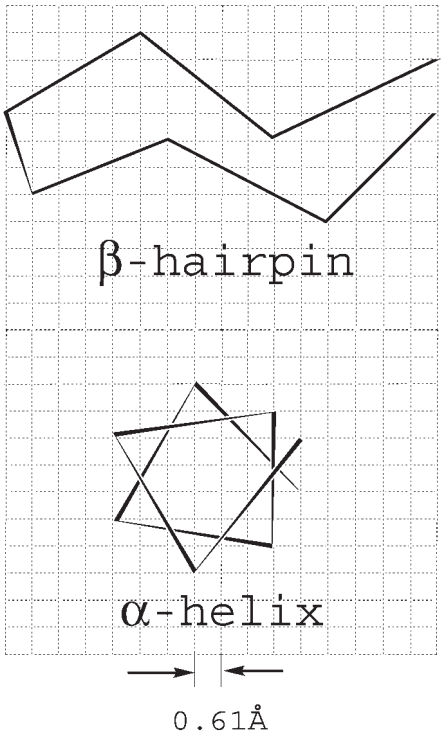

Figure 1. Short fragments of alpha-carbon traces representing a β-sheet and an α-helix confined to the underlying simple cubic lattice with lattice spacing equivalent to 0.61 Å in real proteins.

Positions of three consecutive α-carbons define rather precisely position of the β-carbon for the central residue. The β-carbons are located off-lattice. The position of the center of the remaining portion of the side group corresponds to the most probable (over the database of protein structures) rotamer, given the

type of the main chain conformation, which is assigned as expanded (presumably β-strand or expanded loop) or compact (a helix or a turn), depending on the value of the planar angle of the Cα-trace. Similarly, to the design of the earlier cited UNRES model (Lee *et al.*, 1999), here it is also introduced an additional "united atom" located in the center of Cα–Cα virtual bond. It supports the definition of the main chain hydrogen bonds. The idea of the reduced representation employed in this work is illustrated in Fig. 2.

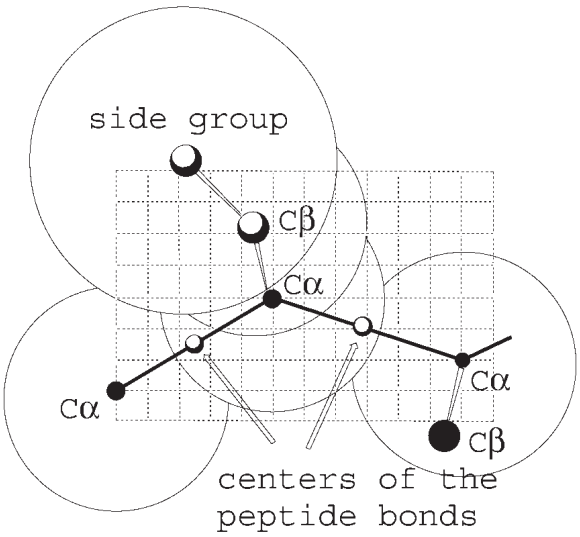

Figure 2. Schematic drawing of the reduced representation of a fragment of a protein chain.

The large open spheres illustrate approximate range of interactions between non-bonded united atoms, except for the centers of the peptide bonds. See the text for more details.

Sampling scheme

Monte Carlo conformational updating of the model chain employs several types of local micromodifications of the α-carbon trace associated with proper displacements of the side chain united atoms. Various micromodifications occur in randomly selected locations and the resulting changes of the local geometry are selected in a random fashion. A single step of the simulation algorithm (a time unit of the Monte Carlo dynamics) consists of 2 attempts to the end moves, 10(N-2) attempts to the two-bond moves, N-3 attempts to the three-bond moves, N-24 attempts to the small distance "rigid-body" type modifications (where the size of the effected portion of the chain is a random variable) and N-24 "reptation" type moves, where a "bubble" on the α-carbon trace is annihilated in one spot and randomly created somewhere else along the chain (the residues between the "bubbles" move down the chain contour without a change of the local geometry of the α-carbon trace). The larger scale moves (the "rigid body" and "reptation") extend over portions of the model chain consisting of 4 to 22 residues. Figure 3 shows examples of the all

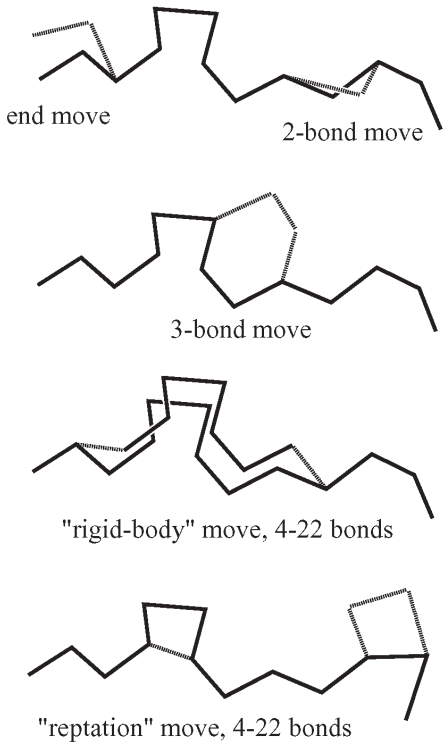

Figure 3. Examples of random micromodification of the model chain conformation.

The positions of the side chain united atoms are defined by the conformation of the main chain, and therefore their changes are not shown for the sake of clarity.

types of the local micromodifications employed in the simulation algorithm of our model. Obviously, due to the excluded volume

clashes the larger scale moves are less frequently successful. The most local moves (two-bond) are attempted more frequently to allow for an equilibration after the larger scale updates. All moves are subject to various conformational restrictions described in the next section.

The asymmetric Metropolis Monte Carlo scheme controls the simulation process (Metropolis *et al.*, 1953). Depending on a purpose, the simulations are carried out in the isothermal conditions, are subject to a simulated annealing procedure or are controlled by the Replica Exchange Monte Carlo (REMC) scheme (Swendsen & Wang, 1986). The REMC technique is significantly more efficient in finding the global energy minima in systems (Hansmann & Okamoto, 1999; Gront *et al.*, 2000) with extremely complex conformational energy landscape (which is exactly the case for the present model). Thus, REMC (and its variants) is a method of choice in the *ab initio* folding (Skolnick *et al.*, 2003), folding with experimental restraints (Li *et al.*, 2003), comparative modeling (Bujnicki *et al.*, 2001; Kihara *et al.*, 2001; Kolinski *et al.*, 2001; Rotkiewicz *et al.*, 2001) and similar applications (Boniecki *et al.*, 2003), where the main goal is to find the global minimum of the conformational energy (Kolinski *et al.*, 2003b), regardless of a rather nonphysical dynamics or distorted the folding pathway.

Generic, sequence independent short range interactions

The lattice-confined chain of the CABS representation is very flexible. Its average distributions of the local conformational characteristics generated in a MC run are far from that seen in real proteins. This is mostly due to the lack of atomic details, resulting in inaccurate excluded volume effects and lack of the specific rotational restrictions of the polypeptide chains. Thus, our first goal is to design a set of sequence independent potentials which correct for a majority of the deficiencies of the reduced representation. These conformational biases are outlined below and a reason for the line of design is provided for each of them. The guideline for these generic interactions comes from the distribution of conformations seen in the real proteins, taken from a statistical analysis of the solved PDB structures. An implicit underlying assumption is that the local conformational characteristics seen in folded structures are not far from the related characteristics of the polypeptide chains in their denatured state. While it is just an assumption, there are numerous reasons that this is a legitimate working hypothesis (Godzik, 1996; Rooman & Gilis, 1998; Mohanty *et al.*, 1999; Shimada *et al.*, 2000; Kuznetsov & Rackovsky, 2002; Evers *et al.*, 2003).

Planar angle restrictions. Figure 4 shows a short fragment of the Cα-trace of the CABS model and provides an explanation of the notation used in this section. In a protein the planar angle for the Cα-trace is restricted due to various short-range (between atoms close along the chain) interactions. Due to the fluctuating length of the virtual backbone vectors **v,** it is better to translate the angular restric-

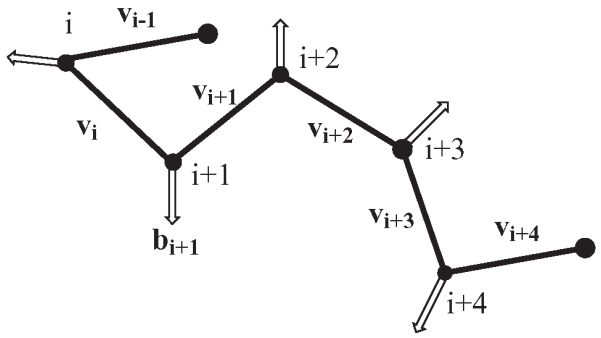

Figure 4. Illustration of the abbreviations used in the definitions of short range conformational biases described in the text, where $\mathbf{r}_i$ stands for the Cartesian coordinate of the i-th Cα and $\mathbf{v}_i$ is the corresponding vector of the virtual Cα chain.

The open arrows indicate "normalized" bisector vectors **b** of the planar angles.

tions onto a restriction on the distances between i-th and i+2nd Cαs. The resulting condition is:

$$4.1 \text{ Å} \leq | \mathbf{r}_{i+2} - \mathbf{r}_i | \leq 7.4 \text{ Å} \quad (1)$$

where $\mathbf{r}_i$ denotes the Cartesian coordinates of the i-th bead (Cα) of the chain. This roughly translates onto $70^o - 150^o$ range for the planar angles in the peptides. The symbol "$| \mathbf{r} |$" in the above formula means the length of the vector $\mathbf{r}$.

Biases towards protein-like chain stiffness. The distribution of the end-to-end distances for a short four-bond generic CABS chain is close to the Gaussian-type distribution. In proteins the corresponding distribution is bimodal, with the short distance peak corresponding to the compact (mostly helical) conformations and the long distance, a more diffused peak corresponding to the expanded conformations. Such distribution can be enforced by the following simple potentials:

$$B_B = 0.5 \times f \times \varepsilon_g \quad (2)$$

when: $(\mathbf{v_{i-1}} \cdot \mathbf{v_{i+3}}) < 0$ and
$| \mathbf{r_{i+4}} - \mathbf{r_i} | < 7.2$ Å,
or when: $(\mathbf{v_{i-1}} \cdot \mathbf{v_{i+3}}) < 0$ and
$| \mathbf{r_{i+4}} - \mathbf{r_i} | > 11.0$ Å

where: symbol "$\cdot$" denotes the dot product of two vectors, ε_g is the scaling factor common for the all "soft" short-range energetic biases B and f is a scaling factor applied to the single domain globular proteins, and reflecting a different flexibility of a polypeptide chain in the protein core and in the surface loops. It scales with the radius of gyration of a globular protein in the following way:

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

where: S is the radius of gyration of the folded protein (assuming a close to spherical shape of the single domain proteins the value of S can be computed from the number of residues in the chain (Kolinski *et al.*, 1993) and s is the mean square distance of the center of mass of a chain fragment from the center of mass of the polypeptide chain in its actual conformation. For the multidomain proteins it is assumed that f = 1 for all residues. In the cases of a comparative modeling or a restrained folding it is also convenient to set f = 0. However, this position dependent scaling of the conformational biases plays some role in the *ab initio* folding simulations, accelerating the chain collapse.

Let us define a normalized "bisector" vectors $\mathbf{b_i}$ of the planar angles:

$$\mathbf{b_i} = (\mathbf{v_{i-1}} - \mathbf{v_i}) / | (\mathbf{v_{i-1}} - \mathbf{v_i}) | \quad (4)$$

and a "normalized" sum of four consecutive vectors $\mathbf{b}$:

$$S4 = \max\{ |(\mathbf{b}_{i+1} + \mathbf{b}_{i+2} + \mathbf{b}_{i+3} + \mathbf{b}_{i+4})|, 0.5\} - 0.5 \quad (5)$$

In protein structures the i-th and i+2nd $\mathbf{b}$ vectors are either almost parallel (expanded conformations) or almost antiparallel (compact, mostly helical conformations). The neighboring vectors are non-parallel. The sum S4 has a "small" value for the majority of protein fragments. Exceptionally, it assumes a larger value in some loops. Therefore, the largest values of the sum were cut to the value of 2.0. These regularities can be encoded in the following potential, which propagates the local protein-like stiffness for a somewhat larger distance down the chain:

$$B_S = \min \{2.0, S4\} \times 0.5 \times f \times \varepsilon_g \quad (6)$$

when: $(\mathbf{b}_{i+1} \cdot \mathbf{b}_{i+2}) + (\mathbf{b}_{i+2} \cdot \mathbf{b}_{i+3}) < 0.25$

$B_S = 0$ otherwise.

Bias towards regular secondary structure. Helices in proteins are usually right handed, with a characteristic value of the distance between the residues along a helix. β-Strands have usually an up-and-down geometry. These conformational properties are formalized in the following potentials:

$$B_H = -0.5 \times f \times \varepsilon_g - \varepsilon_g \quad (7)$$

for: $|\mathbf{r_{i+4}} - \mathbf{r_i}| < 7.2$ 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 $(\mathbf{v_{i+1}} \cdot \mathbf{v_{i+3}}) < 0$ and $(\mathbf{v_i} \cdot \mathbf{v_{i+3}}) > 0$

$$B_E = -0.5 \times f \times \varepsilon_g - \varepsilon_g \quad (8)$$

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 $(\mathbf{b_{i+1}} \cdot \mathbf{b_{i+2}}) < 0$,
and $(\mathbf{b_{i+2}} \cdot \mathbf{b_{i+3}}) < 0$

The scaling factor f has the same meaning as it had in the previously defined potentials, reflecting a higher propensity towards a regular secondary structure in a protein core.

When the secondary structure is dependably assigned (predicted, or taken from a template) a longer fragment could be biased towards a proper geometry, provided that the entire stretch has the same helical or expanded secondary structure assignment:

$$B_{HH} = \delta \times (0.25 \times d \times \varepsilon_g + 0.5 \times \varepsilon_g) \quad (9)$$

for residues i-th to i+7 assigned as helical where: $d = \text{abs}(|\mathbf{r_{i+7}} - \mathbf{r_i}| - 10.75$ Å), and $\delta = 0$ for $d < 0.61$ and $\delta = 1$ elsewhere,

$$B_{EE} = \delta \times (0.25 \times d \times \varepsilon_g + 0.5 \times \varepsilon_g) \quad (10)$$

for residues i-th to i=6 assigned as expanded where: $d = \text{abs}(|\mathbf{r_{i+6}} - \mathbf{r_i}| - 19.1$ Å $) - 1.0$ Å, and $\delta = 0$ for $d < 1.22$ Å and $\delta = 1$ elsewhere.

In the above definitions, 10.75 Å is the average distance between the i-th and the i+7th residues in helices, while 19.1 Å is the average distance between the i-th and the i+6th residues in β-strands. The different cut-off values for B_{HH} and B_{EE} are related to the different variability of the geometry of helices and sheets, respectively.

Bias against "crumpled" structures. An another structural property could be used to regularize the chain conformation and to speed-up the folding simulations by avoiding a non physical local geometry. Namely, highly folded "crumpled" conformations, where the U-turns changing the direction of the chain propagation are very close to each other along the chain, are extremely rare in the protein structures. This is encoded in the following potential, where the minimal length of a fragment between turns (including a turn length) is assumed to be larger than 5 residues:

$$B_C = 4.0 \times \varepsilon_g \quad (11)$$

for: $(\mathbf{r_{i+5}} - \mathbf{r_i}) \cdot (\mathbf{r_{i+10}} - \mathbf{r_{i+5}}) < 0$
and $(\mathbf{r_{i+15}} - \mathbf{r_{10}}) \cdot (\mathbf{r_{i+5}} - \mathbf{r_i}) > 0$

The value of the scaling factor 4.0 is an arbitrary one.

The total energy of the sequence independent (generic) short-range interactions is equal to the sum of the all components:

$$E_g = \Sigma (B_B + B_S + B_H + B_E + B_{HH} + B_{EE} + B_C) \quad (12)$$

where Σ denotes summation along the model chain.

The energy parameter ε_g has the same value for all contributions and its value has to be optimized for a proper balance between the short-range and the long-range interactions in the model. In Monte Carlo simulations controlled just by the above described generic protein-like conformational propensities the model chains (when cooled to a low temperature) adopt conformations with short helical and expanded fragments of a fluctuating length. The design of such protein-like potentials provides a good background of the energy landscape. As a result, the sequence specific potentials are capable to trigger formation of native-like local and global structures. In other words, the generic potentials reduce the conformational entropy, and restrict the conformational space to be searched in the simulations. Moreover, such potentials form a scaffold of a funnel in the energy landscape enabling for a lesser specificity of the sequence-dependent interactions. This seems be a very important feature, since a high specificity of the sequence dependent force field is very difficult to achieve due to the enormous number of various, often competing, interactions in proteins. The idea of such structure-regularizing potentials is typical for our previous work, however recently it became more and more explored by others in various contexts of the protein modeling and structure prediction.

Sequence dependent short range interactions

The sequence dependent short-range potentials were derived basing on a statistics of a non redundant database of the three-dimensional protein structures. These statistical potentials of the mean force reflect a relative frequency of observation of a given local geometry for given pairs of amino acids, in respect to the random amino-acid composition and the random distribution of distances. Three types of potentials contribute to the sequence specific interactions:

$$E_{13} = E_{13}(|\mathbf{r_3} - \mathbf{r_1}|, A_3, A_1) \quad (13)$$

$$E_{14} = E_{14}(|\mathbf{r_4} - \mathbf{r_1}|^*, A_3, A_2) \quad (14)$$

$$E_{15} = E_{15}(|\mathbf{r_5} - \mathbf{r_1}|, A_4, A_2) \quad (15)$$

where A_i is the identity of the amino acid at the i-th position along the chain.

These potentials have the form of histograms, separate for the all possible pairs of amino acids. The E_{13} potentials have 8 bins of $\mathbf{r}_{13}$, from 0 to 8 Å (conformations corresponding to the first four bins are not observed in the database, and an arbitrary high value of the energy is assigned to them), the E_{14} potentials have 24 bins of $\mathbf{r^*}_{14}$, from –12 to 12 Å, where the sign of the "distance" denotes the chirality (left handed conformations are assigned to the negative part of the histogram and the right-handed to the positive part) and the E_{15} potentials have 16 bins of $\mathbf{r}_{15}$, from 0 to 16 Å (again the four first bins are prohibitive). In principle, these potentials should be three amino acid dependent, four and five amino acid dependent, for the E_{13}, E_{14} and E_{15}, respectively. However, the statistics would be too weak for a larger number of amino acids and the resulting data files would be too large. Thus two locations along the relevant fragments were selected, aiming on the highest specificity of the resulting potentials. The full data set for these potentials could be viewed and downloaded from our homepage: www.biocomp.chem.uw.edu.pl and two illustrative examples of the E_{14} potential are given in Fig. 5. The total short-range sequence dependent conformational energy is the weighted sum of the all components along the chain:

$$E_s = \Sigma\,(0.5 \times E_{13} + E_{14} + E_{15}) \quad (16)$$

The lower weighting of the E_{13} components is assumed due to their low specificity – the

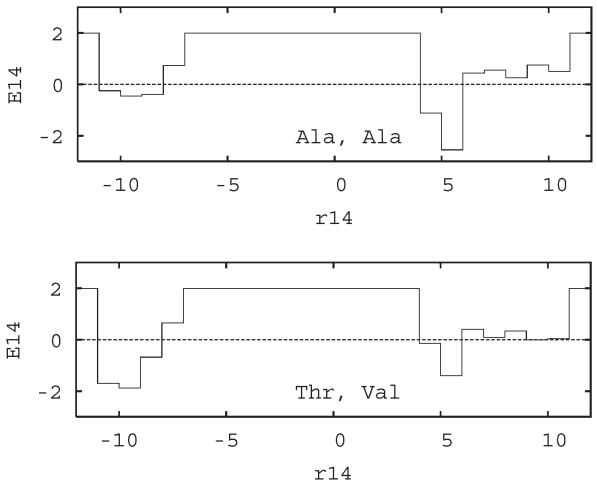

Figure 5. Examples of the E_{14} short range potentials.

The upper panel shows a case of a helical sequence where the helix-forming Ala residue appears on the second and the third position of the four residue (three virtual bonds) fragments. The lowest energy is observed for the range of 5–6 Å, characteristic for helices. The value +2 of the energy corresponds to an arbitrary cut-off for very high or infinite values. The lower panel shows an example of β-sheet forming residues (Thr and Val). A large low energy basin for negative (left handed conformations) values of r14 correspond to the most typical β-sheet conformations.

distances between i-th and i+2 residues do not differ too much for various amino acids. Besides the potentials averaged over the entire database three additional sets of potentials were derived: one for the only helical fragments, one for the sheets and one for the coils. In the cases when the predicted or known secondary structure is available the secondary structure specific potentials can be used for the entire chain or for its fragments. In the case of predicted secondary structure it proven to be productive to use a 50/50 average of the secondary structure specific and the averaged potentials to allow for a correction of the secondary structure prediction errors during the simulations.

Hydrogen bonds

Main chain hydrogen bonds imply a specific geometry of the Cα trace. Due to the reduced representation assumed in the CABS model the H-bonds are defined as directional interactions between the alpha carbons. This approach requires a "renumbering" of the interacting residues. For example, the hydrogen bond between the i-th and i+4th residues in helices is replaced by the Cα–Cα interactions between the i-th and the i+3th Cαs. Figure 6 illustrates such an ersatz of the hydrogen bonds used in the force field of the CABS model. The vectors **h** are orthogonal to the planes formed by the two consecutive Cα–Cα vectors, and their length is equal to 4.6 Å. Residues i and j are considered to be "hydrogen bonded" when the following set of geometrical conditions are fulfilled:

$$
\begin{aligned}
&|\mathbf{r_i} - \mathbf{r_j}| < 6.1 \text{ Å} \\
&\text{abs}(\mathbf{h_i} \cdot \mathbf{h_j}) > 16 \quad (\text{in Å}^2) \\
&(\mathbf{v_{i-1}} \cdot \mathbf{v_{j-1}}) > 0 \quad \text{and} \quad (\mathbf{v_{i+1}} \cdot \mathbf{v_{j+1}}) > 0 \\
&\text{or } (\mathbf{v_{i-1}} \cdot \mathbf{v_{j+1}}) < 0 \quad \text{and} \quad (\mathbf{v_{i+1}} \cdot \mathbf{v_{j-1}}) < 0 \\
&|\mathbf{r_i} - \mathbf{r_j}| - |\mathbf{h_i}| < 1.83 \text{ Å} \qquad (17)
\end{aligned}
$$

The first line defines the cut-off distance for the "hydrogen bonded" residues measured by the corresponding Cα–Cα distance. The second line defines a directional cut-off for the vectors **h**. The angle between two vectors has to be smaller than 40^o or greater than 140^o, i.e. these vectors have to be almost parallel or almost antiparallel, respectively. The third line says that the interacting two bond fragments of the backbone have to be in a roughly parallel (as in helices and parallel β-sheets) or roughly antiparallel mutual orientation (as in antiparallel sheets). The last condition means that the vector $\mathbf{h_i}$ needs to coincide approximately with the relevant Cα–Cα vector. The definition (see Fig. 6) allows for up to two "hydrogen bonds" per one Cα (as in real proteins). The hydrogen bond interactions between the i-th and i+4th Cαs are forbidden in this model. Such a restriction leads to a better geometry of the helices and the tight turns. The strength of the model hydrogen bonds is distance (the first component) and angle dependent (the second component):

$$E_h = \delta_h \times \varepsilon_h \times (1.0 + (4.25/\max\{4.25,\min\{6.01, r_{pp}\}\})^4 - 0.25 + (4.25/\max\{4.25,\min\{6.01, r_{qq}\}\})^4 - 0.25) + \delta_\gamma \times \varepsilon_\gamma \times (2.0 - \max\{(\mathbf{b_i} \cdot (\mathbf{r_i} - \mathbf{r_j})/6.1)^2, 0.125\} - \max\{(\mathbf{b_j} \cdot (\mathbf{r_i} - \mathbf{r_j})/6.1)^2, 0.125\}) \quad (18)$$

where: r_{pp} and r_{qq} denote the proper distance between the centers of the peptide bonds connected with the Cαs of interest (see Fig. 6),

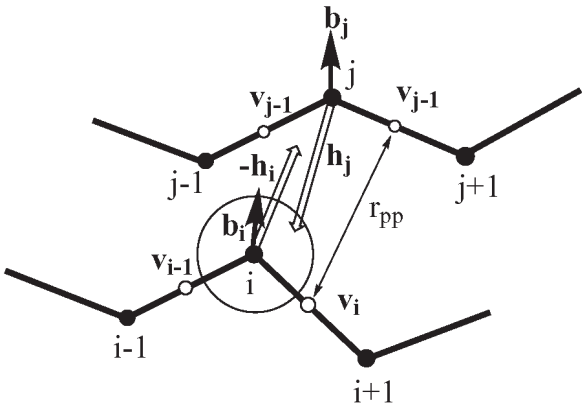

Figure 6. Explanation of the geometry of the model hydrogen bonds.

and 4.25Å is the value of this distance in the hydrogen-bonded elements of a regular secondary structure. $\delta_h = 1$ when all the above listed geometrical condition are satisfied and $\delta_h = 0$ otherwise. $\delta_\gamma = 1$ when the first three conditions are satisfied and $\delta_\gamma = 0$ otherwise. The scaling factors were optimized and their values are: $\varepsilon_h = -1.25$ and $\varepsilon_\gamma = -0.25$. The criterion of the optimization was as good as possible correlation of the model H-bonds with H-bonds assignments in PDB structures by Kabsch and Sander (1983) method DSSP. For well folded model structures about 90% of the model hydrogen bonds coincide with the native H-bonds. The total energy of the hydrogen bonds reads as:

$$E_H = \Sigma\Sigma\ (g_{ij} \times E_h) \quad (19)$$

Where g_{ij} is the factor that increases the strength of the hydrogen bonds in elements of regular elements secondary structure (predicted or assigned) and is equal to 1.5 for the intrahelical hydrogen bonds and the β-sheet assigned residues; otherwise, $g_{ij} = 1.0$. Such a scaling increases the success ratio of a correct topology assembly in an *ab initio* folding. It is known, that strongly predicted β-strands are usually buried inside a globule, while the edges of β-sheets are frequently less regular and more difficult to predict. The stronger hydrogen bonds tend to be saturated first. Consequently, proper strands tend to be buried inside a protein structure. The intrahelical hydrogen bonds propagate helices, and the weaker bonds at the ends of helices can break the pattern leading to formation of a turn or loop.

Long range interactions

The long range interactions between the united atoms of the model consist of the hard core repulsions (infinite energy) between Cα and Cβ units, with the cut-off distances given in Table 1, and pairwise interactions of a finite strength. The simulation algorithm detects potential overlaps before all the remaining energy computations to be done. Other interactions are of a finite magnitude.

Table 1. Cut-off distances for the hard-core repulsive interactions

Interaction	Cut-off (Å)
Cα - Cα	3.05
Cα - Cβ	3.66
Cβ - Cβ	3.05

Generic repulsive interactions. Besides the hard-core repulsions there is a soft-core tail as defined below (for the pairs of united atoms that are not the nearest neighbors along the sequence):

$$E_{C\alpha-C\alpha} = \infty \text{ for } d_{C\alpha-C\alpha} \le 3.05 \text{ Å}$$
$$= \varepsilon_r \times ((3.05/d_{C\alpha-C\alpha})^2 - 0.5)$$
$$\text{for } 3.05\text{Å} < d_{C\alpha-C\alpha} < 4.31\text{Å} \quad (20)$$

Similar soft repulsive interactions are applied between the Cαs and centers of peptide bonds:

$$E_{C\alpha-pb} = \varepsilon_r \times ((4.23/\max\{4.23, d_{C\alpha-pb}\})^2 - 0.75) \quad \text{for } d_{C\alpha-pb} \leq 4.88 \text{ Å} \qquad (21)$$

and between Cαs and the side-groups:

$$E_{C\alpha-SG} = \varepsilon_r \times ((3.66/\max\{4.48, d_{C\alpha-SG}\})^2 - 2/3) \quad \text{for } d_{C\alpha-pb} \leq 4.48 \text{ Å} \qquad (22)$$

Sequence dependent long range interactions. The only sequence dependent long-range interactions are these between the side groups, where the center of interactions coincides with the center of gravity of a side chain that includes the alpha carbon (and the all heavy atoms are treated as identical). A very important difference between the CABS force field and other approaches is the context dependent definition of the pairwise interactions.

$$E_{i,j} = \varepsilon_r \text{ for } d_{i,j} \leq D_{min}(A_i, A_j, \Theta_i, \Theta j, \Phi_{i,j}) = \varepsilon(A_i, A_j, \Theta_i, \Theta j, \Phi_{i,j}) \text{ for } D_{min}(A_i, A_j, \Theta_i, \Theta j, \Phi_{i,j}) < d_{i,j} \leq D_{max}(A_i, A_j, \Theta_i, \Theta j, \Phi_{i,j}) = \varepsilon(A_i, A_j, \Theta_i, \Theta j, \Phi_{i,j}) \times (D_{max}(A_i, A_j, \Theta_i, \Theta j, \Phi_{i,j})/ d_{i,j})^2$$
if A_i, A_j are charged (23)

All pairwise interactions depend on:

- the identity of both involved amino acids A_i and A_j,
- conformations of the interacting segments of the main chain, measured by the values of the planar angles of the Cα-trace, Θ_i and Θj (indices Θ_i and Θj assume only two values: open and compact, and the threshold is equal to 6.0 Å for the $r_{i-1,i+1}$ distance.
- and the mutual orientation of the contacting side groups, measured by the value of the product $bb = (\mathbf{b_i} \cdot \mathbf{b_j})$. Three types of contacts are taken into account: parallel ($bb > 0.5$), antiparalle ($bb < -0.5$) and intermediate ($-0.5 \leq bb \leq 0.5$).

The values of the interaction parameters (let us use a short-hand notation) ε (i,j) and the values of the cut-off distances D_{min}(i,j) and D_{max}(i,j) depend on these attributes. The width of the square-well potentials was adjusted as $D_{min}(i,j) = D_{av}(i,j) - 2.0$ Å and $D_{max}(i,j) = D_{av}(i,j) + 0.5$ Å, where the average contact distances of D_{av}(i,j) were extracted from a statistical analysis of the protein database. In this statistical analysis two side groups were assumed to be "in contact" when any pair of their heavy atoms were closer to each other than 4.5 Å. The numerical values of these potentials can be found in our homepage.

Taking into account the mutual orientation of the side groups and the conformation of the main chain fragments involved is extremely important for the specificity of the resulting potentials (Buchete *et al.*, 2004; Kolinski & Skolnick, 2004). For instance, for a pair of two oppositely charged amino acids the averaged potential (as it can be seen in many interaction scales) has a small, close to zero, value. Statistical potentials that take into consideration mutual orientation of the interacting groups have large negative values for the parallel contacts and positive values for the antiparallel contacts. Indeed, in the globular proteins the charged residues interact on the surface of a protein and thereby are almost always parallel. In order to interact in an antiparallel fashion the side groups have to be buried inside the globule. A numerical example for the Lys–Glu interaction parameters is given in Table 2. It shows clearly a rationale for the approach presented in this work. For the purpose of a faster collapse of the model chain the long range pairwise potentials ε (i,j) were shifted by –0.5.

In cases where it is obvious that a protein consist of a single globule the above defined force field could be supplemented with a weak one-body centrosymmetric potential that fa-

Table 2. Contact potential for Lys–Glu interactions

	P	M	A
CC	–0.9	–0.4	0.9
EE	–1.1	–0.4	0.6
CE	–0.2	0.1	0.8
EC	–0.2	0.0	0.8

P, parallel orientation, M, intermediate, A, antiparallel; CC, residues i and j have compact conformations (see the text); EE, residues i and j have open (expanded) conformations; CE, residue i is in a compact conformation, j in an expanded conformation; EC, residue j is in a compact conformation, i in an expanded conformation.

cilitates faster collapse of a globule and consequently a faster folding.

Total energy

The total conformational energy of the CABS protein is a weighted sum of the all components. An optimization of the weights of particular contributions leds to the following results:

- short range sequence-independent interactions (ε_g = 0.75)
- short range sequence-dependent interaction: scaling factor 0.375 (for the sum given in Eqn. 16)
- hydrogen bonds: scaling factor 1 (for the sum given in Eqn. 19)
- repulsive interactions: ε_r = 5.0
- long range pairwise interactions, scaling factor 2.0 (after summation of all pairwise interactions)

The force field outlined above may appear a bit complex. This is a result of a necessity to correct for the reduced representation. A number of the cut-off parameters had to be carefully derived from the statistical analysis of the known protein structures. The necessity of the weighting of particular components of the force field is a result of the partial over-counting of the real physical interactions. In particular, the effect of a solvent is partially accounted for in several potentials (mostly in the interactions of the side chains) in an implicit way. A priori scaling of the all interaction parameters in such reduced models is extremely difficult, if at all possible. After all, a physical behavior of the model provides a justification for the design of its force field.

PREVIOUS APPLICATIONS IN STRUCTURE MODELING

The force field of the CABS model has been recently updated, using a larger and carefully filtered database of protein structures. Some minor details of the hydrogen bond model and the generic short range interactions were also changed. Nevertheless, the basic concepts of the present model are similar to the assumptions of the models used in several earlier applications (Kolinski *et al.*, 1995; 1998a; 1998b; 1999; 2000; 2001; 2003a; Kolinski & Skolnick, 2004). During the CASP5 (Critical Assessment of the protein Structure Prediction) two different clones of the CABS model were used to predict the structures of all the targets of various degree of difficulty, from the comparative modeling, throughout the fold recognition to the new fold category (Skolnick *et al.*, 2003). Interestingly, the reduced representation CABS algorithm performed very well in the comparative modeling category, in many cases providing better molecular models than the models refined by the detailed all-atom algorithms. Moreover, frequently the resulting models were better (closer to the native structure) than any of the templates used.

Recently, we performed a well-controlled experiment of a computational reconstruction of missing fragments of protein structures (Boniecki *et al.*, 2003). Three reduced models (moderate resolution SICHO model, CABS, and REFINER – a model similar in design to the CABS model, however with a continuous space representation of the protein geometry)

were compared with the classical methods of comparative modeling (Sali, 1995) (MODELLER and SwissModel). The reduced models performed much better. The largest gain in accuracy was observed in the cases where relatively large fragments (20 residues or more) were reconstructed, assuming knowledge of the remaining portion of protein structures. Thus it appears that the reduced models, due to their ability to handle a larger scale structural rearrangements in proteins, are complementary (when compared to the more classical approaches) tools of the protein structure modeling and the structure prediction. In this contribution I illustrate an application of the CABS model to several test cases of an "easy" loop modeling, where a high geometrical accuracy is required.

LOOP MODELING: A TEST OF THE APPLICABILITY OF THE REDUCED MODEL IN THE COMPARATIVE MODELING

Comparative modeling is now the most commonly used approach to the theoretical prediction of protein structure (Holm & Sander, 1996; Sali, 1998; 2001; Clark, 1999). Since a protein structure (and function) is evolutionary more conserved than the protein sequence (Clark, 1999), a sequence similarity of a new protein to a protein (or proteins) of known structure implies their structural similarity. A template protein could be identified by a sequence comparison method (Altschul *et al.*, 1997) or a threading algorithm (Godzik *et al.*, 1992; Rost *et al.*, 1997; Jones, 1999; Skolnick *et al.*, 2000; Schonbrun *et al.*, 2002; Ginalski & Rychlewski, 2003; Kihara & Skolnick, 2004). The most conservative approach to the comparative modeling (sometimes called homology modeling, albeit strictly speaking homology is not required for the sequence and structure similarity) assumes that the most conserved is the hydrophobic core of a protein and that the alignment of the query sequence to the template is the most dependable in these regions of a sequence that corresponds to the core of a protein. It is frequently assumed that within the core region a model resulting from the computational modeling can not be better than its template. This is not always true – methods exist that are capable of building models that are closer in the conformational space to the real structure of a query protein than to its template (Kolinski *et al.*, 2001; Kolinski & Skolnick, 2004). Nevertheless, in order to develop such methods one needs first to make sure that a new methodology works well in these more classical examples where the core is well defined and only the loops need to be added to the model (Fiser *et al.*, 2000). This is the purpose of the test computations presented in this work. The procedure is very similar to that used recently by Fiser and Sali (2003) for evaluation of their new algorithm for the loop modeling. Following the main line of this work I selected PDB structures of few representative globular proteins and assumed that the alignments of their sequences to hypothetical templates are the perfect ones within the protein cores, and correspond to the regions of the regular secondary structure elements. Then all the loops were treated as unknown and were rebuild using the CABS model.

The modeling procedure applied here may be outlined as follows:

- Make the DSSP assignment of the secondary structure of a test protein using its PDB coordinates. The data are given in Table 3.
- Assume that all residues assigned to helices (H) or sheets (E) constitute the modeling template and their coordinates are known. The remaining "loops" are treated as unknown.
- Generate a set of the distance restraints for the template. For each Cα within the core five distances to the residues uniformly distributed along the template part of the chain were stored and applied as a

soft restraints keeping this part of molecule within about 0.5 Å from their original positions (a value typical for average accuracy of very good structural alignments), except the residues flanking the loops where the accuracy was range of 1.0–1.5 Å.

- Generate random conformations of the loops.
- Rebuild the loops and relax the protein core using the CABS model and the REMC optimization technique for the conformational search.
- Select the lowest energy structures (from the lowest temperature replica) for the statistical analysis (involving a comparison to the original PDB structures).

Table 3. Sequences and secondary structures of the test proteins

```
1ten:
1---*----10---*----20---*----30---*----40---*----50
LDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLT
____EEEEEE_____EEEEEE_______EEEEEEEE_______EEEEEE
----*----60---*----70---*----80---*----90
EDENQYSIGNLKPDTEYEVSLISRRGDMSSNPAKETFTT
____EEEE_______EEEEEEEEEE__EE___EEEEEE_

256B:
1---*----10---*----20---*----30---*----40---*----50
ADLEDNMETLNDNLKVIEKADNAAQVKDALTKMRAAALDAQKATPPKLED
__HHHHHHHHHHHHHHHHH___HHHHHHHHHHHHHHHHHH_____HHH__
----*----60---*----70---*----80---*----90---*----100
KSPDSPEMKDFRHGFDILVGQIDDALKLANEGKVKEAQAAAEQLKTTRNA
_____HHHHHHHHHHHHHHHHHHHHHHHHHH___HHHHHHHH_HHHHHHHH
----*----110
YHQKYR
HHHHH_

2fdx:
1---*----10---*----20---*----30---*----40---*----50
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILI
_EEEEE____HHHHHHHHHHHHHHH_____EEEEHHH__HHHH____EEE
----*----60---*----70---*----80---*----90---*----100
LGCSAMGDEVLEESEFEPFIEEISTKISGKKVALFGSYGWGDGKWMRDFE
EEE__E___E_____HHHHHHHHHHH____EEEEEEEE_____HHHHHHH
----*----110--*----120--*----130--*----140
ERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIAAI
HHHHH___EE____EEEE______HHHHHHHHHHHH__

2gb1:
1---*----10---*----20---*----30---*----40---*----50
MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATK
_EEEEEE______EEEEEE___HHHHHHHHHHHH______EEEEE_____
----*----60
TFTVTE
EEEEE_

4mba:
1---*----10---*----20---*----30---*----40---*----50
SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS
___HHHHHHHHHHHHHHH_HHHHHHHHHHHHHHH_HHHHHH_________
----*----60---*----70---*----80---*----90---*----100
VADIKASPKLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGV
HHHHH____HHHHHHHHHHHHHHHHHH___HHHHHHHHHHHHHHHHHH__
----*----110--*----120--*----130--*----140--*----150
GSAQFENVRSMFPGFVASVAAPPAGADAAWTKLFGLIIDALKAAGA
_HHHHHHHHHHHHHHHH________HHHHHHHHHHHHHHHHHH___
```

Actually, the last step could be modified to improve the prediction accuracy. Namely, one can perform a clustering of the all structures from the folding trajectory (Betancourt & Skolnick, 2001), next rebuild the atomic details (Feig *et al.*, 2000) of the cluster's centeroids, and then select the best structure using various bioinformatics tools for evaluation of the structure quality. Here, however the simplest possible energy-based criterion is used. Therefore the presented results provide rather an upper bond for the prediction errors.

Simulations were done for five small proteins selected from various structural classes. These proteins contain significant fraction of the loop residues – up to 46% (41 residues) in the case of the 1ten structure. All loops in a protein were modeled simultaneously. Thus, the present method goes significantly beyond the range of applicability of the recently proposed all-atom procedures. At the same time the accuracy (see Table 4) of the reduced approach is high and similar to the limits of performance of the best existing algorithms. Table 4 provides the cRMSD values of the alpha carbons only. The average accuracy of the all backbone atoms is very similar, usually slightly better, although somewhat dependent on a method of the full backbone reconstruction. Figures 7 and 8, and the left side panel of Fig. 9 show five examples of the obtained models optimally superimposed onto the corresponding crystallographic structures. Clearly, most structural details of the loops have been reproduced with a good accuracy. When only a single loop is modeled (assuming known coordinates of the remaining portion of the structure) the accuracy is always (usually significantly) higher. Simulations for the longest loop of 2gb1 fold is a good example (see the right side panel of Fig. 8).

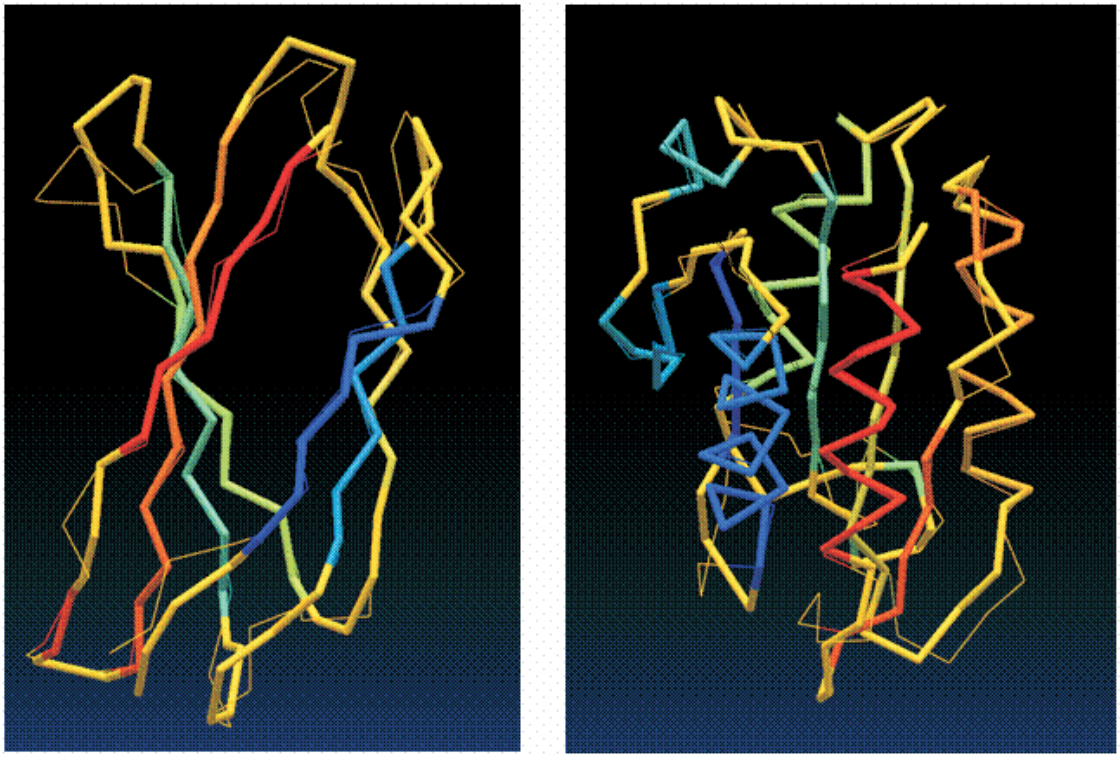

Figure 7. Models (thick lines) of 1ten and 2fxd in the alpha carbon representation superimposed onto crystallographic structures (thin lines).

The colored portions of the chains correspond to the proteins' cores. The yellow fragments represent the loop regions.

Table 4. Statistics for the loop modeling

Name	type	N	N_L	N_{max}	cRMSD (Å)		
					all	*core*	*loops*
1ten	β	89	41	7	1.67	0.54	2.28
256B	α	106	22	7	1.28	0.42	2.32
2fdx	α/β	138	50	6	1.58	0.49	2.17
2gb1	$\alpha+\beta$	56	21	6	1.21	0.57	1.69
2gb1	(single loop modeling)			6	0.81	0.47	1.23
4mba	α	146	34	8	1.34	0.60	2.45

N, protein length (number of residues); N_L, total number of the modeled loop residues; N_{max}, the longest loop in a model; cRMSD, coordinate root mean square deviation from the native structure after the best superimposition; *all*, cRMSD for entire model after the best superimposition with the crystallographic structure; *core*, cRMSD for the core part of the model after best superimposition of the core; *loops*, cRMSD for the all loop residues of the model after best superimposition of the core structure.

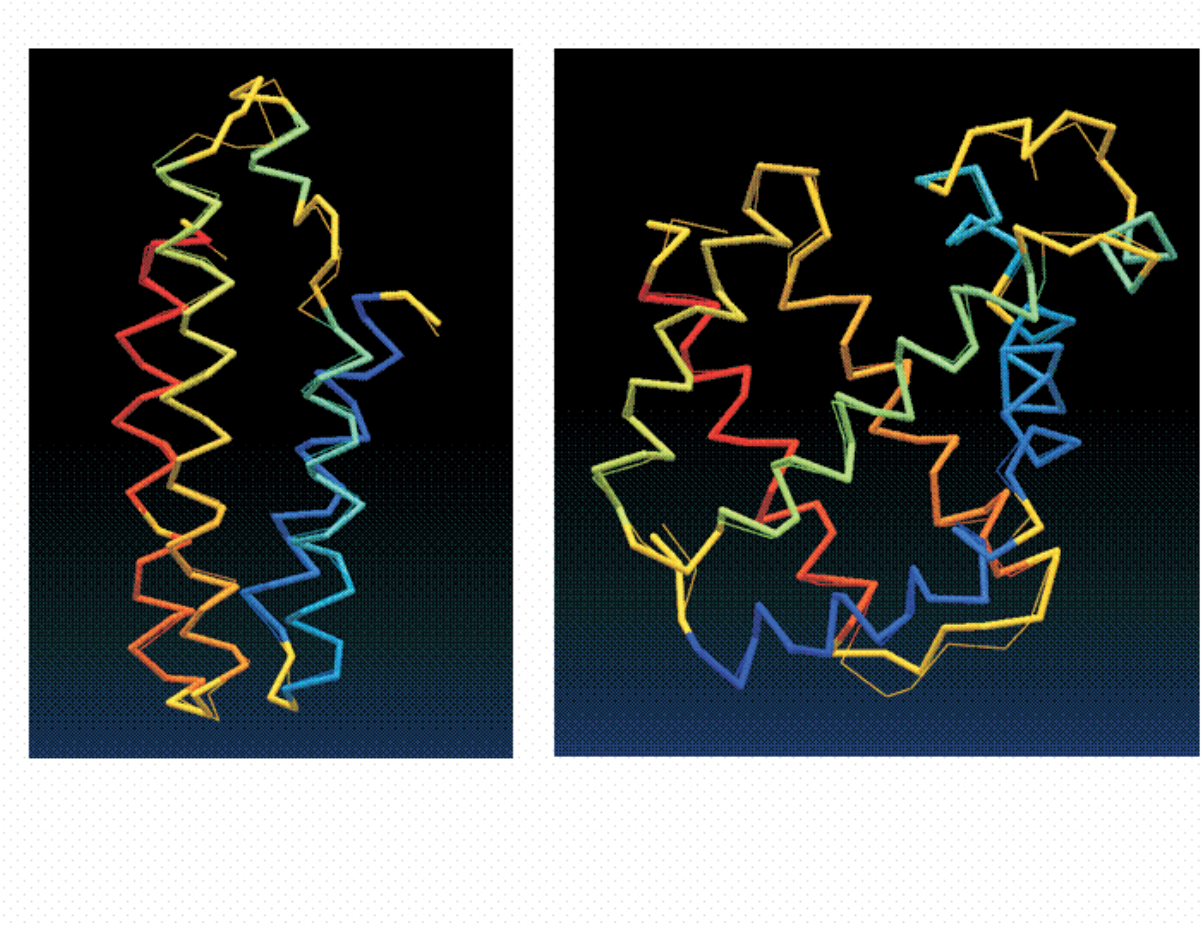

Figure 8. Models (thick lines) of 256B and 4mba in the alpha carbon representation superimposed onto crystallographic structures (thin lines).

The colored portions of the chains correspond to the proteins' cores. The yellow fragments represent the loop regions.

CONCLUSION

This work describes a model of protein structure and dynamics. Protein representation has been reduced to up to four united atoms per residue: Cα, Cβ, center of mass of the remaining portion of a side chain and the center of the virtual Cα–Cα bond. The force field of the model employs knowledge based potentials derived from the statistical analysis of the structural regularities seen in the solved protein structures. The general assumptions of the model of interactions are similar to those of our earlier high coordination lattice models, however the details of the refined force field differ significantly and enable a more accurate modeling.

The examples of the loop modeling in proteins described in this work show that the reduced model described here is an efficient tool of the comparative modeling. Its accuracy is similar to the accuracy of the more traditional methods (Fiser *et al.*, 2000; Fiser & Sali, 2003), however due to the high sampling efficiency the range of applicability of the present approach is significantly broader. Much larger systems (in the present case a larger number of loop residues modeled simultaneously) could be efficiently treated. Recently, we demonstrated that using the CABS model it is possible to predict, with the acceptable accuracy, loops containing up to 20–30 residues each (Boniecki *et al.*, 2003). Applications of the CABS model are not limited to

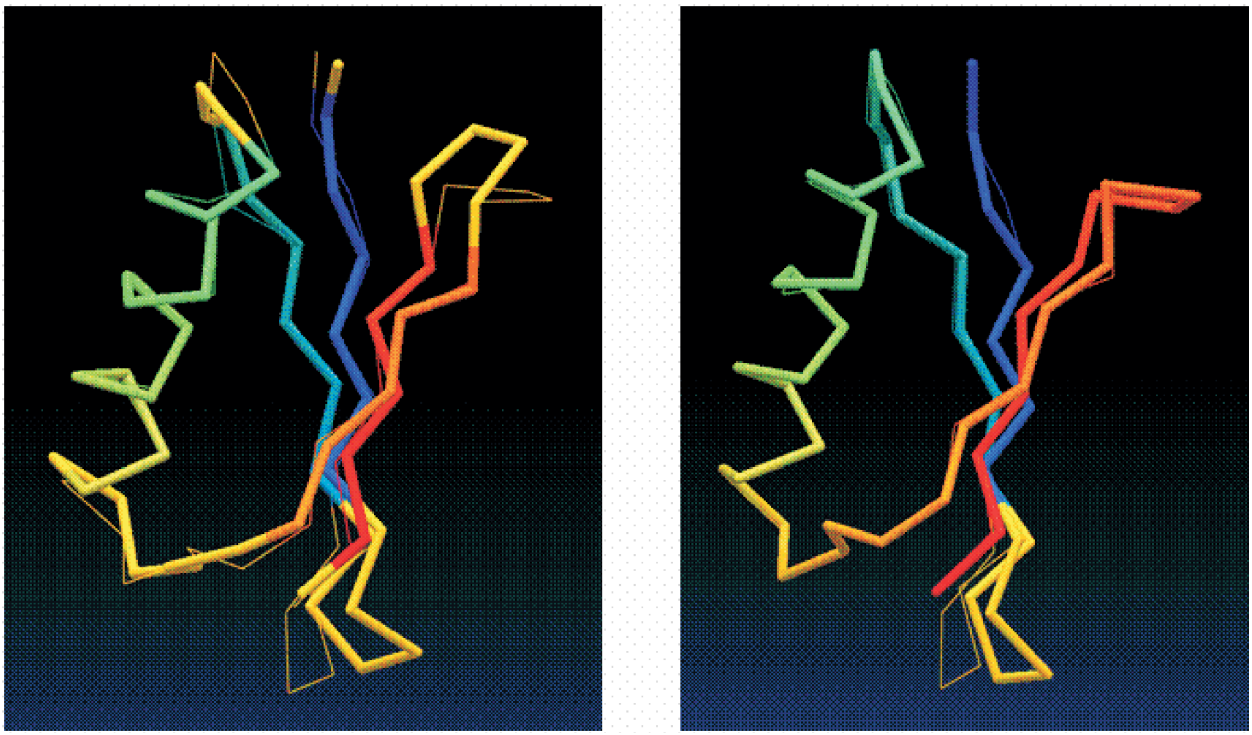

Figure 9. Models (thick lines) of 2gb1 in the alpha carbon representation superimposed onto crystallographic structure (thin lines).

The colored portions of the chains correspond to the protein core. The yellow fragments represent the loop regions. The left side panel illustrates simultaneous modeling of the all 21 loop residues. The right side panel illustrates modeling of the single loop (residues 11–16 of 2gb1).

various problems of comparative modeling. For instance, the model is now being tested within an algorithm for flexible docking of ligands (compare Evers *et al.*, 2003) to both: experimentally determined and theoretically predicted protein structures. Near-future rectifications of the model's force-field include improvements of the selectivity of the short- and – the long-range interactions by statistical analysis accounting for the local-sequence similarity (measured by the similarity of the sequence profiles) of fragments of a protein sequence to the proteins of known structures.

Helpful assistance of Drs. Piotr Rotkiewicz and Dominik Gront during preparation of this manuscript is gratefully acknowledged.
The color figures were prepared with the help of Dr. Rotkiewicz program Biodesigner for sequence and structure analysis and visualization of biomolecules. Biodesigner is freely available for non-commercial users and could be downloaded onto a PC platform *via* either of the following homepages:
www.biocomp.chem.uw.edu.pl or
http://www.pirx.com/biodesigne/
or onto Mac platform (this clone is named iMol) from
http://www.pirx.com/iMol/

REFERENCES

Alm E, Morozov AV, Kortemme T, Baker D. (2002) Simple physical models connect theory and experiment in protein folding kinetics. *J Mol Biol*.; **322**: 463–76.MEDLINE

Altschul SF, Madden TL, Schaefer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*.; **25**: 3389–402.MEDLINE

Anfinsen CB. (1973) Principles that govern the folding of protein chains. *Science*.; **181**: 223–30.MEDLINE

Anfinsen CB, Scheraga HA. (1975) Experimental and theoretical aspects of protein folding. *Adv Protein Chem*.; **29**: 205–300.MEDLINE

Arnold J, Hilton N. (2003) Genome sequencing: Revelations from a bread mould. *Nature*.; **422**: 821–2.MEDLINE

Baker D, Sali A. (2001) Protein structure prediction and structural genomics. *Science*.; **294**: 93–6.MEDLINE

Bernstein FC, Koetzle TF, Williams GJB, Meyer Jr EF, Brice MD, Rodgers JR, Kennard O, Simanouchi T, Tasumi M. (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol*.; **112**: 535–42.MEDLINE

Betancourt MR. (2003) A reduced protein model with accurate native-structure identification ability. *Proteins*.; **53**: 889–907.MEDLINE

Betancourt MR, Skolnick J. (2000) Finding the needle in a haystack. Educing protein native folds from ambigous ab initio folding predictions. *J Comput Chem*.; **22**: 339–53.

Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. (2003) Protein fragment reconstruction using various modeling techniques. *J Comput Aided Mol Des*.; **17**: 725–38.MEDLINE

Brooks CL 3rd, Gruebele M, Onuchic JN, Wolynes PG. (1998) Chemical physics of protein folding. *Proc Natl Acad Sci USA*.; **95**: 11037–8.MEDLINE

Buchete NV, Straub JE, Thirumalai D. (2004) Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci*.; **13**: 862–74.MEDLINE

Bujnicki JM, Rotkiewicz P, Kolinski A, Rychlewski L. (2001) Three-dimensional modeling of the I-TevI homing endonuclease catalytic domain, a GIY-YIG superfamily member, using NMR restraints and Monte Carlo dynamics. *Protein Eng*.; **14**: 717–21.MEDLINE

Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK. (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci*.; **11**: 723–38.MEDLINE

Cherkasov AR, Jones SJ. (2004) Structural characterization of genomes by large scale sequence-structure threading. *BMC Bioinformatics*.; **5**: 37.MEDLINE

Clark MS. (1999) Comparative genomics: the key to understand the Human Genome Project. *Bioessays*.; **21**: 121–30.MEDLINE

Covell DG. (1992) Folding protein alpha-carbon chains into compact forms by Monte Carlo methods. *Proteins*.; **14**:

409–20.MEDLINE

Covell DG, Jernigan RL. (1990) Conformations of folded proteins in restricted spaces. *Biochemistry*.; **29**: 3287–94.MEDLINE

Crippen GM. (1991) Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*.; **30**: 4232–7.MEDLINE

Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. (2002) Direct molecular dynamics observation of protein folding transition state ensemble. *Biophys J*.; **83**: 3525–32.MEDLINE

Evers A, Gohlke H, Klebe G. (2003) Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. *J Mol Biol*.; **334**: 327–45.MEDLINE

Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CLI. (2000) Accurate reconstruction of all-atom protein representation from side chain based low resolution models. *Proteins*.; **41**: 86–97.MEDLINE

Fiser A, Sali A. (2003) ModLoop: automated modeling of loops in protein structures. *Bioinformatics*.; **19**: 2500–1.MEDLINE

Fiser A, Do RK, Sali A. (2000) Modeling of loops in protein structures. *Protein Sci*.; **9**: 1753–73.MEDLINE

Ginalski K, Rychlewski L. (2003) Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins*.; **53** Suppl 6: 410–7.MEDLINE

Go N, Abe H, Mizuno H, Taketomi H. (1980) *Protein Folding*; pp 167–81. Elsevier/North Holland, Amsterdam.

Godzik A. (1996) Knowledge-based potentials for protein folding: what can we learn from known protein structures? *Structure*.; **4**: 363–6.MEDLINE

Godzik A, Kolinski A, Skolnick J. (1992) Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol*.; **227**: 227–38.MEDLINE

Godzik A, Kolinski A, Skolnick J. (1993a) Lattice representation of globular proteins: How good are they? *J Comp Chem*.; **14**: 1194–202.

Godzik A, Skolnick J, Kolinski A. (1993b) Regularities in interaction patterns of globular proteins. *Protein Eng*.; **6**: 801–10.MEDLINE

Godzik A, Kolinski A, Skolnick J. (1995) Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci*.; **4**: 2107–17.MEDLINE

Gront D, Kolinski A, Skolnick J. (2000) Comparison of three Monte Carlo search strategies for a proteinlike homopolymer model: folding thermodynamics and identification of low-energy structures. *J Chem Phys*.; **113**: 5065–71.

Hagler AT, Honig B. (1978) On the formation of protein tertiary structure on a computer. *Proc Natl Acad Sci USA*.; **75**: 554–8.MEDLINE

Hansmann UH, Okamoto Y. (1999) New Monte Carlo algorithms for protein folding. *Curr Opin Struct Biol*.; **9**: 177–83.MEDLINE

Hardin C, Eastwood MP, Prentiss M, Luthey-Schulten Z, Wolynes PG. (2002) Folding funnels: the key to robust protein structure prediction. *J Comput Chem*.; **23**: 138–46.MEDLINE

Harrison PM, Gerstein M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J Mol Biol*.; **318**: 1155–74.MEDLINE

Hinds DA, Levitt M. (1992) A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci USA*.; **89**: 2536–40.MEDLINE

Hinds DA, Levitt M. (1994) Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol*.; **243**: 668–82.MEDLINE

Holm L, Sander C. (1996) Mapping the protein universe. *Science*.; **273**: 595–602.MEDLINE

Jernigan RL. (1992) Protein folds. *Curr Opin Struct Biol*.; **2**: 248–56.

Jones DT. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*.; **287**: 797–815.MEDLINE

Kabsch W, Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*.; **22**: 2577–637.MEDLINE

Kihara D, Skolnick J. (2004) Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins*.; **55**: 464–73.MEDLINE

Kihara D, Lu H, Kolinski A, Skolnick J. (2001) TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci USA*.; **98**: 10125–30.MEDLINE

Klose J. (1989) Systematic analysis of the total proteins of a mammalian organism: principles, problems and implications for sequencing the human genome. *Electrophoresis*.; **10**: 140–52.MEDLINE

Kolinski A, Skolnick J. (1994a) Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins*.; **18**: 353–66.MEDLINE

Kolinski A, Skolnick J. (1994b) Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins*.; **18**: 338–52.MEDLINE

Kolinski A, Skolnick J. (1966) *Lattice Models of Protein Folding, Dynamics and Thermodynamics*; p 200 R.G. Landes, Austin, TX.

Kolinski A, Skolnick J. (1997a) Determinants of seconadry structure of polypeptide chains: interplay between short range and burial interactions. *J Chem Phys*.; **107**: 953–64.

Kolinski A, Skolnick J. (1997b) High coordination lattice models of protein structure, dynamics and thermodynamics. *Acta Biochim Polon*.; **44**: 389–422.MEDLINE

Kolinski A, Skolnick J. (1998) Assembly of protein structure from sparse experimental data: an efficient Monte Carlo Model. *Proteins*.; **32**: 475–94.MEDLINE

Kolinski A, Skolnick J. (2004) Reduced models of proteins and their applications. *Polymer*.; **45**: 511–24.

Kolinski A, Godzik A, Skolnick J. (1993) A General method for the prediction of the three dimensional structure and folding pathway of globular proteins. Application to designed helical proteins. *J Chem Phys*.; **98**: 7420–33.

Kolinski A, Galazka W, Skolnick J. (1995) Computer design of idealized beta-motifs. *J Chem Phys*.; **103**: 10286–97.

Kolinski A, Galazka W, Skolnick J. (1998a) Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: application to small helical, b and a/b proteins. *J Chem Phys*.; **108**: 2608–17.

Kolinski A, Jaroszewski L, Rotkiewicz P, Skolnick J. (1998b) An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side groups centers of mass. *J Phys Chem*.; **102**: 4628–37.

Kolinski A, Ilkowski B, Skolnick J. (1999) Folding dynamics and thermodynamics of b-hairpin assembly: Insight from various simulation techniques. *Biophys, J*.; **77**: 2942–52.MEDLINE

Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. (2000) Protein folding: flexible lattice models. *Progress of Theoretical Physics (Kyoto)*.; **138** Suppl: 292–300.

Kolinski A, Betancourt M, Kihara D, Rotkiewicz P, Skolnick J. (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, lattice and off-lattice modeling for protein structure prediction and refinement. *Proteins*.; **44**: 133–49.MEDLINE

Kolinski A, Klein P, Romiszowski P, Skolnick J. (2003a) Unfolding of globular proteins: Monte Carlo dynamics of a realistic reduced model. *Biophys J*.; **85**: 3271–8.MEDLINE

Kolinski A, Gront D, Pokarowski P, Skolnick J. (2003b) A simple lattice model that exhibits a protein-like cooperative all-or-none folding transition. *Biopolymers*.; **69**: 399–405.MEDLINE

Kuznetsov IB, Rackovsky S. (2002) Discriminative ability with respect to amino acid types: assessing the performance of knowledge-based potentials without threading. *Proteins*.; **49**: 266–84.MEDLINE

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp

D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. (2001) Initial sequencing and analysis of the human genome. *Nature*.; **409**: 860–921.MEDLINE

Lee J, Liwo A, Scheraga HA. (1999) Energy-based *de novo* protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10–55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc Natl Acad Sci USA*.; **96**: 2025–30.MEDLINE

Lee J, Ripoll DR, Czaplewski C, Pilardy J, Wedemeyer WJ, Scheraga HA. (2001) Optimization of parameters in macromolecular potential energy functions by conformational space annealing. *J Phys Chem*.; **105**: 7291–8.

Levitt M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*.; **104**: 59–107.MEDLINE

Levitt M, Warshel A. (1975) Computer simulation of protein folding. *Nature*.; **253**: 694–8.MEDLINE

Li W, Zhang Y, Kihara D, Huang YJ, Zheng D, Montelione GT, Kolinski A, Skolnick J. (2003) TOUCHSTONEX: protein structure prediction with sparse NMR data. *Proteins*.; **53**: 290–306.MEDLINE

Liwo A, Arlukowicz P, Czaplewski C, Oldziej S, Pilardy J, Scheraga HA. (2002) A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field. *Proc Natl Acad Sci USA*.; **99**: 1937–42.MEDLINE

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. (1953) Equation of state calculations by fast computing machines. *J Chem Phys*.; **51**: 1087–92.

Mirny LA, Finkelstein AV, Shakhnovich EI. (2000) Statistical significance of protein structure prediction by threading. *Proc Natl Acad Sci USA*.; **97**: 9978–83.MEDLINE

Miyazawa S, Jernigan RL. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*.; **18**: 534–52.

Miyazawa S, Jernigan RL. (1999a) An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins*.; **36**: 357–69.MEDLINE

Miyazawa S, Jernigan RL. (1999b) Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. *Proteins*.; **36**: 347–56.MEDLINE

Mohanty D, Dominy BN, Kolinski A, Brooks CL 3rd, Skolnick J. (1999) Correlation between knowledge-based and detailed atomic potentials: application to the unfolding of the GCN4 leucine zipper. *Proteins*.; **35**: 447–52.MEDLINE

Monge A, Lathrop EJP, Gunn JR, Shenkin PS, Friesner RA. (1995) Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *J Mol Biol*.; **247**: 995–1012.MEDLINE

Montelione GT, Zheng D, Huang YJ, Gunsalus KC, Szyperski T. (2000) Protein NMR spectroscopy in structural genomics. *Nat Struct Biol*.; **7** Suppl: 982–5.MEDLINE

Pilardy J, Czaplewski C, Liwo A, Wedemeyer WJ, Lee J, Ripoll DR, Arlukowicz P, Oldziej S, Arnautova YA, Scheraga HA. (2001) Development of physics-based energy functions that predict medium-resolution structures for proteins of the

a,b, and a/b structural classes. *J Phys Chem*.; **105**: 7299–311.

Reva BA, Finkelstein AV, Sanner MF, Olson AJ. (1996) Adjusting potential energy functions for lattice models of chain molecules. *Proteins*.; **25:** 379–88.MEDLINE

Rey A, Skolnick J. (1993) Computer modeling and folding of four-helix bundles. *Proteins*.; **16**: 8–28.MEDLINE

Rey A, Skolnick J. (1994) Computer simulation of the folding of coiled coils. *J Chem Phys*.; **100**: 2267–76.

Rooman M, Gilis D. (1998) Different derivations of knowledge-based potentials and analysis of their robustness and context-dependent predictive power. *Eur J Biochem*.; **254**: 135–43.MEDLINE

Rost B, Schneider R, Sander C. (1997) Protein fold recognition by prediction-based threading. *J Mol Biol*.; **270**: 471–80.MEDLINE

Rotkiewicz P, Sicinska W, Kolinski A, DeLuca HF. (2001) Model of three-dimensional structure of vitamin D receptor and its binding mechanism with 1a, 25-dihydroxivitamin D. *Proteins*.; **44**: 188–99.MEDLINE

Sali A. (1995) Comparative protein modeling by satisfaction of spatial restraints. *Mol Med Today*.; **1**: 270–7.MEDLINE

Sali A. (1998) 100,000 protein structures for the biologist. *Nat Struct Biol*.; **5**: 1029–32.MEDLINE

Sali A. (2001) Target practice. *Nat Struct Biol*.; **8**: 482–4.MEDLINE

Schonbrun J, Wedemeyer WJ, Baker D. (2002) Protein structure prediction in 2002. *Curr Opin Struct Biol*.; **12**: 348–54.MEDLINE

Shakhnovich EI. (1997) Theoretical studies of protein-folding thermodynamics and kinetics. *Curr Opin Struct Biol*.; **7**: 29–40.MEDLINE

Shimada J, Ishchenko AV, Shakhnovich EI. (2000) Analysis of knowledge-based protein-ligand potentials using a self-consistent method. *Protein Sci*.; **9**: 765–75.MEDLINE

Simons KT, Strauss C, Baker D. (2001) Prospects for ab initio protein structural genomics. *J Mol Biol*.; **306**: 1191–9.MEDLINE

Skolnick J, Fetrow JS, Kolinski A. (2000) Structural genomics and its importance for gene function analysis. *Nat Biotechnol*.; **18**: 283–7.MEDLINE

Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D. (2003) TOUCHSTONE: a unified approach to protein structure prediction. *Proteins*.; **53** Suppl 6: 469–79.MEDLINE

Sun S. (1993) Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci*.; **2**: 762–85.MEDLINE

Sun Z, Xia X, Guo Q, Xu D. (1999) Protein structure prediction in a 210-type lattice model: parameter optimization in the genetic algorithm using orthogonal array. *J Protein Chem*.; **18**: 39–46.MEDLINE

Swendsen RH, Wang JS. (1986) Relica Monte Carlo simulations. *Phys Rev Lett*.; **57**: 2607–9.MEDLINE

Vajda S, Vakser IA, Sternberg MJ, Janin J. (2002) Modeling of protein interactions in genomes. *Proteins*.; **47**:

444–6.MEDLINE

Wolynes PG, Onuchic JN, Thirumalai D. (1995) Navigating the folding routes. *Science*.; **267**: 1619–20.MEDLINE

Zacharias M. (2003) Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*.; **12**: 1271–82.MEDLINE