Protein Structure Prediction Using Coarse-Grained Models

Maciej Blaszczyk, Dominik Gront, Sebastian Kmiecik, Mateusz Kurcinski, Michal Kolinski, Maciej Pawel Ciemny, Katarzyna Ziolkowska, Marta Panek and Andrzej Kolinski

Abstract The knowledge of the three-dimensional structure of proteins is crucial for understanding many important biological processes. Most of the biologically relevant protein systems are too large for classical, atomistic molecular modeling tools. In such cases, coarse-grained (CG) models offer various opportunities for efficient conformational sampling and thus prediction of the three-dimensional structure. A variety of CG models have been proposed, each based on a similar framework consisting of a set of conceptual components such as protein representation, force field, sampling, etc. In this chapter we discuss these components, highlighting ideas which have proven to be the most successful. As CG methods are usually part of multistage procedures, we also describe approaches used for the incorporation of homology data and all-atom reconstruction methods.

1 Introduction

1.1 Why Do We Need CG Models?

Proteins are key components of all life processes. Thus, the development of relatively cheap and automatic methods for determining amino acid sequences of proteins raised hope for a breakthrough in many branches of science, including pharmacy and biotechnology. However, the knowledge of sequence is insufficient for the majority of

M. Blaszczyk · D. Gront · S. Kmiecik · M. Kurcinski · M. P. Ciemny · K. Ziolkowska · M. Panek
A. Kolinski (✉)
Faculty of Chemistry, Biological and Chemical Research Centre,
University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland
e-mail: kolinski@chem.uw.edu.pl

M. Kolinski
Bioinformatics Laboratory, Mossakowski Medical Research Centre,
Polish Academy of Sciences, Warsaw, Poland

M. P. Ciemny
Faculty of Physics, University of Warsaw, Pasteura 5, 02-093 Warsaw, Poland

A. Liwo (ed.), *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*, Springer Series on Bio- and Neurosystems 8, https://doi.org/10.1007/978-3-319-95843-9_2

applications, and it is necessary to determine three-dimensional structures of proteins and their complexes. Unfortunately, determining the three-dimensional structures has turned out to be a much more demanding problem than studying their sequences. Currently, the Uniprot database [111, 137] contains over 60 million protein sequences, while the total number of known protein structures, or their complexes, in RCSB Protein Data Bank [114] is over 130 thousand. Just a few years ago these numbers were significantly smaller. In May 2012 the RefSeq database [110] contained less than 16 million sequences, while the total number of known structures in Protein Data Bank [7] was about 80 thousand. The main reason for the increasing disproportion is that experimental structure determination may be both expensive and time-consuming. Additionally, in many cases it simply cannot succeed.

Theoretical methods are a potential alternative to the experimental techniques. Over half a century ago, Anfinsen and coworkers [6] showed that the three-dimensional structure of bovine pancreatic ribonuclease is exclusively determined by its sequence. Later, this statement has been generalized for the great majority of globular proteins. Thus, the problem of finding the native state of a protein can be regarded as a problem of free energy minimization [5]. However, due to the number of atoms in protein compounds, their conformational space is often defined by as many as 10^4–10^6 degrees of freedom (DOFs). Therefore, to be able to computationally study biomolecule behaviors, such as large internal motions or conformational changes, we have to reduce the number of DOFs by at least one order of magnitude. This goal can be achieved by building models in which some of the atoms are ignored or grouped to form united pseudo-atoms. Models that follow this paradigm are often called reduced or coarse-grained (CG) models [58].

An ideal coarse graining approach should be able to simplify an atomistic system without losing its important features, such as structural details, characteristic interactions and internal dynamics. Although many CG methods have been created, none of them fully satisfies these requirements. Currently, CG models can be regarded as a part of a multi-scale procedures rather than stand-alone protein prediction methods (Fig. 1). Nevertheless, the idea of CG models was a milestone in protein structure prediction by computational methods [58].

1.2 History of CG Models

Coarse graining is used in biomolecular modeling since the very beginning of this discipline. In their seminal work, Levitt and Warshel [87] started protein simulations with a model where each residue in a peptide chain was represented by its alpha carbon (Cα) and a united atom substituting its side chain. Since then, a huge variety of models have been proposed that cover the whole range of complexity: from the most simplistic Cα-only approaches to all-atom representation [3, 19, 22, 45, 68, 60, 65, 74, 113, 125, 126, 127]. Between these two extreme representations we can find models with or without residue side chains. Each side chain may in turn be represented by one or more interacting centers. A few different methods for the protein backbone

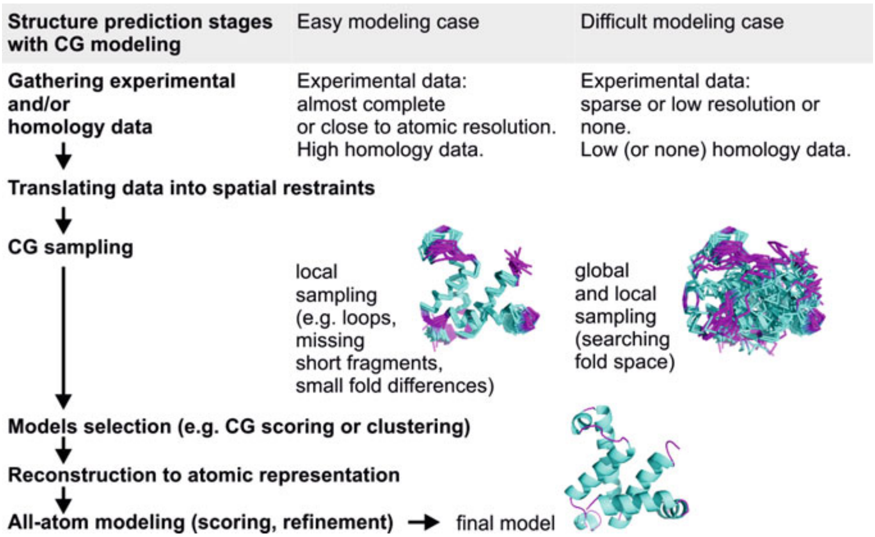
Structure prediction stages with CG modeling
Easy modeling case
Difficult modeling case
Gathering experimental and/or homology data
Experimental data: almost complete or close to atomic resolution. High homology data.
Experimental data: sparse or low resolution or none. Low (or none) homology data.
Translating data into spatial restraints
CG sampling
local sampling (e.g. loops, missing short fragments, small fold differences)
global and local sampling (searching fold space)
Models selection (e.g. CG scoring or clustering)
Reconstruction to atomic representation
All-atom modeling (scoring, refinement)
final model

Fig. 1 Protein structure prediction stages in CG modeling. The diagram presents a general pipeline for multiscale modeling (CG merged with all-atom) and depicts major differences between easy and difficult modeling tasks. Easy or medium-difficulty cases, if necessary, require only limited CG sampling of the conformational space, usually to fill small gaps and quite small uncertainties in available experimental or homology inference data. Extensive CG sampling is required for difficult cases when knowledge about the expected structure is limited

have also been proposed. Finally, discretization may be used to impose additional limits on the space of possible conformations the model can adopt. A quick glimpse of the review articles [20, 41, 58, 67, 140] suggests that most likely all the choices of CG representations have already been explored. This review certainly does not describe all these solutions. Instead, we introduce several important concepts of CG modeling of proteins and other biomolecules and describe the way they have evolved in the past few decades.

One of such inspiring ideas was lattice models [71]. Restricting atomic coordinates to a grid became a very straightforward and simple way to discretize the conformational space. The search space size was greatly reduced and many of its local minima vanished. Atomic coordinates became integer values, which opened many possibilities for use of hash tables. Most importantly, the Cartesian space itself could be stored in a three-dimensional array which resulted in the O(1) time complexity (constant time) of collision detection. Due to these advantages lattice models were at least one or two orders of magnitude faster than their continuous space counterparts. Low resolution grids, however, have a few serious drawbacks. First of all, simple lattice (cubic lattice, face-centered lattice, etc.) representations of protein structures (usually limited to Cα traces, or at least to a few atom centers) were of relatively low resolution, with average errors of such representation of 2–6 Å. Even more risky aspects of low resolution lattice models are related to lattice anisotropy.

Depending on the orientation in respect to the fixed lattice, the resulting model chains changed local geometry and local resolution. Moreover, essentially for all types of interactions the local energy may change with chain orientation. Consequently, the models were highly degenerated, with changing energetic preferences for various orientations of protein fragments.

The first detailed analysis of these problems was published by Godzik and coworkers [34]. Higher resolution reduced protein models, for example Chess-Knight lattice models, led to more accurate representations and smaller anisotropy effects. Probably the best functionality of lattice models was achieved by higher resolutions (more lattice steps per residue) and allowed fluctuation of Cα-Cα distances in the models. This led to a large number of single-bead orientations, higher resolution (1–2 Å in respect to experimental structures) and essentially negligible anisotropy. Obviously, the higher resolution of such models led to a much larger number of allowed structures which caused somewhat higher computational cost. With increasing computing power, the higher resolution lattice models could still deliver fast simulations of protein folding, multibody interactions and related problems. The most important advantage of high-resolution CG protein models (with slightly fluctuating fixed distances in protein chains, usually the Cα-Cα distance) is their computational efficiency. In comparison to continuous models the high-resolution lattice models could be simulated much faster.

Another very prolific concept in protein structure modeling is the use of fragments, that is, short peptides extracted from known protein structures. This concept was originally introduced by Jones and Thirup [50] as a crystallographic method for rapid model building based on experimental electron density. The authors also discussed potential applications of short protein fragments in purely theoretical modeling approaches, which in practice was applied in the late 1990s by J.R. Gunn as well as by Baker and coworkers [13]. The latter application soon became the famous Rosetta program [113], one of the most successful methods in ab initio structure prediction. Later, fragment-based sampling was applied to numerous protein modeling approaches [145, 147]. It should be noted, however, that fragment-based sampling introduces a very strong bias in the dynamics of the sampled chain which makes it unsuitable for studying numerous research problems.

Multibody Force Fields

Interactions within a CG model cannot be directly learned from a physical system. Therefore, they have to be established in the form of a mean field potential. Such a potential can be derived either from statistics extracted from known protein structures [124] or from Molecular Dynamics simulations [90]. Knowledge-based force field models have been actively developed for the past few decades, which resulted in the remarkable improvement of their performance. Among the most important elements we note the proper choice of the reference state [151] and multibody terms. It has recently been shown that two-body potentials are not capable of recognizing all native folds against large datasets of decoy structures [138]. They also cannot properly mimic the cooperativity of the protein folding process [21]. Multibody potentials remediate these problems to some extent and perform significantly better than two-body terms.

2 Necessary Components of CG Models

2.1 Protein Representation and Coordinates

When undertaking a biomolecular modeling study of a particular system, the level of coarse graining must be defined. This includes defining atoms that are explicitly present in the model, atoms that are grouped into united pseudo-atoms and finally atoms that are ignored. By reducing the number of interacting centers we can reduce the cost of energy evaluation. The number of modeled atoms is also related to the number of degrees of freedom (DOFs) of the modeled system, although the dependence is not straightforward.

In most cases, the Cα atom is explicitly defined and serves as the most important point within a residue (perhaps the SICHO model [63, 72, 139] is the only medium-resolution exception to this rule). As for the other backbone atoms, there are three commonly used approaches: Cα only (with N, C and O neglected), all atoms present [48, 113, 145] and a virtual point method. The major issue resulting from removing peptide plate atoms is the problem with the accurate definition of a hydrogen bond between two residues. Cα-only backbone representations attempt to define the hydrogen bonding potential on Cα coordinates; however, such definitions are rather inaccurate [24]. At the same time, hydrogen bonds are crucial for maintaining the proper local geometry of a backbone and for forming secondary structure elements. Thus, a relatively accurate description of this interaction is required. In a virtual point approach, originally proposed by Levitt [86], a point is defined in the geometric center between two subsequent alpha carbons. This approach has been implemented in numerous applications, both in intermediate resolution lattice models [62, 64, 66, 100] and in off-lattice Cartesian space models [12, 94]. Based on its coordinates, a hydrogen bond can be defined with reasonable accuracy. Virtual point also describes the excluded volume effect of the neglected backbone atoms; φ/ψ angles, however, cannot be defined.

Various methods differ in the coarse graining of side chains. It may comprise just one united Side Group (SG) atom, Cβ + united atom (Cβ + SG) or a few united atoms. The simplest case, where the whole side chain is modeled just as a sphere, is also the most inaccurate one. At the next accuracy level, the whole side chain is represented by an ellipsoid [90] or by a Cβ and a united sphere of all the other side chain atoms [68]. These two representations enable satisfactory accuracy with reasonable computational cost. The advantage of the Cβ + SG approach is that of the 20 biogenic amino acids four (G, A, S and C) residues are already accurately represented and for a few others the approximations are rather small. The most challenging, however, are the long side chains incorporating both polar and aliphatic moieties, such as LYS or

TRP. For a better description of these entities, finer coarse graining must be defined with more than two united atoms substituting a side chain [8, 11].

The choice of atoms used to represent a protein chain is strongly connected to the set of the degrees of freedom to be sampled. In Cα-only, Cα + SG and similar approaches, conformational search is done in the Cartesian space. In many cases, however, the conformational space is smaller than 3 N DOFs, because conformations of some these atoms depend on the others. In CABS [68], for example, each residue comprises up to 4 atoms, but only Cα atom is independent. All the remaining atomic positions are unambiguously defined by the Cα trace. To the contrary, in the SICHO model [63, 72, 139] with two interaction centers: Cα and SG, only SG is independent and Cα coordinates based on them are back-calculated. In another example the Rosetta model definition [113] is based on all backbone atoms and SG, but the conformation of a peptide chain is defined by three degrees of freedom only, dihedral angles of each residue: φ, ψ and ω.

Another method used to increase the computational efficiency of a computational model is the discretization of the conformational space. It has been realized since the early days of protein simulations [106] that even a small set of distinct states allowed for a residue can result in the reasonable accuracy of a projected structure. Such a set of selected states can be easily defined when a conformation is described by its internal coordinates. For models defined in the Cartesian space a lattice (grid) is used to limit the search space. In practice a set of basis vectors is defined to connect any two Cα atoms that follow each other in a protein chain. This implies that any conformation of a chain of N residues can be uniquely written as N-1 integer indexes that refer to particular vectors in the basic set. Other atoms of the CG representation (such as SG) may or may not be restricted to the grid.

2.2 Force Field

The already mentioned methods: SICHO, CABS and Rosetta [68, 60, 113] use only three degrees of freedom per residue; however, they employ more than one center to define the interactions of a particular residue. All these atoms, united atoms and virtual points are used to calculate geometric properties, such as distances and planar and dihedral angles. These properties underlie the definition of the energy function of a system. The definition usually assumes a very complex mathematical form of the function which we discuss in detail below. The mathematical formula must be completed with a (possibly large) set of parameters, such as various constants, scaling factors, etc. In the case of all-atom models used for biomolecular studies, the parameters can be derived from experimental data, such as small molecule measurements. This is, however, not possible in the case of a CG model, simply because none of the models reviewed here exists in the real world and many of their properties cannot be measured. Therefore, the energy function for a CG model comprises at least partially statistical potentials of mean force. The construction of such force fields has become a research discipline on its own. Here we provide a very basic description of mean field force fields and focus on differences between particular CG approaches.

The evaluation of energy of a particular conformation requires computation of relevant geometrical properties (for example distances or angles). This enforces recalculation of the Cartesian representation of a biomolecule if a CG model is defined in internal coordinates [107]. Lattice models, on the other hand, use hashing and store some local geometrical properties, such as planar or dihedral angles, vector products etc., in look-up tables. Moreover, the energy function may be conveniently stored in an array and indexed by a distance bin or vector indices.

The hydrogen bonding is one of the indispensable terms of the force field. There were numerous actual attempts proposed in the literature based on different atom types which capture local geometrical properties of the main chain in different ways [24, 35, 73, 102]. For better recapitulation of the local geometry of secondary structure elements, correlations between neighboring hydrogen bonds may be modeled explicitly by an additional potential [68, 88, 92].

Another very important energy component is the one corresponding to hard core repulsion between atoms, often described as an excluded volume term. A rapidly growing function may be used to model this interaction, such as the so-called "12" Lennard-Jones potential term. Relevant radii for united atoms are computed as an average over all the relevant conformations of a group that has been coarse grained into a sphere. Hard core repulsion in low- to medium-resolution on-lattice models may be evaluated instantly just by a single look-up in the 3D matrix that stores the lattice space.

The attractive pairwise potential is established by the Boltzmann inversion of relevant statistics extracted from known protein structures. The potential may depend solely on the distance between interacting partners; in other approaches it takes into account the mutual orientation of the groups and their neighborhood [12, 68].

Local backbone conformation and secondary structure formation is controlled by mean force potentials encoding local correlations between degrees of freedom. Typically, the potentials also depend on amino acid sequences and encode propensities of particular amino acid types to form a given secondary structure. The actual formulation of these potentials depends on how the main chain is represented in the model. In the cases where all backbone atoms are available, Ramachandran-type energy maps are utilized. Otherwise, local interactions depend on local distances, for example between the ith and $i+2$nd Cα, usually denoted as R_{13}, as well as on R_{14} and R_{15}. Another choice is to define energy terms based on planar and dihedral angles between successive Cα atoms.

CG force field is often completed by terms that mimic solvent-induced effects and long range electrostatics. Examples of such terms include centrosymmetric (compacting) potential and various environmental terms.

2.3 Conformational Sampling

Even a CG reduced system is still characterized by a very large number of degrees of freedom (approximately 10^2–10^3). Due to the high density of a molecular system and covalent bonds between atoms (or virtual bonds that connect pseudo- and united atoms), the potential energy hypersurface is extremely rugged (see [142] and reference therein). The motion along many of the DOFs may be impossible due to high energy barriers. A number of methods are used to explore this space, however, the most common are Monte Carlo (MC) and Molecular Dynamics (MD) approaches. Unlike other, specialized methods, these two are general and flexible. Both produce low-energy ensembles, which elucidate protein dynamics.

The general idea of the Monte Carlo [99] may be applied to biomolecular systems in numerous ways. Probably the simplest one is simulated annealing [57] which uses the Metropolis criterion [98] to construct the Boltzmann ensemble of states at an arbitrary temperature. Simulation starts from a high-temperature conformation where the system undergoes large configurational changes but its energy is relatively high. Using gradual cooling leads the system to adopting a conformation in a local energy minimum. Repeating this process leads to the exploration of the energy landscape. However, the chance of finding the global minimum of a biomolecular system is relatively low. This problem can be alleviated to a great extent using the Replica Exchange Monte Carlo (REMC), also known as Parallel Tempering [31, 38, 42, 132]. In this approach many (usually a few to tens of) simulations are run simultaneously. Each simulation runs a separate non-interacting copy (called a replica) of the same system using isothermal Metropolis MC. Occasionally two systems (X_i and X_j) exchange their temperatures. The system X_i which has been so far simulated at T_i goes to T_j and X_j goes from T_j to T_i. The probability p of this exchange is determined by temperatures T_j and T_i as well as by the energies of the two systems:

$$E_i \text{ and } E_j\colon p = \min(1, \exp(\Delta)) \tag{1}$$

where Δ is expressed by:

$$\Delta = \left(1/T_i - 1/T_j\right)\left(E_i - E_j\right) \tag{2}$$

Such an algorithm constructs a Markov chain over a number of Markov chain processes. The exchange between the structures in different replicas facilitates relaxation of structures that might otherwise be trapped in local energy minima. The density of states of the sampled system can be recalculated by a histogram reweighting technique [27, 28, 39, 80]. The Parallel Tempering algorithm can also be applied to Molecular Dynamics simulations [131].

There are also many variants of Molecular Dynamics [51, 82, 101]. In its standard formulation, the trajectory of a molecular system is calculated by solving the Newton's equations of motion at each time step. The forces on the system are computed as the gradient of the potential energy function (the force field) which is dependent on the positions of atoms. To reduce the computational complexity, some limitations may be imposed on the range of interactions between atoms or united atoms.

As in Monte Carlo sampling, Cartesian coordinates may be substituted by generalized variables. Practical examples include all-atom [1, 97] and CG simulations [93]. This approach indeed allows for a significant increase of the integration time step. However, the applicability of this approach is limited, since the forces evaluation requires a recalculation of the Cartesian coordinates of the system (which involves a time-consuming matrix inversion) at every time step.

3 Representative CG Methods

Above, we described all the major components of a coarse-grained model. Now let us summarize a few well-established computational models with particular emphasis on these elements. The models differ in the level of coarse graining and the number of degrees of freedom utilized to define a polypeptide chain. For convenience, the key features of the models are presented in Table 1.

Table 1 Comparison of selected CG methods

Method	Protein representation	Conformational space	Coordinates system	Sampling scheme
Levitt and Warsell model	Cα, SG	Continuous space	Angular	Molecular dynamics
CABS	Cα, Cβ, SG and virtual point at the peptide bond center	Cubic lattice with 0.61 Å spacing which restricts Cα positions	Cartesian	Replica exchange Monte Carlo
SICHO	SG, Cα	Cubic lattice with 1.45 Å spacing which restricts SG positions	Cartesian	Monte Carlo
Rosetta	All backbone atoms and SG or all-atom representation	Continuous space	Angular	Monte Carlo
UNRES	Cα, SG and peptide group	Continuous space	Cartesian/angular	Mesoscopic molecular dynamics and Monte Carlo

3.1 The Original Cα + SG Model

The model originally proposed by Levitt and Warshel in 1975 [87] uses two interaction centers per residue: the Cα atom, which is modeled explicitly, and the side chain, represented by a SG sphere. Each residue is allowed one degree of freedom only: the torsion angle between the 4 successive Cα atoms. Interactions between side chains are modeled by a van der Waals type potential. The radius of each united atom representing a side chain is calculated as the average radius of gyration of the particular group. Another important contribution to the potential energy is the side chain-solvent interaction estimated by the experimental free energy of transfer from water to ethanol. The force field is completed by local interaction expressed as a Fourier expansion function of the torsion angle defined by four Cα atoms. Classical molecular dynamics is used to sample the conformational space. Simulations of the bovine pancreatic trypsin inhibitor sometimes produced structures resembling the native fold, with the best structures having root-mean-square deviation from the native in the range of 6.5 Å. In his later works [86], Levitt introduced an additional degree of freedom for each residue, namely the planar angle between three adjoining Cα atoms. A virtual atom has also been added in the middle of a Cα-Cα vector for a more accurate definition of hydrogen bonding interactions.

3.2 CABS

The coarse-grained representation of the CABS model [68] uses up to four interaction centers per residue: Cα, Cβ, the center of mass of the side group and a virtual point placed at the center of each peptide bond (see Fig. 2a). The Cα trace of the model is restricted to an underlying cubic lattice with a spacing of 0.61 Å. In lattice units, the distance between consecutive Cα atoms varies from $29^{1/2}$ to $49^{1/2}$. This implies that the Cα-Cα distance is allowed to fluctuate between 3.29 and 4.27 Å. There are 800 possible orientations (lattice vectors) of the virtual Cα-Cα. Therefore, the model essentially avoids any lattice-related artifacts. Cβ atoms and side chains are located off-lattice, and their positions are calculated for each residue using the coordinates of three consecutive Cα atoms as a reference frame. For each amino acid, two distinct conformations are defined which mimic the averaged side chain position found in helical and expanded conformations. The rotamer type is uniquely defined by the on-lattice Cα trace; hence, a protein chain comprising N residues has 3 N independent degrees of freedom.

3.3 SICHO

The most unique feature of the Side Chain Only model [63, 72, 139] is the definition of the polypeptide chain. Each residue is represented as a spherical united atom which substitutes its side chain (see Fig. 2b). The united atoms are restricted to a cubic grid with 1.45 Å spacing. The chain vectors representing virtual bonds between interaction centers are of variable length, ranging from $9^{1/2}$ to $30^{1/2}$ lattice units. Unlike other protein models, Cα atom positions are not independent degrees of freedom. Conversely, they are uniquely defined in a local frame of three neighboring side chains and are recalculated after any conformational change. The knowledge-based force field is defined based on both Cα and side chain centers and includes a chain stiffness potential, a secondary structure bias, short-range interactions, hydrogen-bond interactions, and long-range interactions. Such deeply coarse-grained models

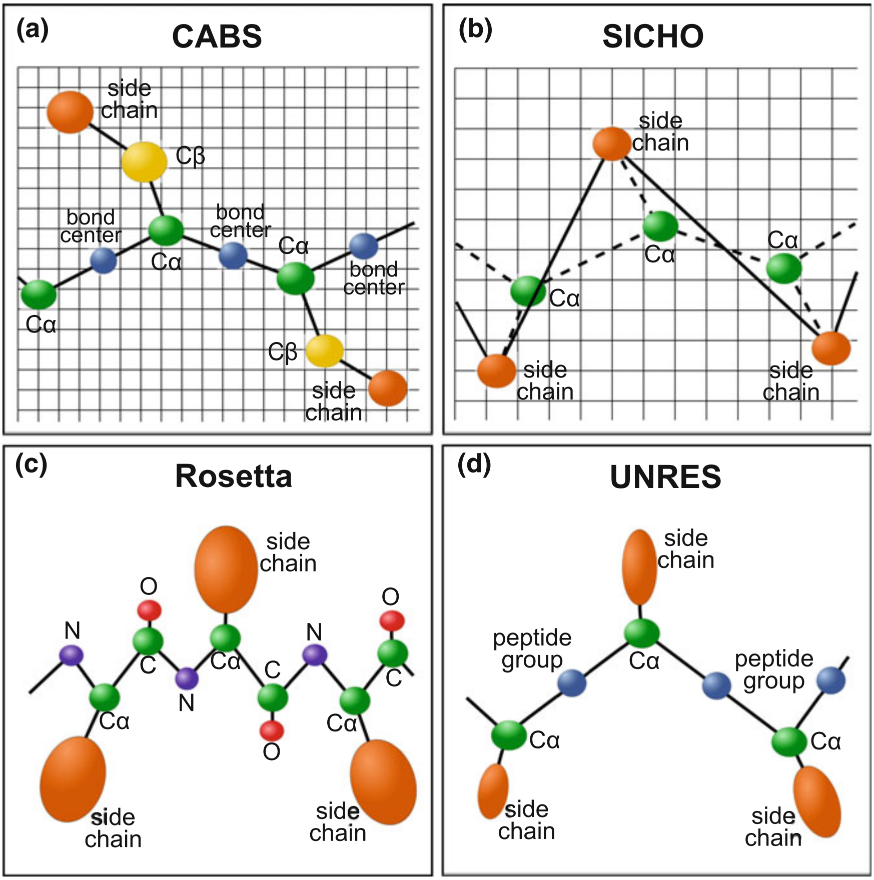

Fig. 2 Comparison of representations of four CG models

are computationally very effective [130], and they can be effective in difficult tasks of structure prediction and studies of large scale protein dynamics if the model structures resolution allows for the atom-level reconstruction. Even lower resolution realistic models of proteins can be designed if the crude structure representation can be compensated by specific patterns of knowledge based statistical potentials [23].

3.4 Rosetta

Rosetta [113, 129] utilizes a library of short peptide fragments (typically 3 and 9 residue long) as a Monte Carlo moves set. In practice a fragment is defined by three internal coordinates (φ, ψ and ω backbone dihedral angles) per residue. Each time a fragment is inserted, a number of subsequent DOFs (9 or 18, for 3mers and 9mers, respectively) are affected in the simulated polypeptide chain. The fragments themselves are extracted from known protein structures [40]. Such a sampling method reduces the conformational space, changes the respective DOFs in a correlated manner and introduces a strong bias toward protein-like geometries. Rosetta utilizes two representations: a coarse-grained, termed "centroid" (shown in Fig. 2c) and an all-atom one. In both representations the protein backbone is treated explicitly. In the centroid mode, each side chain is represented by a united atom located at the side-chain center of mass. In the high-resolution mode, atomic coordinates for all side-chain atoms, including hydrogens, are utilized. Side chains are restricted to discrete conformations as described by a backbone-dependent rotamer library. The Rosetta energy function is different for the two representations and in both cases it comprises numerous mean-field terms.

3.5 UNRES

In the UNited RESidue model [90] the protein backbone is reduced to a sequence of Cα atoms and a united peptide group (p) connected by virtual bonds (Fig. 2d). United side-chains are attached to the α-carbons (SG). In the most recent version of UNRES, the positions of these atoms are defined by internal Cartesian coordinates (vectors of the virtual bonds). Previously, planar and torsion angles were used as a set of generalized coordinates [91]. UNRES employs a physics-based mean-field force field for simulations of protein structure and dynamics. The energy function definition and conformational space sampling methods have evolved over time. Initially the effective energy function was described as a restricted free energy (RFE) function or the potential of mean force (PMF) of polypeptide chains in water. Currently, it is defined as an approximate cumulant expansion of restricted free temperature-dependent energy whose calibration is based on protein-folding thermodynamic data. UNRES is the only coarse-grained force field which explicitly depends on temperature and can compute thermodynamic quantities of protein folding.

In UNRES the conformational space search was initially based on the global optimization of the potential-energy function to find the lowest-energy conformation. It was performed by stochastic Monte Carlo-based algorithms, namely Monte Carlo plus Energy Minimization (MCM) [89] and hybrid approaches, such as Conformational Space Annealing (CSA) [85] which turned out to be the most effective. Later, UNRES was extended to mesoscopic Molecular Dynamics (MD) to study pathways and kinetics of the protein folding process. This implementation of MD reformulates the conformational sampling as a search for the most probable conformational ensembles with the lowest free energy at temperatures below the folding transition temperature. The UNRES extension of MD can also be used to simulate multichain proteins. To improve the conformational space search, UNRES can use Replica Exchange Molecular Dynamics (REMD) and Replica Exchange Monte Carlo (REMC) sampling.

The UNRES coarse-grained model has been successfully applied to the protein structure prediction problem [76, 78, 105] to study folding trajectories [118] and to investigate folding process thermodynamics [77, 152].

4 Reconstruction of an All-Atom Representation, Post-processing and Analysis

A coarse grained computational model provides description of the modeled structure at a limited resolution. To infer the biological function of the investigated system or to use the produced model in virtual docking procedures, it is crucial to obtain its atomistic representation. Figure 3 shows a common, two-step all-atom chain reconstruction approach that consists of (i) generation of backbone coordinates and (ii) reconstruction of residue side chains.

The first group of the backbone reconstruction methods [37, 46, 96, 115] relies on an assembly of fragments derived from Protein Data Bank [7]. In this approach, the most probable fragments are selected using energy-based, homology-based or geometric criteria. Such algorithms can be fast and accurate. However, they have to

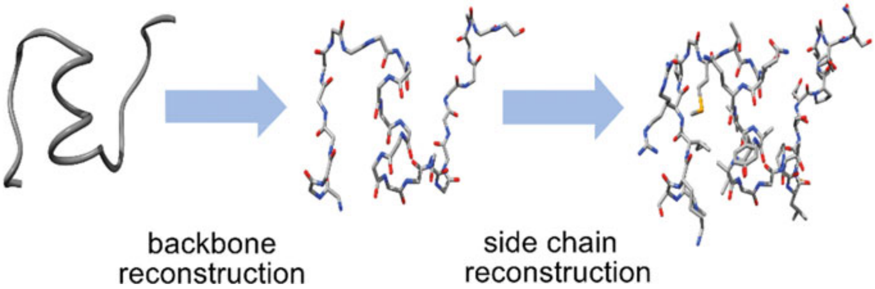

Fig. 3 Illustration of a two-step procedure of all-atom representation reconstruction

maintain large and up-to-date collections of fragments. The second group of methods utilizes averaged knowledge about backbone geometry. The computations are performed based on the statistics of backbone atom positions derived from representative known protein structures.

Methods for side chain position prediction [46, 75, 115] are based on sampling the conformational space by a rotamer library. This involves statistical clustering of observed side chain conformations in known structures. Other algorithms use conformer libraries which contain samples of side chains from known protein structures. In both approaches a scoring function is required to evaluate the quality of the sampled conformations.

The reconstruction to an all-atom representation from a reduced CG representation of the protein is an important part of structure modeling pipelines. Such all-atom models may be directly used for further refinement with molecular mechanics programs [36] and are essential for later structural studies. Most of the post-processing applications, such as structure quality assessment, protein-protein interaction prediction, protein function analysis or ligand docking, require an all-atom model of the protein [121, 143]. There are many tools available for such model conversion [2, 10, 43, 52, 53, 108], but only a small number of them is commonly used. Below, we describe selected servers and applications freely accessible for use online. The time in which all computations are performed by these methods is a matter of seconds to minutes.

4.1 BBQ

The Backbone Building from Quadrilaterals program [37] is a stand-alone application for protein backbone reconstruction from the α-carbon trace. It is available for download from the BioShell website (bioshell.pl). The method uses statistics of backbone atoms positions extracted from a non-redundant database of protein chains to determine backbone coordinates. In this approach, the Cα trace is divided into four residue fragments—quadrilaterals. The quadrilateral conformation is described by three internal coordinates: distances between the four Cα atoms. The coordinates for all four-residue sets of a protein trace are discretized with a mesh size of 0.2 Å. These three distances define a three-dimensional grid in which the average positions of C, O, and N atoms are measured in a local Cartesian coordinate system. The protein sequence is not taken into account in the reconstruction process. The BBQ package was designed to be a fast, robust and as accurate as possible tool for backbone atom reconstruction.

4.2 SABBAC

This online service provides Structural Alphabet-based protein BackBone reconstruction from Alpha-Carbon trace [96]. It is based on a specific approach to protein structure fragment selection and assembly.

The Cα trace is encoded in the Hidden Markov Model-derived structural alphabet which describes conformations of four-residue long fragments [14]. Then, candidate fragments at each position of the structure are chosen from sets of coordinates precomputed in a local reference frame. A full-protein backbone reconstruction is done by joining fragments using a greedy algorithm and searching for the best combination of fragments compatible with the Cα trace. The energy criterion is used to determine the optimality of the combination of fragments.

The SABBAC service has been proven to be fast owing to its fragment library of reduced size. It can be accessed at http://bioserv.rpbs.jussieu.fr/SABBAC.html. During computation, side chains can also be added to the model using Scite [30] which is conveniently combined with SABBAC service.

4.3 SCWRL4

The SCWRL4 algorithm [75] is a method that reconstructs sidechains, based on an input all-atom protein backbone. For each residue type, the input rotamer library provides statistics such as rotamer frequencies and average dihedral angles. Firstly, the input backbone coordinates are checked and side-chain coordinates are built for all rotamers and subrotamers (conformations with dihedral angles ± one standard deviation from the library). Then, self and pairwise energies are computed and rotamers with high self-energy are removed from the reconstruction. To represent the side-chain placement problem, SCWRL4 uses an interaction graph, where vertices represent residues and edges indicate nonzero interactions between them. A Dead End Elimination method is used to find the best rotamer assignment. SCWRL4 is available at http://dunbrack.fccc.edu/scwrl4/.

4.4 MaxSprout

This automatic database procedure [46] for generating the all-atom representation of a protein requires the input Cα trace and amino acid sequence. The computations are split into two basic steps: backbone reconstruction using the Cα trace and side-chain coordinates prediction using the reconstructed backbone.

During backbone construction, a protein structure database is scanned for fragments that locally fit the alpha carbon trace and candidates for a complete overlapping cover of the chain are matched. The optimal continuous path is then found by a dynamic programming algorithm which minimizes the mismatch at protein fragment joints. Final backbone coordinates are taken from fragments superposed on the Cα trace.

Side chain construction starts by generating sets of plausible coordinates from a library of frequently occurring rotamers based on backbone coordinates. Subsequently, all the rotamer-rotamer interaction energies are calculated. To minimize intramolecular energy by an optimized choice of the rotamer, a simple and fast Monte Carlo procedure with simulated annealing is used. When the lowest energy configuration is found, the program returns the coordinates of all-atom representation of the protein. The MaxSprout algorithm is available on-line at http://www.ebi.ac.uk/Tools/maxsprout/.

4.5 PULCHRA

PULCHRA ("Protein Chain Reconstruction Algorithm") [115] is a standalone program for the reconstruction of full-atom protein models from input α-carbon trace and amino acid sequence. The backbone reconstruction step in this approach is very similar to BBQ as both PULCHRA and BBQ implement the same algorithm. BBQ uses the backbone and side chain rotamer libraries, which have been generated from representative protein crystallographic structures.

The side-chain reconstruction procedure uses the same set of distances and coordinates as the backbone reconstruction method. There is a list of possible side-chain conformations which is sorted by the decreasing probability of occurrence in the PDB database, for each combination of calculated distances. The procedure places side-chain heavy atoms on the backbone and optimizes their positions to avoid clashes. In the final step, hydrogen atoms can optionally be added to the full-atom representation. PULCHRA is freely available for download at http://cssb.biology.gatech.edu/PULCHRA.

5 Combining CG Models with Comparative Modeling Methods

For very small proteins, CG methods of structure prediction may provide satisfactory models. However, for the great majority of targets it is necessary to use additional sources of information. The databases of known protein structures are the most easily available among them—e.g., the Protein Data Bank (PDB) [114].

As during the evolution protein structure has become much more strongly conserved than sequence [47], the most straightforward approaches use comparison of sequences of known protein structures (templates) with the query sequence. However, the inability to detect sequence similarity with any of the known structures does not exclude the existence of a good template. The solution in such cases can be so-called threading methods which compare predicted structural features (for example the secondary structure, burial) of the target and the template [36, 122, 128]. Regardless which approach is chosen to detect homology, the aim of this method is to create an alignment, which highlights the similarities between the query and templates.

Obviously, the level of similarity affects the correctness of template selection and the quality of the alignments. For easy cases (high similarity), classical approaches such as PsiBlast [4] almost always provide sufficiently accurate alignment. Therefore, it is relatively easy to build a good model for the query. However, even in those cases CG methods can be useful for the local sampling of some regions, such as loops, which are not defined by the alignment. The difficulty of the problem rapidly increases with the decreasing level of similarity, not only due to the ambiguity of the alignments, but also because of differences in the geometry of correctly aligned regions or suboptimal template selection. One of the most effective approaches to those problems is the incorporation of CG models. Below we present certain strategies that incorporate the information obtained with comparative modeling into the CG-based protocols.

5.1 Reduction of the Sampled Conformational Space

In one of the most straightforward approaches, the query chain is allowed to move in a tube formed by a chain of spheres surrounding the template structure [70]. In this method, the query chain is confined within the tube by imposing energetic penalties for any excursion outside. Thus, the disadvantage of this approach is the limited degree of possible improvement of the initial model.

The answer to this limitation was the application of a more complex set of restraints within GENECOMP [60], a method in which the energy function is constructed in a way which allows two-residue shifts of the target chain along the template. This feature enables changing the initial alignment, and thus correction of possible errors. Additionally, the GENECOMP restraining scheme includes two types of restraints: (i) based on the predicted contacts in the target and (ii) target distances predicted from the fragment threading procedure.

In the more recent studies, the pairwise distances observed in the templates are a source for deriving restraints for the CABS modeling tool [68]. For the number of templates given by comparative modeling procedures, distances between all pairs of Cα atoms are calculated and the minimum and maximum distances between equivalent pairs of residues are taken as limits of the restraint. The restraints are included in the CABS energy function as trapezoid-shaped potential wells, where the gradient of the lateral sides depends on the weight of the restraint. The spatial restraints significantly reduce conformational space, which decreases computation time and increases the probability of obtaining a successful model (see Fig. 4a).

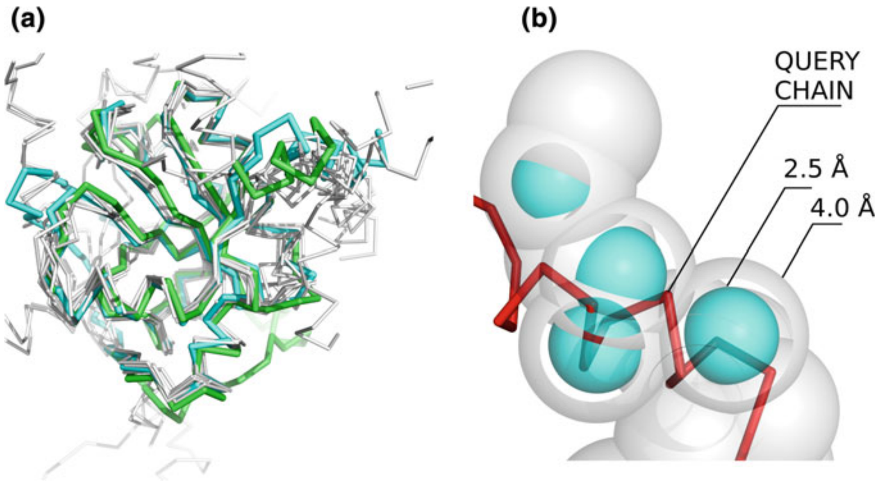

Fig. 4 Sample strategies of combining comparative modeling methods with CG models. **a** T0592 target from CASP9, templates (in gray) define conformational space sampled with CABS. The final model (navy) is more similar (in terms of GDT_TS) to the native (green) than any of the templates. **b** The idea of TRACER. Template scaffold is represented as spheres, query Cα trace as red lines. Query residues within the gray sphere satisfy the free criteria of the query-template pseudo energy, while those within the navy sphere satisfy the additional secondary structure identity criterion (see the text for details)

5.2 Application of the Probability Density Function

A more sophisticated technique was originally used in the Modeller method [117]. In this approach, spatial restraints are defined in terms of a probability density function (PDF). The PDF used for restraining a certain feature x (distance or angle, for instance) can be written as P ($x|A, B \ldots C$). This formula gives a probability density for x when $A, B \ldots C$ are known. For instance, in Rosetta [134], the feature which is restrained is the distance between pairs of Cα atoms (r) and PDF is given as a Gaussian and defined as P($r|G, L, B, D$), where G, L, B and D are predictor variables (see Table 2).

As we know, Gaussian can be defined by two parameters: mean and standard deviation. The latter was calculated using a non-redundant database of nearly 8,000 known protein structures. The HHSearch algorithm [128] was employed to align all pairs of proteins. The standard deviations of r were computed for 10,000 combinations of different G, L, B, D based on differences in the equivalent atoms distances in the aligned structures and put into the four-dimensional table spanned by the values of the predictor variables.

Such a table of standard deviations enables prediction of restraints for a query sequence aligned with the template. For each pair of Cα atoms (apart from those closer than 10 Å or separated by less the 10 residues along the query sequence) the values of four predicting variables are calculated. Then, pairwise distance Gaussian

Table 2 Predictor variables used for deriving restraints for the ROSETTA modeling tool

	Feature	Value
G	Global alignment quality	$-\log(E)$ where E is HHsearch e-value
L	Residue-pair alignment quality	Blosum62 [44] score
B	Burial in the template structure	Number of Cβ's within 8 Å of the template residue Cβ
D	Average distance to an alignment gap	Distances in a number of residues from the aligned pair to the nearest gap in the sequence alignment

L, *B* and *D* are averaged over the pairs of aligned residues, *G* is constant for the given alignment

restraints are assigned: the mean is given by a distance between the equivalent atoms in the template structure, and the standard deviation is taken from the table according to the calculated predictor variables.

It is also possible to combine prediction from the multiple templates as weighted mixture of the Gaussians. Such restraints can be combined with the Rosetta energy function by adding a component equal to $\sum_{i,j} -\ln(P(d_{i,j}))$ where summation is done over pairs of residues, and $P(d_{i,j})$ is the probability of the distance $d_{i,j}$ given by the calculated PDF.

5.3 Unification of Comparative Modeling Methods with CG Models

In the above-mentioned strategies homology inference data (usually in the form of distance restraints) are used as input for CG methods. TRACER [69, 136] is an approach which unifies those two steps. The method uses CABS representation of the protein conformational space and its force field. The most important extension of the model is incorporation of the α-carbon trace template, represented as a fuzzy three-dimensional scaffold with assigned multi-featured properties (Fig. 3b). The query chain is forced to "align" with the template by an additional query-template similarity pseudo energy component introduced to the CABS energy function. This component is a sum over pairs of residues of the query chain and the template that are not further apart than a certain cut-off. The value of query-template similarity pseudo energy for the ith query residue and the jth template residue depends on:

amino acid similarity (quarter of the negative value of the BLOSUM62 substitution matrix [44]; cut-off: 4 Å)
similarity of hydrophobic/hydrophilic features (quarter of the negative value of the product of Kyte-Dollitle indexes [83]; cut-off: 4 Å)

similarity of the orientation and directions of the chains in the vicinity of the ith and jth residue. (−1 if the angle between the flanking Cα-Cα vector is smaller than 90°; cut-off: 4 Å)
identity of the secondary structures (helical or extended) of fragments consisting ith and jth residues (−1 if identical; cut-off: 2.5 Å)

As in the CABS model, the conformational space is sampled by the REMC scheme. The conformational updates include those originally applied in CABS modifications of small fragments (2–4 pseudo-bonds of the Cα trace) and, additionally, rearrangements of larger parts of the chain consisting of up to 22 residues. These larger-scale modifications enable effective sampling of the scaffold, which corresponds to changing the alignments between the query and the template.

TRACER significantly extends the application of comparative modeling methods, especially to regions of very low or even undetectable sequence identity. However, the major drawback of the current version of TRACER, in comparison to some other methods described in this section, is inability to use more than one template.

6 Evaluation of CG Models in CASP9

CASP (Critical Assessment of Techniques for Protein Structure Prediction) [104] is a unique opportunity to evaluate the performance of computational methods in protein structure prediction. CASP is a blind experiment, since the target structures are not published until the end of the prediction period. Therefore, it is possible to fairly assess and compare different prediction methods under the same conditions. The most successful groups, which took part in the CASP9 experiment, employed multistage methodologies, which typically utilize several independent methods, such as: consensus homology detecting tools, modeling methods, quality assessment procedures, optimization and refinement methods. For example, the top ranked group in CASP9, MUFOLD [148], used techniques such as consensus constraints-based model construction and the Multi Dimensional Scaling Technique—a machine learning method for quality assessment—and, finally, model refinement by the combination of model and template information.

On the other hand, if one uses the number of the best models to rank the methods (among all predictions submitted to the CASP9 as the top model—each group may send up to 5 models), the best four methods are based on CG modeling tools described before in Sects. 1, 2 and 3 of this chapter (see Table 3). Below we attempt to briefly evaluate the CG-based methods performance dependence on the difficulty of the targets. Figure 5 shows the comparison of single-method groups presented in Table 3 (except for PRLMS which also utilizes a non-CG Modeller method) taking into account target annotation into categories: FM (Free Modeling), TBM (Template-Based Modeling) and FM/TBM. In the FM category the leading groups (ZHANG_AB_INITIO, BAKER) use fragment-assembling approaches (Rosetta, QUARK). In the intermediate difficulty category, FM/TBM, the LTB (CABS) group

Table 3 Top groups in CASP9 in terms of the number of models with the highest GDT_TS score submitted to CASP9 as first models

Group name (number)	Method	Number of models with the highest GDT_TS	Mean GDT_TS for all server/human targets	Rank in CASP9
PRMLS (65)	Rosetta/Modeller	7	54.10	12
LTB (400)	CABS	5	51.86	28
BAKER (172)	Rosetta	5	51.77	29
ZHANG_AB_INITIO (418)	QUARK	4	52.95	18

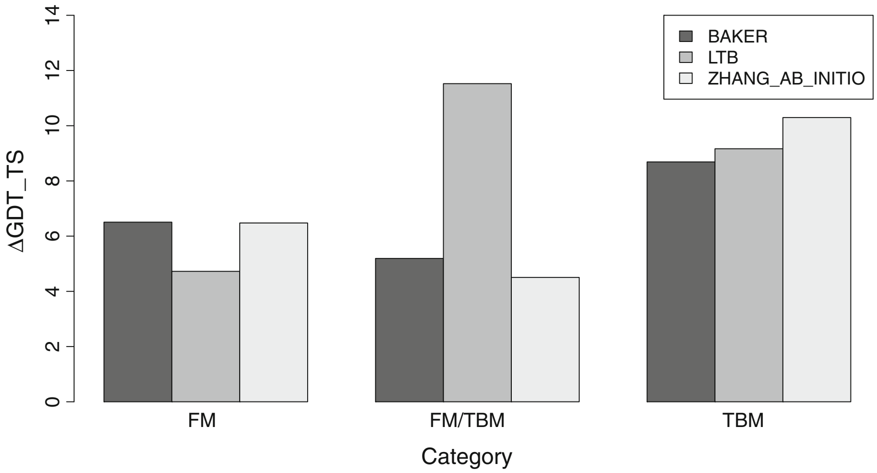

Fig. 5 Differences for three difficulty categories between mean GDT_TS for a particular group and the mean GDT_TS for models submitted to CASP9 by all groups

significantly outperforms two other methods; however, it is necessary to note that this category contains only three targets.

The statistics of the TBM category show that using CG methods in easy cases of comparative modeling is not the best choice. In this category the highest mean quality of the targets was achieved by the ZHANG_AB_INITIO group, which used models provided earlier from automated prediction servers instead of using QUARK. However, it does not mean that CG methods cannot provide successful models. For instance, all the five targets, for which the best model was submitted by the LTB group, belong to the TBM category. The lower mean quality of the models is the effect of the low consistency of prediction quality.

CASP is a biannual experiment initiated over 20 years ago. One of the most intriguing questions regarding this undertaking is the progress in the field. Unfortunately, the evaluation of the progress is not an easy task due to the differences in the difficulty of the targets in various CASP editions. However, a general tendency can be observed that after dramatic improvements in early editions, in the last ones the progress is modest [79].

The latest CASP experiments confirm this relatively slow, however permanent progress in theoretical structure prediction [26, 55, 56, 103]. Combinations of coarse-grained modeling strategies with careful bioinformatics analysis of sequence similarities and final selection/refinement prove to be the most efficient [54, 133, 146, 149].

7 Example Case: CG Prediction of Loop Conformations

The so-called loop closure problem has been a focus of research from the earliest days of computational protein modeling [32, 33, 144]. The prediction of loop structure is often the most difficult challenge in comparative modeling efforts [36]. The accuracy of homology models is usually the lowest in loop regions. Since loop regions often exhibit very low sequence conservation, they have to be modeled without a structural template. In that case, simple homology modeling methods cannot be used. To illustrate some of the applications of the CG approach to protein structure prediction, we briefly review recent modeling efforts using the CABS CG model toward the accurate prediction of protein loops conformation.

In the benchmark study of loop modeling methods [49] the performance of the following tools was compared: MODELLER, ROSETTA, CABS and a combination of MODELLER with CABS. MODELLER [25] is commonly considered a standard comparative modeling package. It employs explicitly designed loop modeling strategies relying on the optimization-based approach (conjugate gradients and molecular dynamics with simulated annealing). ROSETTA and CABS, in turn, employ a knowledge-based driven search of the discretized conformational space. These methods were tested on a large set of loops of various lengths (4–25 residues). The tests showed that classical modeling with MODELLER gives more accurate predictions for short loops, while CG de novo modeling by CABS performs better for longer loops. In the cases of long gaps in protein structures (~20 residues), loops were predicted by CABS with medium or medium-low resolution (RMSD on the level of 2–6 Å from the native). Results of similar quality were obtained for the structure prediction of three extracellular loops of 13 G-protein coupled receptors (GPCRs) by a de novo CABS procedure [59, 61]. This modelling task was particularly challenging for the de novo blind prediction method, as all three extracellular loops were fully flexible during the prediction procedure. Still, the best resulting conformations showed RMSD values lower than 3 Å from the experimental structure (see Fig. 6). Previous benchmark studies, aimed at the prediction of missing protein structure fragments, also indicated that the CG models (an early version of CABS and two other tools based on similar principles) performed relatively well in the range of large fragments [11].

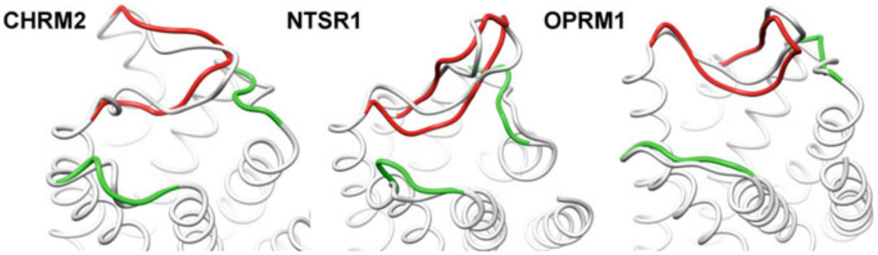

Fig. 6 Structure prediction of GPCR loops using the de novo CABS method [59]. The picture shows the best models for second extracellular loop (EL2) in muscarinic acetylcholine receptor M2 (CHRM2), neurotensin receptor type 1 (NTSR1) and mu-type opioid receptor (OPMR1). The predicted loops are shown in red (EL2) and green (EL1 and EL3), the reference loops (crystallographic structure) and the adjacent intracellular receptor structures in silver. The resulting conformations of the longest EL2 exhibited the following RMSD values from the experimental conformation: 2.65 Å for CHRM2 (15 residue long), 2.99 Å for NTSR1 (21 residues) and 1.92 Å for OPMR1 (17 residues)

As shown by Jamroz and Kolinski [49] CG models can be effectively used for the prediction of loop structures in combination with other techniques. Namely, top ranked models generated by MODELLER were used as multiple templates for CABS modeling. As a result of such a hybrid procedure, the predicted models were on average more accurate than those from the single individual methods.

8 Example Case: CG Molecular Docking of Peptides to Proteins Receptors

Molecular docking is a challenging problem of structural biology and medicine [18, 29]. The subtle energetic effects usually play the main role in docking small ligands to protein or biomolecular complexes. In these cases, a straightforward application of CG models may be difficult or not practical. Docking of large molecules, however, in which the conformational effects are the most important, seems to be a perfect task for CG and/or multiscale modeling strategies. A good example is the flexible and unrestrained docking of peptides to protein receptors. It was possible to allow for significant fluctuations of protein structures, unlimited flexibility of peptide ligands and unrestrained search for docking sites by employing the CABS-based modeling scheme [9, 17, 16, 81, 141]. The CABS-dock protocol is both very efficient and allows for higher flexibility of entire modeled structures than other available tools [84, 95, 112, 120, 119, 135]. The CABS-dock method generates moderate resolution protein-peptide structures for significant fraction of test cases [81]. The resulting lower resolution, coarse-grained structures can be easily refined by classical MD simulations or local docking methods. An example of peptide docking using CABS-dock is illustrated in Fig. 7.

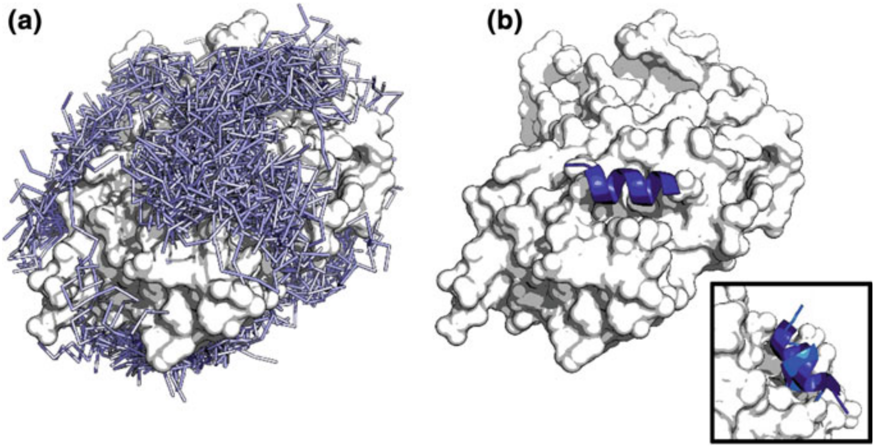

Fig. 7 The figure presents the results of docking nuclear receptor coactivator 1 (sequence: HKLVQLLTTT) to peroxisome proliferator-activated receptor gamma (PDB code 2FVJ:A) without using prior knowledge about the binding site. The docking was performed with CABS-dock method [81]. Panel **a** shows 1000 lowest energy models (light blue, best model RMSD to native pose is 1.43 A) while panel **b** shows the top scored model (dark blue, RMSD to native pose is 3.46 A) together with the experimental structure of bound peptide (light blue) in the close up frame (native complex PDB code: 2FVJ). The protein receptor is presented in surface representation

Protein-peptide docking strategies can also serve as powerful supporting tools for protein-protein docking. Providing a contacting structural fragment from one of the complex components can be predicted with a reasonable fidelity, it may be extracted as a “peptide” fragment. This short linear interacting motif may be docked to the second complex component with protein-peptide docking tools. In some of the modeling cases, this fragmentary template may be successfully used to reconstruct the entire complex [15]. This strategy for hierarchical protein-protein docking is now being intensively studied [109], since protein-protein docking can be of great importance for new directions in drug design [123].

9 Conclusions and Perspectives

One of the main purposes of this chapter was to demonstrate that the most interesting CG models are based on quite complex sets of assumptions, such as protein representation, force field, coordination system and sampling scheme. Obviously, the accuracy of particular assumptions of CG protein models defines the range of applicability of modeling procedures. It seems to be reasonable to state that the future development of CG models will focus on a more accurate reconstruction of real physical effects. Increasing computational power should lead to a considerable decrease in the assumed simplifications of the existing models, and, therefore, provide a more accurate description of the observed physics of biomacromolecules.

Another promising direction of the development of CG models is a more effective combination of existing CG methods with comparative modeling approaches [58, 116, 147]. Perhaps, the term "unification" would be more accurate as we believe that the incorporation of comparative modeling methods should go further than mere utilization of information provided by stand-alone comparative modeling tools. Such a precursor approach has been shown in Sect. 5.3.

Finally, we expect that the development of integrative approaches which use experimental data from various sources together with different computational techniques, as well CG models, will be critical. The most recent (and spectacular) examples of the integrative structure determination include the use of Cryo-Electron Microscopy (cryo-EM) in combination with CG modeling techniques. One of the biggest advantage of Cryo-EM experiments is the fact that, contrary to the popular X-ray crystallography, specimens can be observed in their native environment, which enables the exploration of conformational states. The main problem for Cryo-EM maps is their low resolution which can be solved by the application of CG computational techniques for fitting high-resolution protein structures [150]. Probably, such integrative approaches will become widespread in the near future.

Acknowledgements Maciej Blaszczyk, Sebastian Kmiecik, Katarzyna Ziolkowska and Marta Panek acknowledge support from Foundation for Polish Science TEAM project (TEAM/2011-7/6) co-financed by the European Regional Development Fund operated within the Innovative Economy Operational Program. We also acknowledge support from the National Science Center (NCN Poland) Grant (MAESTRO2014/14/A/ST6/00088).

References

1. Abagyan, R.A., Mazur, A.K.: New methodology for computer-aided modelling of biomolecular structure and dynamics. 2. Local Deformations Cycles J. Biomol. Struct. Dyn. **6**, 833–845 (1989). doi: citeulike-article-id:673543
2. Adcock, S.A.: Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield. J. Comput. Chem. **25**, 16–27 (2004). https://doi.org/10.1002/jcc.10314
3. Altschul, M., Simpson, K.W., Dykes, N.L., Mauldin, E.A., Reubi, J.C., Cummings, J.F.: Evaluation of somatostatin analogues for the detection and treatment of gastrinoma in a dog. J. Small Anim. Pract. **38**, 286–291 (1997)
4. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402 (1997b). doi: gka562 [pii]
5. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science **181**, 223–230 (1973)
6. Anfinsen, C.B., Haber, E., Sela, M., White Jr., F.H.: The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. USA **47**, 1309–1314 (1961)
7. Berman, H., Henrick, K., Nakamura, H.: Announcing the worldwide protein data bank. Nat. Struct. Biol. **10**, 980 (2003). https://doi.org/10.1038/nsb1203-980 nsb1203-980 [pii]

8. Betancourt, M.: A reduced protein model with accurate native-structure identification ability. Proteins **53**, 889–907 (2003). doi: citeulike-article-id:5200969
9. Blaszczyk, M., Kurcinski, M., Kouza, M., Wieteska, L., Debinski, A., Kolinski, A., Kmiecik, S.: Modeling of protein-peptide interactions using the CABS-dock web server for binding site search and flexible docking. Methods **93**, 72–83 (2016). https://doi.org/10.1016/j.ymeth.2015.07.004
10. Blundell, T., et al.: 18th Sir Hans Krebs lecture. Knowl.-Based Protein Model. Design Eur. J. Biochem. **172**, 513–520 (1988)
11. Boniecki, M., Rotkiewicz, P., Skolnick, J., Kolinski, A.: Protein fragment reconstruction using various modeling techniques. J. Comput. Aided Mol. Des. **17**, 725–738 (2003). doi: citeulike-article-id:668480
12. Buchete, N.V., Straub, J.E., Thirumalai, D.: Orientation-dependent coarse-grained potentials derived by statistical analysis of molecular structural databases Polymer **45**, 597–608 (2004). doi: citeulike-article-id:10750645
13. Bystroff, C., Baker, D.: Prediction of local structure in proteins using a library of sequence-structure motifs. J. Mol. Biol. **281**, 565–577 (1998). doi: citeulike-article-id:669894
14. Camproux, A.C., Gautier, R., Tuffery, P.: A hidden markov model derived structural alphabet for proteins. J. Mol. Biol. **339**, 591–605 (2004). https://doi.org/10.1016/j.jmb.2004.04.005s0022283604004085 [pii]
15. Ciemny, M.P., Kurcinski, M., Blaszczyk, M., Kolinski, A., Kmiecik, S.: Modeling EphB4-EphrinB2 protein-protein interaction using flexible docking of a short linear motif. Biomed. Eng. Online **16**, 71 (2017). https://doi.org/10.1186/s12938-017-0362-7
16. Ciemny, M.P., Kurcinski, M., Kozak, K.J., Kolinski, A., Kmiecik, S.: Highly flexible protein-peptide docking using CABS-Dock. Methods Mol. Biol. **1561**, 69–94 (2017). https://doi.org/10.1007/978-1-4939-6798-8_6
17. Ciemny, M.P., Debinski, A., Paczkowska, M., Kolinski, A., Kurcinski, M., Kmiecik, S.: Protein-peptide molecular docking with large-scale conformational changes: the p 53-MDM2 interaction. Sci. Rep. **6**, 37532 (2016). https://doi.org/10.1038/srep37532
18. Ciemny, M., Kurcinski, M., Kamel, K., Kolinski, A., Alam, N., Schueler-Furman, O., Kmiecik, S.: Protein–peptide docking: opportunities and challenges. Drug Discov. Today **23**(8), 1530–1537, ISSN 1359-6446 (2018). https://doi.org/10.1016/j.drudis.2018.05.006
19. Covell, D.G.: Folding protein alpha-carbon chains into compact forms by Monte Carlo methods. Proteins **14**, 409–420 (1992). https://doi.org/10.1002/prot.340140310
20. Czaplewski, C., Liwo, A., Makowski, M., Ołdziej, S., Scheraga, H.A.: Coarse-grained models of proteins: theory and applications. In: Kolinski, A. (ed.) Multiscale approaches to protein modeling, pp. 85–109. Springer, New York (2011)
21. Czaplewski, C., Rodziewicz-Motowidlo, S., Liwo, A., Ripoll, D.R., Wawak, R.J., Scheraga, H.A.: Molecular simulation study of cooperativity in hydrophobic association. Protein Sci. **9**, 1235–1245 (2000). https://doi.org/10.1110/ps.9.6.1235
22. Dashevskii, V.G.: [Lattice model for globular protein three-dimensional structure] Mol. Biol. (Mosk) **14**, 105–117 (1980)
23. Dawid, A.E., Gront, D., Kolinski, A.: SURPASS low-resolution coarse-grained protein modeling. J. Chem. Theor. Comput. **13**, 5766–5779 (2017). https://doi.org/10.1021/acs.jctc.7b00642
24. De Sancho, D., Rey, A.: Evaluation of coarse grained models for hydrogen bonds in proteins. J. Comput. Chem. **28** (2007). doi: citeulike-article-id:1127406
25. Eswar, N., Eramian, D., Webb, B., Shen, M.Y., Sali, A.: Protein structure modeling with MODELLER. Methods Mol. Biol. **426**, 145–159 (2008). https://doi.org/10.1007/978-1-60327-058-8_8
26. Feig, M., Mirjalili, V.: Protein structure refinement via molecular-dynamics simulations: what works and what does not? Proteins **84**(Suppl 1), 282–292 (2016). https://doi.org/10.1002/prot.24871
27. Ferrenberg, A., Landau, D.P., Swendsen, R.: Statistical errors in histogram reweighting. Phys. Rev. E **51**, 5092 (1995). doi:citeulike-article-id:875595

28. Ferrenberg, A., Swendsen, R.: Optimized Monte Carlo data analysis. Phys. Rev. Lett. **63**, 1195–1198 (1989). doi:citeulike-article-id:774372
29. Fosgerau, K., Hoffmann, T.: Peptide therapeutics: current status and future directions. Drug Discov. Today **20**, 122–128 (2015). https://doi.org/10.1016/j.drudis.2014.10.003
30. Gautier, R., Camproux, A.C., Tuffery, P.: SCit: web tools for protein side chain conformation analysis. Nucleic Acids Res. **32**, W508–511 (2004). https://doi.org/10.1093/nar/gkh38832/suppl_2/w508 [pii]
31. Geyer, C.J.: Markov chain Monte Carlo maximum likelihood. In: Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation. Fairfax Station, pp. 156–163 (1991). doi: citeulike-article-id:606345
32. Go, N., Scheraga, H.: Ring closure and local conformational deformations of chain molecules. Macromolecules **3**, 178–187 (1970)
33. Go, N., Scheraga, H.A.: Ring-Closure in Chain Molecules with Cn, I, or S2n Symmetry. Macromolecules **6**, 273–281 (1973)
34. Godzik, A., Kolinski, A., Skolnick, J.: Lattice representations of globular proteins: how good are they? J. Comput. Chem. **14**, 1194–1202 (1993). https://doi.org/10.1002/jcc.540141009
35. Grishaev, A., Bax, A.: An empirical backbone–backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. J. Am. Chem. Soc. **126**, 7281–7292 (2004). doi: citeulike-article-id:1896684
36. Gront, D., Kmiecik, S., Blaszczyk, M., Ekonomiuk, D., Koliński, A.: Optimization of protein models Wiley interdisciplinary reviews: computational molecular. Science **2**, 479–493 (2012). https://doi.org/10.1002/wcms.1090
37. Gront, D., Kmiecik, S., Kolinski, A.: Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. J. Comput. Chem. **28**, 1593–1597 (2007). https://doi.org/10.1002/jcc.20624
38. Gront, D., Kolinski, A., Skolnick, J.: Comparison of three Monte Carlo conformational search strategies for a proteinlike homopolymer model: Folding thermodynamics and identification of low-energy structures. J. Chem. Phys. **113**, 5065–5071 (2000). doi: citeulike-article-id:606324
39. Gront, D., Kolinski, A., Skolnick, J.: A new combination of replica exchange Monte Carlo and histogram analysis for protein folding and thermodynamics. J. Chem. Phys. **115**, 1569–1574 (2001). doi: citeulike-article-id:876359
40. Gront, D., Kulp, D., Vernon, R., Strauss, C., Baker, D.: Generalized fragment picking in rosetta: design, protocols and applications. PLoS ONE **6**, e23294 (2011). doi: citeulike-article-id:9705043
41. Guardiani, C., Livi, R., Cecconi, F.: Coarse Grained Modeling and Approaches to Protein Folding. Curr. Bioinform. **5**, 217–240 (2010)
42. Hansmann, U.: parallel tempering algorithm for conformational studies of biological molecules. Chem. Phys. Lett. **281**, 140–150 (1997). doi: citeulike-article-id:715765
43. Heath, A.P., Kavraki, L.E., Clementi, C.: From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. Proteins **68**, 646–661 (2007). https://doi.org/10.1002/prot.21371
44. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**, 10915–10919 (1992)
45. Hinds, D.A., Levitt, M.: A lattice model for protein structure prediction at low resolution. Proc. Natl. Acad. Sci. USA **89**, 2536–2540 (1992)
46. Holm, L., Sander, C.: Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. J. Mol. Biol. **218**, 183–194 (1991). doi: 0022-2836(91)90883-8 [pii]
47. Illergard, K., Ardell, D.H., Elofsson, A.: Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. Proteins **77**, 499–508 (2009). https://doi.org/10.1002/prot.22458
48. Irbäck, A., Mohanty, S.: PROFASI: A Monte Carlo simulation package for protein folding and aggregation. J. Comput. Chem. **27**, 1548–1555 (2006). doi: citeulike-article-id:7290910

49. Jamroz, M., Kolinski, A.: Modeling of loops in proteins: a multi-method approach. BMC Struct. Biol. **10**, 5+ (2010)
50. Jones, T.A., Thirup, S.: Using known substructures in protein model building and crystallography. EMBO J. **5**, 819–822 (1986). doi: citeulike-article-id:705742
51. Karplus, M., McCammon, J.A.: Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. **9**, 646–652 (2002). https://doi.org/10.1038/nsb0902-646nsb0902-646 [pii]
52. Kazmierkiewicz, R., Liwo, A., Scheraga, H.A.: Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a Monte-Carlo method. J. Comput. Chem. **23**, 715–723 (2002). https://doi.org/10.1002/jcc.10068 [pii]
53. Kazmierkiewicz, R., Liwo, A., Scheraga, H.A.: Addition of side chains to a known backbone with defined side-chain centroids. Biophys. Chem. **100**, 261–280 (2003). doi: S0301462202002855 [pii]
54. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J.: The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. **10**, 845–858 (2015). https://doi.org/10.1038/nprot.2015.053nprot.2015.053 [pii]
55. Kim, H., Kihara, D.: Protein structure prediction using residue- and fragment-environment potentials in CASP11. Proteins **84**(Suppl 1), 105–117 (2016). https://doi.org/10.1002/prot.24920
56. Kinch, L.N., Li, W., Monastyrskyy, B., Kryshtafovych, A., Grishin, N.V.: Evaluation of free modeling targets in CASP11 and ROLL. Proteins **84**(Suppl 1), 51–66 (2016). https://doi.org/10.1002/prot.24973
57. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**, 671–680 (1983). doi: citeulike-article-id:379797
58. Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A.E., Kolinski, A.: Coarse-grained protein models and their applications. Chem. Rev. **116**, 7898–7936 (2016). https://doi.org/10.1021/acs.chemrev.6b00163
59. Kmiecik, S., Jamroz, M., Kolinski, M.: Structure prediction of the second extracellular loop in G-protein-coupled receptors. Biophys. J. **106**, 2408–2416 (2014). https://doi.org/10.1016/j.bpj.2014.04.022
60. Kolinski, A., Betancourt, M.R., Kihara, D., Rotkiewicz, P., Skolnick, J.: Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins **44**, 133–149 (2001)
61. Kolinski, M., Filipek, S.: Study of a structurally similar kappa opioid receptor agonist and antagonist pair by molecular dynamics simulations. J. Mol. Model. **16**, 1567–1576 (2010). https://doi.org/10.1007/s00894-010-0678-8
62. Kolinski, A., Galazka, W., Skolnick, J.: Computer design of idealized beta-motifs. J. Chem. Phys. **103**, 10286–10297 (1995)
63. Kolinski, A., Ilkowski, B., Skolnick, J.: Dynamics and thermodynamics of beta-hairpin assembly: insights from various simulation techniques. Biophys. J. **77**, 2942–2952 (1999)
64. Kolinski, A., Milik, M., Rycombel, J., Skolnick, J.: A reduced model of short-range interactions in polypeptide-chains. J. Chem. Phys. **103**, 4312–4323 (1995)
65. Kolinski, A., Milik, M., Skolnick, J.: Static and dynamic properties of a new lattice model of polypeptide-chains. J. Chem. Phys. **94**, 3978–3985 (1991)
66. Kolinski, A., Skolnick, J.: Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins **18**, 338–352 (1994). https://doi.org/10.1002/prot.340180405
67. Kolinski, A., Skolnick, J.: Reduced models of proteins and their applications. Polymer **45**, 511–524 (2004). https://doi.org/10.1016/j.polymer.2003.10.064
68. Kolinski, A.: Protein modeling and structure prediction with a reduced representation. Acta Biochimica. Polonica **51**, 349–371 (2004). doi: citeulike-article-id:606304
69. Kolinski, A., Gront, D.: Comparative modeling without implicit sequence alignments. Bioinformatics **23**, 2522–2527 (2007). doi: btm380 [pii]https://doi.org/10.1093/bioinformatics/btm380
70. Kolinski, A., Rotkiewicz, P., Ilkowski, B., Skolnick, J.: A method for the improvement of threading-based protein models. Proteins **37**, 592–610 (1999b). https://doi.org/10.1002/(sici)1097-0134(19991201)37:4%3c592::aid-prot10%3e3.0.co;2-2 [pii]

71. Kolinski, A., Skolnick, J.: Lattice Models of Protein Folding, Dynamics and Thermodynamics. Landes (1996). doi: citeulike-article-id:877252
72. Kolinski, A., Skolnick, J.: Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins **32**, 475–494 (1998). https://doi.org/10.1002/(sici)1097-0134(19980901)32:4%3c475::aid-prot6%3e3.0.co;2-f [pii]
73. Kortemme, T., Morozov, A.V., Baker, D.: An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J. Mol. Biol. **326**, 1239–1259 (2003). doi: citeulike-article-id:556189
74. Krigbaum, W.R., Lin, S.F.: Monte-Carlo simulation of protein folding using a lattice model. Macromolecules **15**, 1135–1145 (1982)
75. Krivov, G.G., Shapovalov, M.V., Dunbrack Jr., R.L.: Improved prediction of protein side-chain conformations with SCWRL4. Proteins **77**, 778–795 (2009). https://doi.org/10.1002/prot.22488
76. Krupa, P., Mozolewska, M.A., Joo, K., Lee, J., Czaplewski, C., Liwo, A.: Prediction of protein structure by template-based modeling combined with the UNRES force field. J. Chem. Inf. Model. **55**, 1271–1281 (2015). https://doi.org/10.1021/acs.jcim.5b00117
77. Krupa, P., Sieradzan, A.K., Mozolewska, M.A., Li, H., Liwo, A., Scheraga, H.A.: Dynamics of disulfide-bond disruption and formation in the thermal unfolding of ribonuclease A. J. Chem. Theor. Comput. **13**, 5721–5730 (2017). https://doi.org/10.1021/acs.jctc.7b00724
78. Krupa, P., et al.: Performance of protein-structure predictions with the physics-based UNRES force field in CASP11. Bioinformatics **32**, 3270–3278 (2016). doi:btw404 [pii]https://doi.org/10.1093/bioinformatics/btw404
79. Kryshtafovych, A., Fidelis, K., Moult, J.: CASP9 results compared to those of previous CASP experiments. Proteins **79**(Suppl 10), 196–207 (2011). https://doi.org/10.1002/prot.23182
80. Kumar, S., Rosenberg, J., Bouzida, D., Swendsen, R., Kollman, P.: Multidimensional free-energy calculations using the weighted histogram analysis method. J. Comput. Chem. **16**, 1339–1350 (1995). doi: citeulike-article-id:774417
81. Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A., Kmiecik, S.: CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. Nucleic Acids Res. **43**, W419–424 (2015). https://doi.org/10.1093/nar/gkv456gkv456 [pii]
82. Kwak, W., Hansmann, U.H.: Efficient sampling of protein structures by model hopping. Phys. Rev. Lett. **95**, 138102 (2005). https://doi.org/10.1103/PhysRevLett.95.138102
83. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**, 105–132 (1982). doi: 0022-2836(82)90515-0 [pii]
84. Lee, H., Heo, L., Lee, M.S., Seok, C.: GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. Nucleic Acids Res. **43**, W431–W435 (2015). https://doi.org/10.1093/nar/gkv495
85. Lee, J., Scheraga, H.A., Rackovsky, S.: New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J. Comput. Chem. **18**, 1222–1232 (1997)
86. Levitt, M.: A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. **104**, 59–107 (1976). doi: citeulike-article-id:4000523
87. Levitt, M., Warshel, A.: Computer simulation of protein folding. Nature **253**, 694–698 (1975). doi: citeulike-article-id:4275709
88. Levy-Moonshine, A., Amir, E-a. D., Keasar, C.: Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. Bioinformatics **25**, 2639–2645 (2009). doi: citeulike-article-id:7012194
89. Li, Z., Scheraga, H.A.: Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc. Natl. Acad. Sci. USA **84**, 6611–6615 (1987)
90. Liwo, A., He, Y., Scheraga, H.A.: Coarse-grained force field: general folding theory. Phys. Chem. Chem. Phys. **13**, 16890–16901 (2011). https://doi.org/10.1039/c1cp20752k
91. Liwo, A., et al.: Simulation of Protein Structure and Dynamics with the Coarse-Grained UNRES Force Field. Coarse-Graining of Condensed Phase and Biomolecular Systems. CRC Press (2008). doi: citeulike-article-id:3822586

92. Liwo, A., Czaplewski, C., Pillardy, J., Scheraga, H.: Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. J. Chem. Phys. **115**, 2323–2347 (2001). doi: citeulike-article-id:715745
93. Liwo, A., Khalili, M., Scheraga, H.: Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc. Natl. Acad. Sci. U.S.A. **102**, 2362–2367 (2005). doi: citeulike-article-id:1365687
94. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S., Scheraga, H.A.: Prediction of protein conformation on the basis of a search for compact structures: test on avian pancreatic polypeptide. Protein Sci.: Publ. Protein Soc. **2**, 1715–1731 (1993). doi: citeulike-article-id:7558759
95. London, N., Raveh, B., Cohen, E., Fathi, G., Schueler-Furman, O.: Rosetta FlexPepDock web server–high resolution modeling of peptide-protein interactions. Nucleic Acids Res. **39**, W249–W253 (2011). https://doi.org/10.1093/nar/gkr431
96. Maupetit, J., Gautier, R., Tuffery, P.: SABBAC: Online structural alphabet-based protein BackBone reconstruction from alpha-carbon trace. Nucleic Acids Res. **34**, W147–151 (2006). doi: 34/suppl_2/W147 [pii]https://doi.org/10.1093/nar/gkl289
97. Mazur, A.K., Dorofeev, V.E., Abagyan, R.A.: Derivation and testing of explicit equations of motion for polymers described by internal coordinates. J. Comput. Phys. **92**, 261–272 (1991). doi: citeulike-article-id:10750684
98. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953). doi: citeulike-article-id:531300
99. Metropolis, N., Ulam, S.: The Monte Carlo method. J. Am. Stat. Assoc. **44**, 335–341 (1949). doi: citeulike-article-id:1886002
100. Milik, M., Kolinski, A., Skolnick, J.: Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. J. Comput. Chem. **18**, 80–85 (1997)
101. Mitsutake, A., Sugita, Y., Okamoto, Y.: Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers **60**, 96–123 (2001). https://doi.org/10.1002/1097-0282(2001)60:2%3c96::aid-bip1007%3e3.0.co;2-f [pii]https://doi.org/10.1002/1097-0282(2001)60:2%3c96::AID-BIP1007%3e3.0.CO;2-F
102. Morozov, A., Lin, S.: Accuracy and convergence of the Wang-Landau sampling algorithm. Phys. Rev. E (Statistical, Nonlinear, and Soft Matter Physics) **76** (2007). doi: citeulike-article-id:3802626
103. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A.: Critical assessment of methods of protein structure prediction (CASP)-round XII. Proteins (2017). https://doi.org/10.1002/prot.25415
104. Moult, J., Fidelis, K., Kryshtafovych, A., Tramontano, A.: Critical assessment of methods of protein structure prediction (CASP)–round IX. Proteins **79**(Suppl 10), 1–5 (2011). https://doi.org/10.1002/prot.23200
105. Mozolewska, M.A., Krupa, P., Zaborowski, B., Liwo, A., Lee, J., Joo, K., Czaplewski, C.: Use of restraints from consensus fragments of multiple server models to enhance protein-structure prediction capability of the UNRES force field. J. Chem. Inf. Model. **56**, 2263–2279 (2016). https://doi.org/10.1021/acs.jcim.6b00189
106. Park, B.H., Levitt, M.: The complexity and accuracy of discrete state models of protein structure. J. Mol. Biol. **249**, 493–507 (1995). doi: citeulike-article-id:5845728
107. Parsons, J., Holmes, B., Rojas, M., Tsai, J., Strauss, C.: Practical conversion from torsion space to Cartesian space forin silico protein synthesis. J. Comput. Chem. **26**, 1063–1068 (2005). doi: citeulike-article-id:1036763
108. Payne, P.W.: Reconstruction of protein conformations from estimated positions of the C-alpha coordinates. Protein Sci. **2**, 315–324 (1993)
109. Peterson, L.X., et al.: Modeling the assembly order of multimeric heteroprotein complexes. PLoS Comput. Biol. **14**, e1005937 (2018). https://doi.org/10.1371/journal.pcbi.1005937pcompbiol-d-17-00872 [pii]
110. Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **35**, D61–65 (2007). doi: gkl842 [pii]https://doi.org/10.1093/nar/gkl842

111. Pundir, S., Martin, M.J., O'Donovan, C.: UniProt protein knowledgebase methods. Mol. Biol. **1558**, 41–55 (2017). https://doi.org/10.1007/978-1-4939-6783-4_2
112. Raveh, B., London, N., Zimmerman, L., Schueler-Furman, O.: Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. PLoS ONE **6**, e18934 (2011). https://doi.org/10.1371/journal.pone.0018934
113. Rohl, C., Strauss, C., Misura, K., Baker, D.: Protein structure prediction using Rosetta. In: Numerical Computer Methods, Part D, vol. 383, pp. 66–93. Elsevier (2004). doi: citeulike-article-id:441859
114. Rose, P.W., et al.: The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. **45**, D271–D281 (2017). https://doi.org/10.1093/nar/gkw1000
115. Rotkiewicz, P., Skolnick, J.: Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem. **29**, 1460–1465 (2008). https://doi.org/10.1002/jcc.20906
116. Sali, A., et al.: Outcome of the first wwPDB hybrid/integrative methods task force workshop. Structure **23**, 1156–1167 (2015). https://doi.org/10.1016/j.str.2015.05.013
117. Sali, A., Blundell, T.L.: Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**, 779–815 (1993). doi: S0022-2836(83)71626-8 [pii] https://doi.org/10.1006/jmbi.1993.1626
118. Scheraga, H.A., Khalili, M., Liwo, A.: Protein-folding dynamics: overview of molecular simulation techniques. Annu. Rev. Phys. Chem. **58**, 57–83 (2007). https://doi.org/10.1146/annurev.physchem.58.032806.104614
119. Schindler, C.E., de Vries, S.J., Zacharias, M.: iATTRACT: simultaneous global and local interface optimization for protein-protein docking refinement. Proteins **83**, 248–258 (2015). https://doi.org/10.1002/prot.24728
120. Schindler, C.E., de Vries, S.J., Zacharias, M.: Fully blind peptide-protein docking with pepATTRACT. Structure **23**, 1507–1515 (2015a). https://doi.org/10.1016/j.str.2015.05.021s0969-2126(15)00224-5 [pii]
121. Shenoy, S.R., Jayaram, B.: Proteins: sequence to structure and function–current status. Curr. Protein Pept. Sci. **11**, 498–514 (2010)
122. Shi, J., Blundell, T.L., Mizuguchi, K.: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. **310**, 243–257 (2001). https://doi.org/10.1006/jmbi.2001.4762s0022-2836(01)94762-x [pii]
123. Shin, W.H., Christoffer, C.W., Kihara, D.: In silico structure-based approaches to discover protein-protein interaction-targeting drugs. Methods **131**, 22–32 (2017). doi: S1046-2023(17)30208-6 [pii]https://doi.org/10.1016/j.ymeth.2017.08.006
124. Sippl, M.J.: Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J. Comput. Aided Mol. Des. **7**, 473–501 (1993)
125. Skolnick, J., Kolinski, A.: Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six-member, Greek key beta-barrel proteins. J. Mol. Biol. **212**, 787–817 (1990a). doi:0022-2836(90)90237-G [pii]
126. Skolnick, J., Kolinski, A.: Simulations of the folding of a globular protein. Science **250**, 1121–1125 (1990b). doi: 250/4984/1121 [pii]https://doi.org/10.1126/science.250.4984.1121
127. Skolnick, J., Kolinski, A., Brooks III, C.L., Godzik, A., Rey, A.: A method for predicting protein structure from sequence. Curr. Biol. **3**, 414–423 (1993). doi:0960-9822(93)90348-R [pii]
128. Soding, J.: Protein homology detection by HMM-HMM comparison. Bioinformatics **21**, 951–960 (2005). doi: bti125 [pii]https://doi.org/10.1093/bioinformatics/bti125
129. Stein, A., Kortemme, T.: Improvements to robotics-inspired conformational sampling in rosetta. PLoS ONE **8**, e63090 (2013). https://doi.org/10.1371/journal.pone.0063090pone-d-13-06862 [pii]

130. Stumpff-Kane, A.W., Maksimiak, K., Lee, M.S., Feig, M.: Sampling of near-native protein conformations during protein structure refinement using a coarse-grained model, normal modes, and molecular dynamics simulations. Proteins **70**, 1345–1356 (2008). https://doi.org/10.1002/prot.21674
131. Sugita, Y., Okamoto, Y.: Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. **314**, 141–151 (1999). doi:citeulike-article-id:197524
132. Swendsen, R., Wang, J.: Replica Monte Carlo simulation of spin-glasses. Phys. Rev. Lett. **57**, 2607–2609 (1986). doi: citeulike-article-id:773436
133. Tai, C.H., Bai, H., Taylor, T.J., Lee, B.: Assessment of template-free modeling in CASP10 and ROLL. Proteins **82**(Suppl 2), 57–83 (2014). https://doi.org/10.1002/prot.24470
134. Thompson, J., Baker, D.: Incorporation of evolutionary information into Rosetta comparative modeling. Proteins **79**, 2380–2388 (2011). https://doi.org/10.1002/prot.23046
135. Trabuco, L.G., Lise, S., Petsalaki, E., Russell, R.B.: PepSite: prediction of peptide-binding sites from protein surfaces. Nucleic Acids Res. **40**, W423–W427 (2012). https://doi.org/10.1093/nar/gks398
136. Trojanowski, S., Rutkowska, A., Kolinski, A.: TRACER. A new approach to comparative modeling that combines threading with free-space conformational sampling. Acta Biochim. Pol. **57**, 125–133 (2010)
137. UniProt: the universal protein knowledgebase. Nucleic Acids Res. **45**, D158-D169 (2017) https://doi.org/10.1093/nar/gkw1099
138. Vendruscolo, M., Najmanovich, R., Domany, E.: Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins **38**, 134–148 (2000). https://doi.org/10.1002/(sici)1097-0134(20000201)38:2%3c134::aid-prot3%3e3.0.co;2-a [pii]
139. Vinals, J., Kolinski, A., Skolnick, J.: Numerical study of the entropy loss of dimerization and the folding thermodynamics of the GCN4 leucine zipper. Biophys. J. **83**, 2801–2811 (2002). doi: S0006-3495(02)75289-2 [pii]https://doi.org/10.1016/s0006-3495(02)75289-2
140. Voth, G. (ed): Coarse-Graining of Condensed Phase and Biomolecular Systems. CRC Press Taylor & Francis, Farmington, CT (2008)
141. Wabik, J., Kurcinski, M., Kolinski, A.: Coarse-grained modeling of peptide docking associated with large conformation transitions of the binding protein: Troponin I fragment-Troponin C system. Molecules **20**, 10763–10780 (2015). https://doi.org/10.3390/molecules200610763
142. Wales, D.: Energy Landscapes: Applications to Clusters, Biomolecules and Glasses (Cambridge Molecular Science). Cambridge University Press (2004). doi: citeulike-article-id:755112
143. Wang, T., Wu, M.B., Zhang, R.H., Chen, Z.J., Hua, C., Lin, J.P., Yang, L.R.: Advances in computational structure-based drug design and application in drug discovery. Curr. Top Med. Chem. **16**, 901–916 (2016). doi: CTMC-EPUB-69847 [pii]
144. Wedemeyer, W.J., Scheraga, H.A.: Exact analytical loop closure in proteins using polynomial equations. J. Comput. Chem. **20**, 819–844 (1999)
145. Xu, D., Zhang, J., Roy, A., Zhang, Y.: Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins **79**(Suppl 10), 147–160 (2011). https://doi.org/10.1002/prot.23111
146. Yan, C.H., et al.: Minimal residual disease- and graft-vs.-host disease-guided multiple consolidation chemotherapy and donor lymphocyte infusion prevent second acute leukemia relapse after allotransplant. J. Hematol. Oncol. **9**, 87 (2016). https://doi.org/10.1186/s13045-016-0319-5
147. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., Zhang, Y.: The I-TASSER Suite: protein structure and function prediction. Nat. Methods **12**(1), 7–8 (2015). https://doi.org/10.1038/nmeth.3213
148. Zhang, J., He, Z., Wang, Q., Barz, B., Kosztin, I., Shang, Y., Xu, D.: Prediction of protein tertiary structures using MUFOLD methods. Mol. Biol. **815**, 3–13 (2012). https://doi.org/10.1007/978-1-61779-424-7_1

149. Zhang, Y.: Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins **82**(Suppl 2), 175–187 (2014). https://doi.org/10.1002/prot.24341

150. Zheng, W.: Accurate flexible fitting of high-resolution protein structures into cryo-electron microscopy maps using coarse-grained pseudo-energy minimization. Biophys. J. **100**, 478–488 (2011). doi: S0006-3495(10)05186-6 [pii]

151. Zhou, H., Zhou, Y.: Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. **11**, 2714–2726 (2002). https://doi.org/10.1110/ps.0217002

152. Zhou, R., et al.: Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements. Proc. Natl. Acad. Sci. U.S.A. **111**, 18243–18248 (2014). https://doi.org/10.1073/pnas.1420914111 1420914111 [pii]