Computational Methods for Protein Fold Prediction: an Ab-initio Topological Approach

G. Ceci[1,5], A. Mucherino[1,5], M. D'Apuzzo[1,4,5], D. Di Serafino[1,4,5*], S. Costantini[2,3,5], A. Facchiano[3,4,5], and G. Colonna[2,4,5]

[1] Department of Mathematics, Second University of Naples
via Vivaldi 43, I-81100 Caserta, Italy
[2] Department of Biochemistry and Biophysics, Second University of Naples,
via Costantinopoli 16, I-80138 Naples, Italy
[3] Institute of Food Science, CNR, via Roma 52 A/C, I-83100 Avellino, Italy
[4] Research Center of Computational and Biotechnological Sciences (CRISCEB),
Second University of Naples, via Costantinopoli 16, I-80138 Naples, Italy
[5] Computational Biology Doctorate, Second University of Naples, Italy

Summary. The prediction of protein native conformations is still a big challenge in science, although a strong research activity has been carried out on this topic in the last decades. In this chapter we focus on ab-initio computational methods for protein fold predictions that do not rely heavily on comparisons with known protein structures and hence appear to be the most promising methods for determining conformations not yet been observed experimentally. To identify main trends in the research concerning protein fold predictions, we briefly review several ab-initio methods, including a recent topological approach that models the protein conformation as a tube having maximum thickness without any self-contacts. This representation leads to a constrained global optimization problem. We introduce a modification in the tube model to increase the compactness of the computed conformations, and present results of computational experiments devoted to simulating α-helices and all-α proteins. A Metropolis Monte Carlo Simulated Annealing algorithm is used to search the protein conformational space.

Key words: Protein fold prediction, Ab-initio methods, Native state topology, Tube thickness, Global optimization, Simulated annealing

1 Introduction

Proteins are heteropolymers that control and regulate many vital functions [66, 67, 68], hence they are considered the building blocks of living organisms. A protein is made of a sequence of amino acid residues connected by peptide

* Corresponding author. Email: daniela.diserafino@unina2.it.

bonds, called *primary structure*, which folds into a unique three-dimensional conformation, called *tertiary structure* or *native state*. The biological function of a protein is largely determined by its native state; the knowledge of the native state is therefore critical in understanding the role of the protein in the cell and the related molecular mechanisms. Levinthal's paradox [48] and Anfinsen's experiment [5] suggest that the Nature applies an "algorithm" to drive a protein from its primary structure to its own tertiary structure, and that the information needed to perform this algorithm is contained in the primary structure. Understanding the *protein folding problem* means understanding and reproducing this algorithm.

Many scientists have been working on the protein folding problem for nearly half a century. A growing interest in its solution has been observed during the years, because of its impact in several research fields, such as genetic disease treatment, drug design, and the emerging structural and functional genomics. However, despite the research has been very active, we are still far from a clear and full explanation of the protein folding mechanisms and this problem is still considered a big challenge in science.

Different computational approaches to the protein fold prediction have been developed. We focus our attention on the so-called *ab-initio* methods that do not rely heavily on comparisons with known protein structures and appear to be the most promising for determining three-dimensional conformations that have not yet observed experimentally. These methods are usually based on suitable representations of the polypeptide chain and on suitable energy functions reproducing physicochemical interactions among protein atoms. According to Anfinsen's hypotesis, the native state corresponds to the minimum energy of the system and its determination requires the solution of a (computationally demanding) global constrained optimization problem.

Recent studies have emphasized the role of the *topology of the native state* in the protein folding process [11, 42, 69, 80]. In this context, an ab-initio method has been developed that takes into account mainly topological rather than physicochemical features of the protein [7, 8, 9, 10, 54]. This method is based on a very simplified model that represents the polymer chain as a tube of nonzero thickness, without self-contacts. As in other approaches, this formulation leads to a constrained global optimization problem. In this chapter we present a modified version of this model, discuss the choice of model parameters and show results of computational experiments devoted to simulating α-helices and all-α proteins.

The chapter is organized as follows. In Section 2 we provide a very short description of the chemical structure of a protein to better understand the terminology used in the remainder of the chapter. In Section 3 we introduce the three main computational approaches to the protein fold prediction problem: *homology modeling*, *fold recognition* and *ab-initio methods*. In Sections 4 and 5 we provide a brief description of energy functions and global optimization techniques, that characterize a variety of ab-initio approaches. Following Klepeis and Floudas [37], ab-initio methods can be further classified as ab-initio meth-

ods that require database information and "true" ab-initio methods, that are based only on information obtained from physicochemical principles. A survey of methods falling into both classes is provided in Sections 6 and 7. Among the true ab-initio methods, we present also recent approaches based on topological features of the proteins. This survey is not meant to be exhaustive; it rather gives an idea of the evolution of main trends in the ab-initio protein folding research, along with successes and limitations. In Section 8 we focus on a specific topological model and give the mathematical description of the corresponding constrained global optimization problem, while in Section 9 we discuss how the values of the model parameters have been chosen. In Section 10, after a short presentation of the Simulated Annealing algorithm used to solve the optimization problem, we report results of our computational experiments. A few concluding remarks are given in Section 11.

2 The Chemical Structure of a Protein

A protein is a polymer composed by a sequence of genetically driven amino acid residues. Proteins in living cells are built from a set of only 20 different amino acids, all having two main substructures: a common basic substructure composed by an amide group (NH_2), a carboxyl group (COOH) and a hydrogen atom (H), all linked to a central carbon atom called C_α, and a substructure that differentiates each amino acid, called *side chain* or *R-group*, composed by chemically different residues. A schematic representation of an amino acid is given in Figure 2. The carbon atom of the carboxyl group is usually called C'. Consecutive amino acids are connected by a *peptide bond*, i.e. the carboxyl group of the i-th amino acid of the sequence is linked, through a covalent bond, to the amide group of the $(i+1)$-th amino acid and a H_2O molecule is released, as shown in Figure 2. Therefore, the whole structure of the protein consists of a "main chain" of atoms, made of the linked $NC_\alpha C'O$ components of amino acids, and a number of side chains, with a shape similar to a fishbone. For this similarity, the main chain is also called *backbone*. The sequence of amino acids specific of each different protein is called *primary structure*.

As previously observed, the information contained into the chain of amino acid residues determines the unique three-dimensional conformation of a protein, i.e. its own *native state* or *tertiary structure*. Folded proteins usually contain one or more local, repetitive spatial arrangements of amino acid residues, with characteristic conformations, called *secondary structures*. The most common secondary structures found in proteins are α-helices, β-sheets and loop/turns. Examples of α-helices and β-sheets are given in Figure 2.

Protein tertiary structures can be described in terms of bond lengths (i.e. distances between two atoms connected with a covalent bond), bond angles (i.e. angles between two adiacent bond vectors, where a bond vector is identified by two atoms connected with a covalent bond) and dihedral angles

(i.e. angles between the normals to the planes defined by suitable consecutive triplets of atoms). When the protein is at its equilibrium state, the bond lengths and bond angles can be considered approximately fixed, so that the three-dimensional conformation is determined by the dihedral angles. These angles are conventionally denoted with the letters Φ, Ψ, ω and χ. The former three angles characterize the protein backbone, while the latter is related to the side chains. A representation of Φ, Ψ and ω is given in Figure 2, where the indices $i-1$, i and $i+1$ identify three consecutive amino acid residues. For more details the reader is referred, for example, to [53].

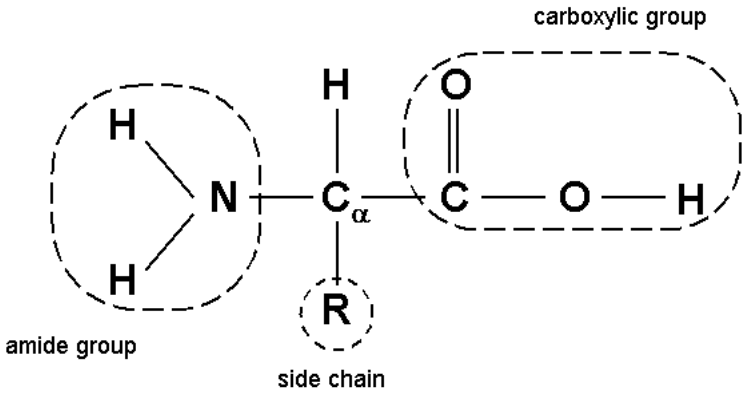

Fig. 1. Schematic representation of an amino acid.

3 Computational Approaches to Protein Fold Prediction

Computational approaches to predict protein three-dimensional conformations are usually classified as *homology modeling* (or *comparative modeling*), *fold recognition* (or *threading*) and *folding ab initio* (see, for example, [18]).

Homology modeling is based on the idea that proteins having strong sequence similarity have also strong structure and function similarity. Given a sequence of amino acid residues, homology modeling methods essentially try to align the target sequence to suitable structure templates, stored in protein databases, and build a three-dimensional conformation by using alignment information (see, for example, [14, 17, 79]). Different alignment methods have been developed, such as BLAST [3], PSI-BLAST [4] and the profile-profile method [41]. The main limitation of the homology modeling methods is that they work effectively only for sequences with at least 30-40% identity. For smaller identity percentages, they have a low reliability (see, for example,

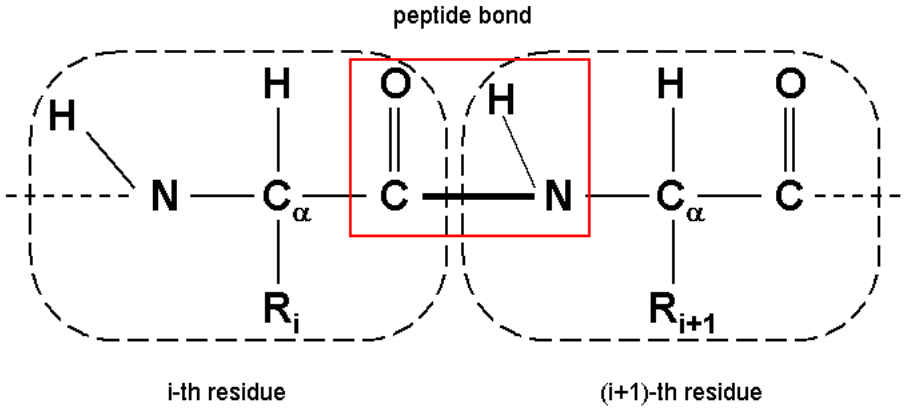

Fig. 2. A peptide bond between two amino acid residues.

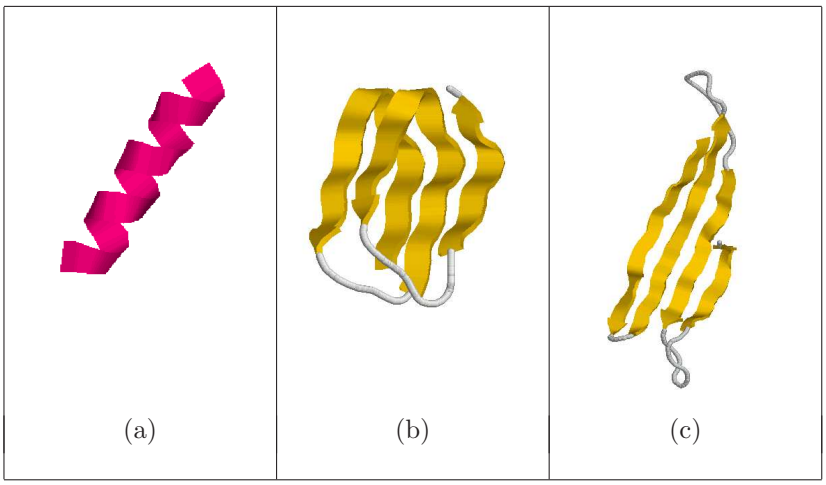

Fig. 3. Examples of protein secondary structures: α-helix (a), parallel β-sheet (b), anti-parallel β-sheet (c).

[22]). A further limitation is that only 15-25% of sequences have homologous proteins with known three-dimensional conformation in a given genome.

Fold recognition methods are based on the idea that there may be only a limited number of different protein folds. Therefore, they try to predict the protein conformation from known three-dimensional structures that do not have homologous characteristics. To this aim, a library of structure templates is defined, then the target sequence is fitted to each library entry and an

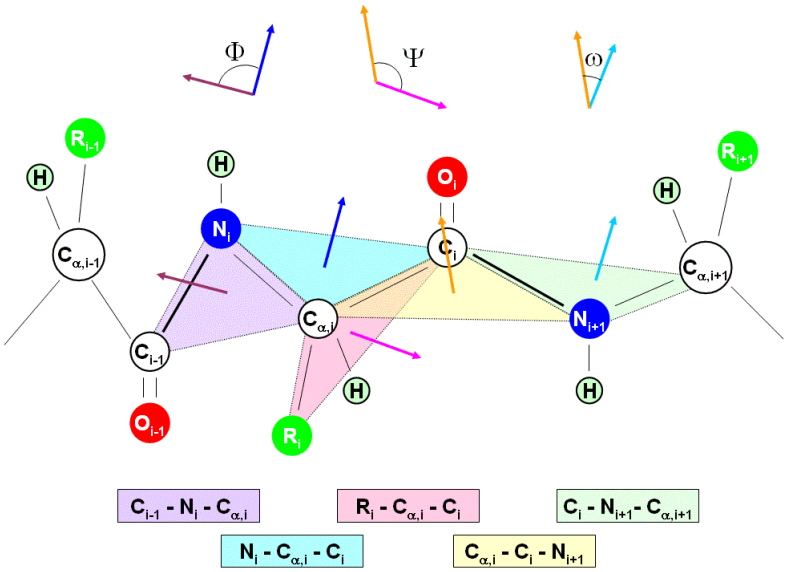

Fig. 4. The protein dihedral angles Φ, Ψ and ω.

energy function is used to evaluate the fit and hence to determine the most suitable template. Obviously, the quality of the obtained model is limited by the actual presence of the correct template into the database and by the actual similarity of the selected templates.

Ab-initio methods are potentially able to predict three-dimensional conformations not yet been observed experimentally. The basic idea behind these methods is that, according to thermodynamic principles, a protein spontaneously folds into its native state, which corresponds to a global minimum of free energy. As already observed, ab-initio techniques can be divided into two categories, one including the methods that use knowledge-based information, such as secondary structure information stored in databases, the other including the methods that do not exploit structural databases during folding predictions.

As discussed in [27, 61], ab-initio methods are generally characterized by suitable protein representations, by energy functions that take into account physicochemical interactions, and by efficient algorithms to search the feasible conformational space. Computational models of proteins explicitly treating all degrees of freedom are currently impractical because of the huge size of the conformational space, of the large number of intramolecular/intermolecular interactions and of the protein complex topology. Both high-resolution and low-resolution models introduce simplifications. High-resolution models taking into account detailed information about the protein conformation are more

rigorous, but lead to problems that are more difficult to be solved. On the other hand, low-resolution models, based on simplified molecular descriptions or structural restraints, can provide only simplified fold descriptions, but are able to give insights into thermodynamic and kinetic properties of the protein folding process.

4 Energy Functions

Energy modeling plays a critical role in protein folding simulations. A large number of energy functions, also called force fields, has therefore been developed to represent the interactions among protein atoms. To better understand the ab-initio methods presented in Sections 6 and 7 we give a short description of energy functions. This description follows [21]; for more details the reader is referred there and to references therein.

Over the years, a large number of energy models has been empirically developed for the protein folding problem, such as AMBER [93], CHARMM [15], ECEPP [57, 58], ECEPP/2 [59], ECEPP/3 [60], MM2 [1] and MM3 [2]. These models are typically expressed as the sum of potential energy terms representing *bonded interactions*, i.e. related to bonds, bond angles and dihedral angles, and *nonbonded interactions*, such as van der Waals and electrostatic ones. These potentials are usually described in terms of relative distances of atoms or atom aggregates.

A simple model of bond potential energy is

$$E^{bond} = k^{bond}(r - r_0)^2,$$

which measures how much the bond length r is far from its ideal value r_0. The constant $k^{bond} > 0$ is called "spring constant", in analogy with Hooke's law. This model provides a good approximation of the bond potential just on small motions around the equilibrium configuration. A more detailed representation of bond stretching is obtained by considering the so-called Morse potential:

$$\tilde{E}^{bond} = \tilde{k}^{bond}(1 - e^{a(r-r_0)})^2,$$

with $\tilde{k}^{bond}, a > 0$. However, the first potential is usually considered because it is simpler to evaluate than Morse potential. Small protein structures obtained by X-ray crystallography are typically used to compute r_0.

Angle bending energy is associated with vibrations around the equilibrium bond angle θ_0, therefore its potential can be modeled by Hooke's law too:

$$E^{angle} = k^{angle}(\theta - \theta_0)^2.$$

The value of θ_0 depends on the triplet of atoms defining the bond angle θ and k^{angle} and controls the angle stiffness.

Tortional energy potentials are used to describe the internal rotation energy of dihedral angles. These potentials are usually modeled as

$$E^{dihedral} = \sum_{n=1}^{3} \frac{V_n}{2}[1 + cos(n\psi - \gamma)]$$

where the V_i's are rotation energy barriers, ψ is the torsion angle and γ is the angular offset. Note that some force fields neglect bond stretching and angle bending energies, thus taking into account only torsonal energy.

Nonbonded interactions involve atoms that are not linked by covalent bonds. Usually, non bonded energy terms account for the electrostatic energy and the van der Waals energy.

On each peptide bond between two amino acid residues there is a dipole which is orthogonal to the $N-C$ bond. The energy of this dipole is described by the Coulomb law:

$$E_{ij}^{elect} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

where q_i and q_j are the magnitudes of the two charges of the dipole, r_{ij} is the distance between the charges and ϵ_0 is the dielectric constant.

The main energy involved in the protein stabilization is the non-bonded van der Waals energy, arising from a balance between attractive and repulsive subatomic forces. Attractive forces are longer range than repulsive forces, but, if the distance among atoms is short, they become dominant. This leads to an equilibrium distance in which repulsive and attractive forces are balanced. The van der Waals interaction between two atoms i and j is often modeled through a Lennard-Jones potential, which includes attraction and repulsion terms:

$$E_{ij}^{vdW} = \frac{a_{ij}}{(r_{ij})^{12}} - \frac{b_{ij}}{(r_{ij})^{6}}.$$

The constants a_{ij} and b_{ij} control the depth and the position of the potential energy well.

The solvent, usually water, has a fundamental influence on the structure, dynamics and thermodynamics of biological molecules, both locally and globally. One of the most important solvent effects is the screening of electrostatic interactions. This can be taken into account implicitly, by including a further dielectric constant ϵ_r in the electrostatic energy potential:

$$E^{elect+solv} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}$$

A more rigorous treatment of solvent effects can be obtained by considering the Poisson-Boltzmann equations. As an alternative, the solvent is explicitly taken into account, by using models based on the assumption that solvation energy is proportional to the protein surface area exposed to the solvent, or to the solvent accessible volume of a hydration layer. These models account also for cavity formations [39].

From a thermodynamic point of view, the difference between two molecular conformations is determined by their difference in free energy, that is defined

in terms of enthalpy, H, entropy, S, and absolute temperature, T, of the molecular system:

$$E^{free} = H - T \cdot S.$$

A direct computation of the free energy requires detailed molecular dynamics simulations and hence is too costly. A generally accepted alternative approach based on statistical mechanics describes the free energy contributions by using harmonic approximations [31].

5 Optimization Solvers

As previously observed, ab-initio methods search for conformations corresponding to the global minimum of some energy function, under suitable constraints, i.e. lead to constrained global optimization problems. Hence, it is useful to give a brief description of optimization solvers applied in this context. We focus here on the solvers that are used in the protein fold prediction methods described in the next sections. For more details the reader is referred to [21, 33, 62, 63, 64].

The global optimization solvers can be divided into two main classes: *heuristic* and *deterministic*. The former includes methods based on probabilistic descriptions, for which convergence to a solution is not ensured, or only a convergence in probability is demonstrated. The latter contains methods that, under suitable hypotheses, provide convergence to a solution of the global optimization problem.

Monte Carlo (MC) methods [19] are heuristic methods that simulate the evolution of a system in terms of probability distribution functions. They generate many approximate solutions by random sampling from a probability distribution and get the target solution as an average over the generated samples. In many applications, the variance corresponding to the average solution can be predicted, obtaining an estimate of the number of samples needed to achieve a given error. Enhancements of the basic MC strategy have been developed to reduce the possibility of getting trapped into local minima. They include *Replica Exchange Monte Carlo* (REM) [91], *Parallel Hyperbolic Sampling* [96] and *Electrostatically-Driven Monte Carlo* (EDMC) [72] methods.

A further improvement over MC methods is provided by *Simulated Annealing* (SA) methods [40, 52]. They are based on an analogy with the annealing physical process that consists in decreasing slowly the temperature of a given system (e.g. a liquid metal) in order to obtain a crystalline structure. SA methods are iterative procedures that, at each step, execute a *Metropolis Monte Carlo* algorithm that generates a new candidate approximation of the solution, by applying a random perturbation to the previous one. Through a random mechanism controlled by a parameter called temperature, it is decided whether to move to the candidate approximation or to stay in the current one at the next iteration. The acceptance/rejection of the new approximation is

usually based on the evaluation of the so-called Metropolis acceptance function, that is a probability function based on the Boltzmann distribution [55]. Higher temperatures correspond to a larger number of accepted conformations. The temperature parameter plays a crucial role in the whole process; it must be decreased very slowly, to avoid the simulation gets trapped in a local minimum close to the initial state. A modification of the SA strategy called *Monte Carlo Minimization* (MCM) has been also developed, that applies a local Monte Carlo minimization to the current conformation, before checking if the Metropolis acceptance criterion is satisfied [73, 74]. We come back to Metropolis Monte Carlo Simulated Annealing in Section 10.1, since we used this method in our experiments.

Genetic Algorithms (GAs) are heuristic methods based on principles from the evolution theory. Indeed, they represent each feasible point in the conformational space as a chromosome and mimic the evolution of a population of chromosomes. Two chromosomes can generate child chromosomes (crossover operation) and a chromosome can undergo mutations. Furthermore, chromosomes are selected depending on their fitness value, which is defined taking into account the objective function to be minimized. Starting from an initial population, GAs set up an iterative process, where a child population is generated at each step from a parent one, by applying the above evolutionary mechanisms, until suitable termination criteria are satisfied. GAs differ by the mechanisms used to simulate mutation and crossover and by the fitness function. As noted in [21], the choice of these mechanisms greatly influences the ability of finding global minimum energy configurations. A review on GAs is given in [87].

Conformational Space Annealing (CSA) methods work with typical concepts of SA, GAs and MCM. As in GAs, an initial population of variables called first bank is generated and then a subset of bank conformations called seeds are selected. The seeds are perturbed, by replacing (typically small) seed portions with the corresponding portions of bank conformations, and are used as trial conformations, to obtain a new bank. As in MCM, a local minimization is applied to all conformations to work only with the space of local minima. The diversity of sampling is controlled by comparing a suitable distance measure between two conformations with a cutoff value, D_{cut}. A trial conformation is compared with the closest one in the current bank. If their distance is smaller than D_{cut}, they are considered similar and the one with lower energy is chosen. Otherwise, the highest energy conformation in the bank plus the trial one is discarded. The cutoff value is slowly decreased during the simulation process and hence acts as the temperature parameter in SA. The algorithm usually stops when all the bank conformations have been used as seeds and the cutoff parameter has reached a suitably small value. More details can be found in [32, 47].

An example of deterministic global optimization strategy is provided by *Molecular Dynamics* (MD) simulations. MD methods simulate the evolution of a molecular system by applying the equation of motion to the atoms of

the system. They have been able to provide detailed information about heteropolymers and to give insights into complex dynamic processes occurring in biological systems, such as protein folding [16].

Branch and Bound (BB) methods fall into the class of deterministic global optimization methods too. These are iterative methods that, at each step, find lower and upper bounds on the global minimum objective value. The iterations are stopped when the difference between the bounds is smaller than a given tolerance. Recently, a deterministic BB algorithm named αBB has been developed by Floudas et al. and applied to molecular conformation problems [6, 35, 36]. αBB determines the upper bounds by function evaluation or local minimization of the original objective function, while the lower bounds are computed by minimizing convex lower-bounding functions obtained by adding a convex term to the original one. The lower-bounding functions depend on a parameter that controls their shape and must be properly chosen to guarantee convexity. Lower-bounding functions are built in such a way that they have properties ensuring the convergence of the algorithm to a global minimum.

6 Ab-initio Methods Using Knowledge-Based Information

Ab-initio methods with knowledge-based information usually build template models by extracting from databases fragments with sequence or structural similarity to fragments of the target sequence. Therefore, there is no clearly defined separation between these methods and the homology modeling or fold recognition ones. Ab-initio methods exploiting both approches are discussed in the next two Sections.

6.1 Lattice models

To reduce the degrees of freedom of the conformational space, models have been developed that are based on a simplified representation of the protein chain over a lattice. These *lattice models* use secondary structure predictions and threading techniques to derive some constraints; then, they search the conformational space by applying Monte Carlo procedures to the lattice. Because of these semplifications, lattice models are generally two orders of magnitude faster than high-resolution models [45]. On the other hand, simplified models of proteins lead to a loss of dynamic mechanisms, so that often predicted conformations do not fit native structures suitably. First lattice studies did not focus on protein structure prediction, but rather on understanding thermodynamic and kinetic properties of protein folding. Indeed lattice models have a long history in modeling polymers, due to their analytical and computational simplicity.

Early in the '90s, Levitt et al. [49] developed a low-resolution method, based on a simple representation of the protein backbone as a self-avoiding

chain of connected vertices on a tetrahedral lattice, with several amino acid residues assigned to each lattice vertex. To reduce the space of feasible lattice structures, this model requires the final conformations to be compact and globular. Effects of solvent interactions are not considered, because the lattice model did not represent accurately the exposed surface of a conformation. Starting from observed contact frequencies in X-ray structures, the energy of contact between two lattice vertices is defined and a dynamic programming strategy is applied to find the best conformational energy. This model was validated on real proteins with 52-68 amino acid residues and correct low-resolution structures were found [49]. A drawback is that it can be applied only to proteins with a small number of residues; furthermore, it does not consider interatomic interactions.

Lattice models have undergone an evolution over the years. In [77, 94] Levitt and co-workers presented a lattice-based hierarchical approach. In this case, starting from the sequence of amino acid residues, all feasible compact conformations are identified by using a highly simplified tetrahedral lattice model; a lattice-based scoring function is used to select a subset of these conformations and to build high-resolution (all-atom) models. Then, by using a knowledge-based scoring function, three small subsets are extracted from the set of all-atom models and a procedure based on distance geometry is used to generate the best conformations from each of the subsets. Using this approach, structures of proteins with at most 80 residues were predicted, obtaining RMSD values ranging from 4.1 to 7.4 $\mathring{\text{A}}$ [77]. Unfortunately, the method failed for proteins with complex supersecondary structures.

Lattice models have been also studied by Skolnick and co-workers [43, 44]. They developed a lattice model of the protein structure and dynamics in which the polypeptide chain is represented with a simple cubic lattice. The emphasis is on the side chain role, rather than on geometry of the backbone. The backbone is treated implicitly, since the C_α coordinates are computed by considering the positions of three consecutive side chains. The energy function takes into account sequence independent properties, such as interactions between the i-th and the $(i+4)$-th residues in the α-helix side chains or long distance interactions in the β-sheets, and sequence dependent properties, such as long-range pairwise and multibody interactions that simulate hydrofobic effects. The lowest energy conformation corresponding to the native state is searched by a Replica Exchange Monte Carlo (REM) procedure [91]. The model was tested on small and structurally simple single-domain proteins considering two sets of sequences, one corresponding to single fragments of known structures, the other to known protein tertiary structures. The best results, evaluated by using the RMSD values of the predicted versus the original conformations, were obtained for the set of single fragments. The method evolved into a hierarchical ab-initio lattice approach that uses a combination of multiple sequence comparison, threading, clustering and refinement [83]. In this approach, the starting fragmentary templates for the lattice model

are provided by a threading algorithm and a reduced representation of the protein conformational space is used, where the center of mass of the C_α and side-chain atoms are the interaction centers. The energy function is defined through a statistical analysis of known protein structures, leading to statistical potentials for pairwise and multiboby side-chain interactions. The conformational space is sampled by the REM procedure. This method is called SICHO (SIde CHain Only). Results presented at CASP4 meeting [70] showed that it is able to obtain good results on small proteins of not too complex topology [83].

Another structure prediction lattice-method that combines homology and ab-initio modeling is TOUCHSTONE, developed by Skolnick et al. [84]. A first version of this method is based on the SICHO lattice model, with force field including short-range structural correlations, hydrogen-bonding interactions and long-range pair-wise potential. Two threading restraints are used to reduce the conformational search space, concerning side-chain contacts and local distances. The former restraint is obtained by using the PROSPECTOR threading algorithm [71], while the latter is derived from sequence alignments and threading of short sequence fragments. REM is used to search the conformational space. To generate another set of indipendent trajectories, a Monte Carlo sampling scheme, called Parallel Hyperolic Sampling (PHS) [96], is used. Then the structures generated by the simulations are rebuilt at an atomic detail. This method was applied to the genome of *Mycoplasma genitalium* bacterium, that has one of the smallest known genomes among living organisms [85]. 85 proteins with at most 150 amino acid residues were examined, obtaining a correct prediction of the topology of 63% of the proteins.

As discussed in [85], the potential function used in TOUCHSTONE is not suitable for predicting multiple-domain structures. To overcome this limitation, both the lattice representation and the force field have been modified [86, 97]. The SICHO model has been replaced by the CABS one, in which the C_α trace is confined to a lattice system, while the group made by the side chain and the C_β carbon are off-lattice, with positions determined from three adjacent C_α atoms. The energy function takes into account pairwise and multiboby side-chain interactions, short- and long-range hydrogen-bond interactions, contact and local distance restraints obtained through PROSPECTOR, burial and electrostatic interactions, global propensities to predicted contact orders and contact numbers, and local stiffness of global proteins. The conformational space search method is PHS, as in the previous TOUCHSTONE version.

Experiments were carried out on a set of 125 proteins (43 all-α proteins, 41 all-β proteins and 51 α/β-proteins, according to Kabsch and Saunder classification [30]), with lengths ranging from 36 to 174 amino acid residues. By using PROSPECTOR restraints, 83 proteins were successfully folded. Comparisons with the previous TOUCHSTONE version showed the efficiency of CABS versus SICHO. Furthermore, it was observed that short-range restraints considerably speedup local structure formations.

Recently, a high-resolution lattice model has been developed by Kolinski [45] that is based on a representation of the protein backbone over a lattice and on the REM searching procedure. For each residue, this model takes into account the C_α and C_β carbons, the side-chain and an additional atom located along the $C_\alpha - C_\alpha$ virtual bound. Only the C_α coordinates are explicitly computed and are used, together with amino acid properties, to calculate the coordinates of off-lattice elements. The force field is based on the CABS model and the potential used takes into account short- and long-range interactions. The simulation process is based on Metropolis Monte Carlo scheme, subject to a simulated annealing procedure or controlled by REM. This lattice model can be applied to perform ab-initio structure predictions as well as in multi-template comparative modeling [45].

6.2 Methods Based on Fragment Assembly

The idea behind these methods is to build protein tertiary structures from small protein segments or secondary structures, obtained through sequence alignment or threading.

Such an approach is implemented, for example, in FRAGFOLD, developed by Jones et al. [26, 28]. In FRAGFOLD simulations, the first step is the selection from a library of protein structures of suitable supersecondary structural fragments at the position of each residue of the target sequence, and hence the prediction of secondary structures by using PSIPRED [25], which applies neural-network techniques and PSI-BLAST sequence alignments. The predicted secondary structures are used as input to FRAGFOLD. Random conformations are then generated until a conformation with no steric clashes is obtained. Starting from this one, a Simulated Annealing algorithm is applied to minimize an energy function, which is a weighted sum of terms expressing short- and long-distance pair potentials, single-residue solvation energy, steric interactions (such as the van der Waals energy), and hydrogen-bond interactions. Results presented at CASP4 and CASP5 [26, 28] showed that FRAGFOLD can correctly predict local domains, but fails in predicting entire three-dimensional structures. In particular, there are problems with the prediction of β-structures, since the formation of these structures is a cooperative process requiring the convergence of many substructures.

Another method which exploits sequence alignment and fragment assembly is Rosetta, developed by Baker et al. [12, 13, 81, 82]. This method is based on the assumption that the distribution of conformations of each three- and nine-residue segment can be reasonably approximated by the distribution of structures adopted by the corresponding sequence (or closely related ones) in known protein conformations. Therefore, Rosetta breaks the target sequence into three- and nine-residue segments and applies a profile-profile comparison procedure to extract fragment libraries from protein structure databases. The fragments are assembled to build three-dimensional structures by using a

fragment insertion Metropolis Monte Carlo procedure. Many of such template-based models are generated and then clustered. For sequences with less than 100 residues, an all-atom refinement is used instead of clustering. The energy function used in searching the conformational space describes sequence-dependent properties, such as non-local interactions (e.g. disulfide bonding, backbone hydrogen bonding, electrostatics) and sequence-independent properties, connected to the formation of α-helices, β-strands and to the assembly of β-strands into β-sheets. Only the backbone atoms are considered explicitly, while the side chains are represented as centroids.

Rosetta underwent a significant evolution since its development. The improvements concern the application of filters to reject non-protein-like conformations (local low-order contact conformations and β-strands not properly assembled into β-sheets) [76], the modifications of the methodology for picking up fragments from the structure database, in order to ensure a remarkable diversity of secondary structures when dealing with segments with a weak propensity to fold into a single secondary structure, the use of a new prediction method, JUFO [29], and the exploitation of quantum chemistry calculations, traditional molecular mechanics approaches and protein structural analysis to compute parameters in the energy function [12, 13]. A neural network method is under development with the aim of identifying strand-loop-strand motifs starting from the protein primary structure [46].

Rosetta was applied to CASP5 targets. In particular, for α- or α/β-proteins Rosetta generated models with a correct overall topology and RMSD values ranging from 2.8 to 4.2 $\mathring{\text{A}}$. Rosetta method failed for proteins having more than 280 residues and a complex topology; furthermore, it sometimes generated models being too globular or having β-strands less exposed than in the native conformation.

7 Ab-initio Methods Without Knowledge-Based Information

Knowledge-based ab-initio methods are dependent on the information stored in structural databases and on statistical analysis of this information; hence they can produce inaccurate predictions of new folds. A way to overcome this problem is offered by "true" ab-initio methods which simulate the folding process by using only protein models based on physicochemical principles. These methods are obviously more challenging, since they require "realistic" representations of atomic interactions and powerful algorithms and computational resources to search the feasible conformational space. A few examples of ab-initio methods without database information are discussed in the next sections.

7.1 Hierarchical Approaches

Hierarchical approaches start from a reduced representation of protein atoms and their interactions and then refine computed reduced conformations to obtain all-atom structures to be optimized.

A simple hierarchical approach to protein folding is given by LINUS (Local Independently Nucleated Units of Structure), developed by Srinivasan and Rose [89, 90]. This procedure has been used to predict secondary structures and to capture a physical interpretation of protein secondary elements. Indeed, Srinivasan and Rose used LINUS to support the physical theory that secondary structure propensities are mainly determined by competing local effects, involving conformational entropy and hydrogen bonding.

A Metropolis Monte Carlo procedure is applied to search the conformational space. The amino acid sequence is considered as an extended chain, where the backbone atoms are represented as points, while the side chains are modeled as different nonoverlapping spheres, according to amino acid type and size. The degrees of freedom are the dihedral angles, Φ, Ψ and χ. A Metropolis Monte Carlo procedure is used to search the conformational space. More precisely, the extended chain is subdivided into subsequences of three consecutive residues, proceeding from the N-terminus to the C-terminus, that are perturbed by using a predefined set of random moves to obtain a new configuration. This configuration is accepted or rejected according to a Metropolis acceptance criterion based on attractive and repulsive contributions [90]. This cycle is completed when all the chain residues have been processed.

LINUS was used also by Maritan and co-workers in order to estimate the rate of successful secondary structure predictions as a function of the temperature [24]. In particular, they showed that at low temperatures local interactions are facilitated and stabilized, leading to α-helices and turns; consequently, β-strands are favoured at high temperatures. At intermediate temperatures some protein subsequences tend to fold into β-strands, while others into α-helices and turns. They also found that α-helices and β-strands can be predicted with an accuracy greater than 40% [24].

A different hierarchical approach has been developed by Scheraga and coworkers [78] to capture pairwise and multibody interactions during the folding process. In this approach, a set of low-energy structures is computed first, by using a reduced model based only on the C_α trace and on the so-called UNRES (UNited-RESidue) potential force field [50, 51], to describe intra-protein interactions and hydrogen bonding. The conformational space is searched by a Conformational Space Annealing (CSA) algorithm [47]. The virtual-bond chains of these low-energy structures are converted to an all-atom backbone, by using the dipole-path method based on alignment of peptide-group dipoles [50]. The backbone conformation is optimized by using EDMC [72], a procedure that iteratively looks for low-energy structures in the conformational space and takes into account electrostatic interactions and thermal effects. All-atom side chains are added to the previous model under constraints of

non-overlap; loop and disulfide-bonds are then treated explicitly. The final conformation is obtained by using the ECEPP/3 all-atom energy function [60], with gradual reduction of the C_α-C_α distance of the parent united-residue structure. The ECEPP/3 energy function is the sum of electrostatic, hydrogen-bonded, torsional and non-bonded terms.

The method above described was successfully applied to single-chain proteins as well as to multiple-chain ones. In the latter case, in order to obtain correct predictions, interchain interactions were taken into account by suitably modyfing UNRES and CSA [78].

7.2 A Combinatorial and Global Optimization Approach

A novel true ab-initio approach for the prediction of three-dimensional structures of proteins is implemented in ASTRO-FOLD, developed by Floudas and co-workers [33, 35, 36, 38, 39]. ASTRO-FOLD combines the classical hierarchical view of protein folding, in which the folding process starts from rapid formation of secondary structures and then proceeds to the slower tertiary structure arrangement, with the hydrophobic-collapse view, in which secondary and tertiary structures are formed concurrently. The prediction of a protein conformation is performed into four steps. First, initiation and termination sites of α-helices are identified, then β-strands are identified and β-sheet topologies are predicted, and, later, constraints on the protein structure and information on loop segments are derived. Based on the previous information, the overall protein tertiary structure is predicted by using a model that combines both the above views of the protein folding process and by applying deterministic global optimization, stochastic optimization and torsion-angle dynamics. Therefore, ASTRO-FOLD can be defined as combinatorial and global optimization framework based on a four-step approach.

The main idea behind α-helix determination is that the fold of such secondary structure is based on local interactions. Hence, in order to identify local sites of helix formation, the amino acid sequence is segmented into overlapping oligopeptides and ensembles of low potential states are computed, along with a global minimum energy state, using a detailed atomic level model based on the ECEPP/3 force field [60]. The determination of these state is performed by applying the deterministic branch-and-bound algorithm α-BB [20] and the stochastic CSA algorithm [32, 47]. Free energy calculations are then performed, with a force field which is the sum of potential, entropic, solvation, ionization and internal cavity contributions. The energy values are used to compute the probability that each oligopeptide folds into a helix and to define a helix propensity for each residue.

Once α-helices have been identified, the remaining residues are analyzed to identify the locations of β-strands and β-sheets, and to predict β-sheet topologies as well as disulfide bridges. Since the formation of such structures is driven by long-distance interactions, a different approach is used. The key assumption is that β-structure formation depends on hydrophobic forces [37];

to model them, the prediction of hydrofobic residue contacts is required. To predict a β-sheet, β-strand superstructures are postulated that encompass all the β-strand substructures that may constitute the β-sheet topology. The mathematical model of the superstructures is formulated as a global optimization problem, whose solution maximizes contacts between hydrophobic residues, subject to constraints enforcing physically meaningful configurations for β-strands and disulfide bridges. This approach is used to identify a rank-ordered list of possible β-sheet structures.

Once α-helices and β-sheets have been identified, secondary structure restraints are defined. Dihedral angles, atomic distance and $C_\alpha - C_\alpha$ distance bounds are defined according to the main properties of corresponding secondary structures. Restraints for unassigned residues are also defined either through an analysis of overlapping oligopeptides, such as for α-helices identification, or through predictions of entire loop fragments. Both approaches are implemented by exploiting deterministic and stochastic optimization solvers.

The final stage of ASTRO-FOLD is the prediction of the protein tertiary structure. This problem is formulated as the global minimization of a suitable potential energy, subject to the restraints above discussed. This problem is solved by a combination of α-BB and torsion angle dynamics [35].

As reported in [36], ASTRO-FOLD was tested on CASP5 targets of at least 150 residues, obtaining accurate α-helix and β-strand and impressive β-sheet predictions. Indeed, RMSD values ranging between 4.1 $\mathring{\text{A}}$ and 6.9 $\mathring{\text{A}}$, and SOV [95] values corresponding to more than 80% accuracy have been obtained for the computed conformations.

As noted in [36], the application of ASTRO-FOLD to medium-size proteins was made possible by using distributed computing environments. The framework was parallelized by taking into account the different problems and solvers at each stage the prediction process.

7.3 Topological Approaches

Experimental and theoretical studies have shown that the folding process is widely influenced by topological properties of the native state. For example, by analyzing a small set of non homologous simple single domain proteins, Baker and co-workers revealed that a statistically significant correlation exists between folding kinetics and native state topological complexity [69]. Starting from their results, Koga and Takada studied the relationships between native topology and folding pathways [42]. By using a simple representation of the polypeptide chain through its C_α trace and a free-energy functional approach, that takes into account chain connectivity, contact interactions and entropy, they were able to correctly describe folding pathways of small single-domain proteins. The correlation between the topology of the native state and the folding pathways was confirmed by Maritan et al. [80], by performing molecular dynamics simulations of the immunoglobulin, using a model that represents only the C_α carbons and an energy function that includes bonding

and non-bonding terms. Other studies suggest that folding rates are correlated to topological parameters such as contact order and cliquishness [56].

An interesting topological approach to the protein folding problem has been proposed by a research group led by Banavar and Maritan [7, 8, 9, 10, 54]. In this approach, a protein is modeled as a tube of nonzero thickness without any self-contacts (see Figure 7.3). The axis of the tube is a suitable curve interpolating the C_α carbons and the thickness is expressed in terms of a metric that measures the "distance" among any three points on the curve, x_i, x_j, x_k, as the radius $r(x_i, x_j, x_k)$ of the circle passing through them (r is assumed to be infinity if the points are aligned). Note that $1/(r(x_i, x_j, x_k)^p$, $p > 0$, can be regarded as a three-body potential and hence the tube thickness is related to a certain interaction energy among chain particles [23]. Indeed, the modeled structure is energetically stable, i.e. its conformation corresponds to a minimum of free energy, when it achieves a maximum thickness under constrains preventing self-intersection and aligned triplets of amino acids. As pointed out in [7], despite its simplicity, this model is able to capture the physical thickness of the protein chain, that is due to the presence of the R-groups. Furthermore, a nonzero thickness implies that the interactions between two spatially close tube segments do not depend only on their distance, but also on their relative orientation, so the tube model is able to represent the inherent anisotropy associated with the local directionality of the chain.

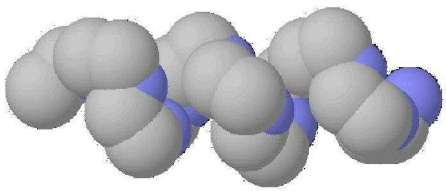

Fig. 5. The sequence of N-C_α-C' units of the crambin helix composed of the amino acids $7 \div 17$. The picture is similar to a thickened tube.

Numerical simulations based on the above model are reported in [7, 10, 92]. Different constraints have been considered to take into account the compactness of a polymer chain, such as a pairwise attractive potential with a given range [7], or suitable bounds on the global and the local gyration radius or on the contact distance and the number of allowed contacts [92]. A Metropolis Monte Carlo procedure has been used to search the conformational space, obtaining helix- and hairpin-like structures.

We have focused our attention on the tube model, because it appears both simple and capable of representing significant features of the protein chain.

Next sections are devoted to describe a modified version of it and related computational experiments.

8 A Modification of the Tube Model

Following [10, 23], we provide a more detailed description of the tube model, which is the basis of our computational approach. Let $X = (x_1, x_2, \ldots, x_n)$ be a n-ple of different points called *conformation*, where each $x_i \in \Re^3$ represents the position of the C_α atom of the i-th amino acid residue of the polypeptide chain. The interaction among any three non-aligned points x_i, x_j, x_k can be measured by the radius of the unique circle among them, which has the following expression:

$$r(x_i, x_j, x_k) = \frac{||x_i - x_j||||x_i - x_k||||x_j - x_k||}{4A(x_i, x_j, x_k)} = \frac{||x_i - x_j||}{2|\sin\theta|}$$

where $||\cdot||$ is the Euclidean norm, $A(x_i, x_j, x_k)$ is the area of the triangle with vertices x_i, x_j and x_k, and θ is the angle between the vectors $x_i - x_k$ and $x_j - x_k$. If the three points are aligned, $A(x_i, x_j, x_k)$ and $\sin\theta$ are null, hence the above definition can be extended to these points by setting $r(x_i, x_j, x_k) = \infty$. Note that $r(x_i, x_j, x_k)$ can be viewed as an approximation of the standard radius of curvature. Indeed, if the three points vary over a simple (i.e. without knots) and smooth curve C, then

$$\lim_{\substack{x_j, x_k \to x_i \\ x_j, x_k \in C}} r(x_i, x_j, x_k) = \rho(x_i),$$

where $\rho(x_i)$ is the radius of curvature of C at x_i. In the following, the radius $r(x_i, x_j, x_k)$ is referred to as three-body radius.

The thickness of the conformation X can be defined as:

$$D(X) = \min_{\substack{1 \le i,j,k \le n \\ i \ne j, j \ne k, k \ne i}} r(x_i, x_j, x_k). \tag{1}$$

$D(X)$ is a "discrete version" of the thickness $\Delta(C)$ of a simple and smooth curve C, which is defined as the maximum thickness of a tube with axis C and circular section, that does not exhibit any self-contacts. $\Delta(C)$ has the following expression:

$$\Delta(C) = \min\left\{\min_{x \in C} \rho(x), \frac{1}{2} \min_{(x,y) \in \Omega} ||x - y||\right\},$$

where Ω is the set of all pairs of points of C such that $x \ne y$ and the vector $x - y$ is orthogonal to the tangents to C at both x and y. In other words, in the continuous case, the tube thickness is the smallest value between the

minimum radius of curvature of C and half the minimum distance of closest approach over C. It can be proved that

$$\Delta(C) = \min_{x,y,z \in C} r(x, y, z),$$

where the definition of r is extended by continuity to coinciding points [23].

As pointed out in [10], the three-body radius is able to distinguish among local and non-local interactions along the protein chain. When three consecutive particles are considered, a discrete version of the radius of curvature is used to measure their interaction; when the particles are non-consecutive, the distance of approach between two parts of the chain is taken into account (see Figure 8). The thickness takes into account that the protein backbone cannot have self-contacts and that the side chains cannot overlap; furthermore, it provides a global measure of the free space in the protein conformation.

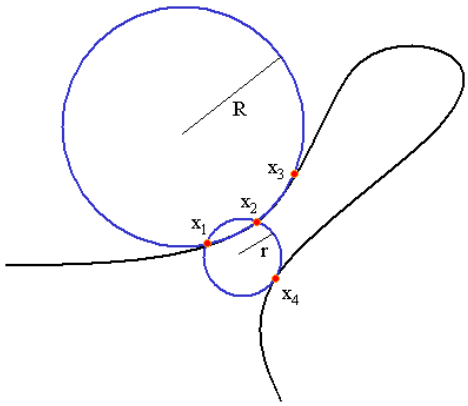

Fig. 6. Three-body radii of consecutive and non-consecutive points.

As observed in Section 7.3, finding an energetically stable conformation can be achieved by maximizing the thickness under suitable constraints. On the other hand, the tube model can be used to predict and analyze compact tube-shaped conformations of given thickness. The latter conformations can be obtained by maximizing a function counting the number of triplets having a three-body radius close to a given thickness value $\bar{D}$:

$$f(X) \equiv f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} f_{\bar{D}}(r(x_i, x_j, x_k)), \tag{2}$$

where

$$f_{\bar{D}}(r(x_i, x_j, x_k)) = \begin{cases} 1 \text{ if } \quad r(x_i, x_j, x_k) \in [\bar{D} - \epsilon, \bar{D} + \epsilon] \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

and ϵ is a real positive constant. As shown in Section 9, typical values of thickness, characterizing protein structures, can be obtained by analyzing existing protein structure data sets; therefore, maximizing $f(X)$ under suitable constraints, using these typical values of thickness, can provide a means to predict meaningful protein-like three dimensional conformations.

To increase global protein compactness, we have modified f by adding a term forcing the points x_i to be inside an ellipsoid, whose surface is thought as a rough approximation of the protein surface shape. By changing the lengths of the ellipsoid axes, different shapes can be approximated. The added term has the following form:

$$g(X) \equiv g(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} g_{(a,b,c)}(x_i) \tag{4}$$

where

$$g_{(a,b,c)}(x_i) = \begin{cases} 1 \text{ if } \dfrac{(x_i^1 - x_G^1)^2}{a^2} + \dfrac{(x_i^2 - x_G^2)^2}{b^2} + \dfrac{(x_i^3 - x_G^3)^2}{c^2} \leq 1 \\ 0 \text{ otherwise} \end{cases}, \tag{5}$$

$x_G = (x_G^1, x_G^2, x_G^3)$ is the barycenter of X, a, b and c are the lengths of the ellipsoid semiaxes, and the superscripts are used to denote the Cartesian coordinates of a point.

Constraints have been imposed to explicitly take into account that two consecutive α-carbons are virtually bonded, hence their Euclidean distance can have only slight variations, and that the Euclidean distance between any two non-consecutive amino acid residues cannot fall below a certain threshold. Furthermore, starting from the observation that in α-helices amino acid residues with positions i and $i + 2$ along the chain are closer than in other structures, a contraint on the Euclidean distance between x_i and x_{i+2} has been imposed to specifically simulate all-α conformations.

The global constrained optimization problem described so far has the following formulation:

$$\max F(X) = \max[f(X) + g(X)] \tag{6}$$

subject to

$$c_1 \leq d(x_i, x_{i+1}) \leq c_2, \qquad \forall i \in \{1, 2, \ldots, n-1\}, \tag{7}$$

$$c_3 \leq d(x_i, x_j), \qquad \forall i, j : \; i > j + 1, \tag{8}$$

$$c_4 \leq d(x_i, x_{i+2}) \leq c_5, \qquad \forall i \in \{1, 2, \ldots, n-2\}. \tag{9}$$

where c_1, c_2, c_3, c_4 and c_5 are real positive constants chosen on the base of experimental observations (see Section 9). The costraints (9) are specifically related to all-α structures.

9 Choice of Model Parameters

The problem (1)-(9) requires the choice of some parameters: the thickness $\bar{D}$ and the related value ϵ in the definition of f (see (2)-(3)), the semiaxis lengths a, b, c in the definition of g (see (4)-(5)) and the constants c_i in the constraints (see (7)-(9)).

The values of $\bar{D}$ and ϵ have been chosen by performing an analysis of a set of 3639 protein structures available in the *PDBSELECT* data collections with *R-factor* < 0.25 and *Resolution* < 2.5 [65]. The thickness of each structure has been evaluated, obtaining the thickness frequency distribution shown in Figure 9.

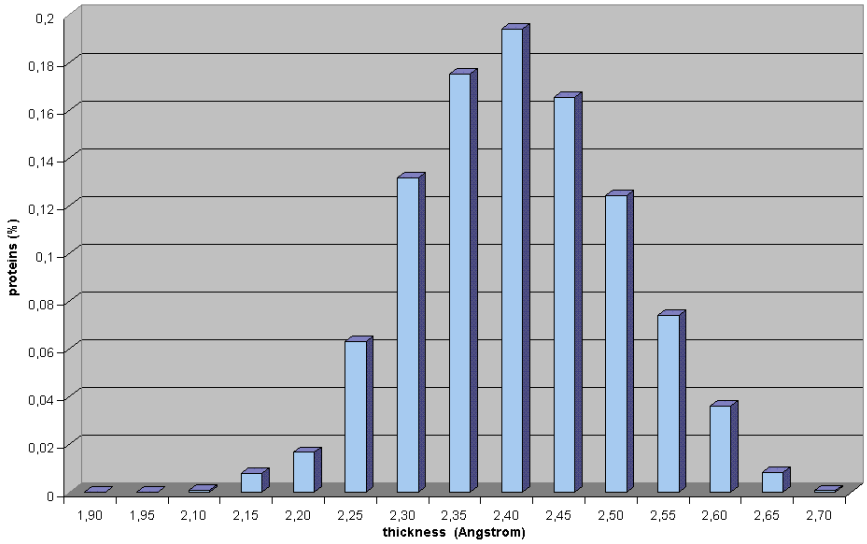

Fig. 7. Frequency distribution of the thickness for a set of 3639 proteins from PDBSELECT.

The thickness mean value is 2.40 Å, with a standard deviation of 0.10 Å; the minimum thickness is 1.91 Å (achieved by only one structure), while the maximum is 2.67 Å. The same analysis has been performed considering all the α-helices (14592 structures) and all the β-sheets (13070 structures) separately. The mean thickness value of the α-helices is 2.65 Å, with a standard deviation of 0.07 Å, a minimum of 2.26 Å and a maximum of 4.58 Å. However, according to the small standard deviation value, more than 98.5% of the α-helices have a thickness ranging between 2.50 Å and 2.90 Å. The frequency

distribution of the thickness of the α-helices in the interval $[2.50, 2.90]$ is reported in Figure 9. The previous results agree with the fact that the α-helices have very similar geometries. The mean value of the β-structures is 2.65 $\mathring{A}$ too, but with a standard deviation of 0.46 $\mathring{A}$, a minimum of 2.12 $\mathring{A}$ and a maximum of 9.75 $\mathring{A}$. Taking into account the low variability of the α-helices thickness, in our experiments we focused our attention on α-structures.

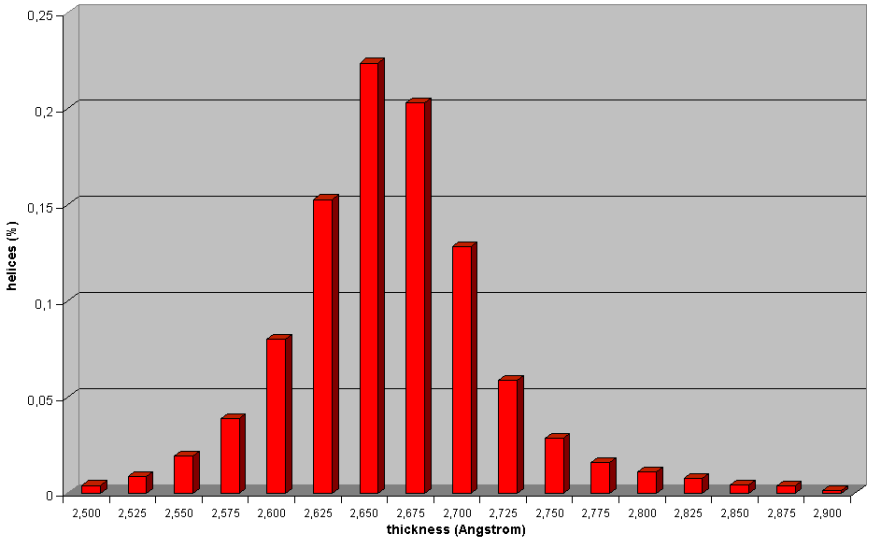

Fig. 8. Frequency distribution of the thickness for a set of 14592 α-helices from PDBSELECT.

A deeper analysis has shown that in the α-helices only few triplets of α-carbons have a three-body radius equal to the thickness. For example, the helix of the crambin (PDB code `1crn`) composed by the amino acid residues $7 \div 17$ has a thickness equal to 2.66 $\mathring{A}$, but just the α-carbons 15, 16 and 17 have this three-body radius, while all the other triplets have a three-body radius of at least 2.71 $\mathring{A}$.

Since the term f in the objective function (2) counts the number of triplets having a three-body radius close to $\bar{D}$, we made some more studies to find out frequent values of the three-body radius. We first analyzed the so-called perfect helix, that is the stable conformation of the amino acid sequence made only by alanine. In this helix, all the triplets (x_i, x_j, x_k), with constant $i - j$ and $j - k$, have the same radius. All the triplets (x_i, x_{i+h}, x_k), with $h > 0$ and $i + h < k$, and (x_i, x_{k-h}, x_k), with $h > 0$ and $k - h > i$ have the same radius too. The most frequent triplets with the same radius are of the type

(x_i, x_{i+1}, x_{i+3}) and (x_i, x_{i+2}, x_{i+3}), but the minimum radius, i.e. the thickness, is achieved by the triplets (x_i, x_{i+1}, x_{i+2}). The corresponding values, reported in Table 9, show that the difference between the minimum radius and the most frequent one amounts to 0.13 Å. We then considered the same type of triplets in the PDBSELECT set and computed the mean radius values, and the corresponding standard deviations, obtaining the results reported in Table 9. In this case, the difference between the mean thickness values (2.65 Å) and the triplet mean values is 0.10 Å for the triplets (x_i, x_{i+1}, x_{i+2}) and 0.23 Å for the triplets (x_i, x_{i+1}, x_{i+3}) and (x_i, x_{i+1}, x_{i+3}).

	perfect helix	**PDBSELECT set**
(x_i, x_{i+1}, x_{i+2})	2.71	2.75 (0.12)
(x_i, x_{i+1}, x_{i+3})	2.84	2.88 (0.28)
(x_i, x_{i+2}, x_{i+3})	2.84	2.88 (0.28)

Table 1. Three-body radii (Å) of selected triplets of α-carbons in the perfect helix and in a set of 3639 proteins from PDBSELECT. Mean and standard deviation (in brackets) of radius values are reported for the PDBSELECT triplets.

Taking into account the previous analysis, we set $\bar{D} = 2.70$ and $\epsilon = 0.20$, i.e. $[\Delta - \epsilon, \Delta + \epsilon] = [2.50, 2.90]$. This value of $\bar{D}$ is very close to the thickness of the perfect helix (2.71 Å); furthermore, the interval $[2.50, 2.90]$ contains most of the thickness values of the α-helices from PDBSELECT and includes also the most frequent three-body radii of both the perfect helix and the PDBSELECT α-helices.

The semiaxis lengths a, b and c that define the function g have been determined taking into account the volumes of the single amino acids, that are reported in Table 9. For each protein chain, we computed the sum of the volumes of the amino acids, then we increased this sum by 3.8%, to take into account that proteins have cavities [75], and, finally, we set a, b and c in such a way that their products was equal to the cube of the radius s of the sphere with volume equal to the increased sum of amino acid volumes, i.e.

$$a \cdot b \cdot c = s^3, \tag{10}$$

where

$$s = \frac{3}{4\pi} \sqrt[3]{1.038 \cdot \sum_{i=1}^{n} vol_i} \tag{11}$$

and vol_i is the volume of the i-th amino acid in the protein chain. Obviously, the single values of a, b and c are not univocally determined by (10)-(11);

by varying these values, theoretically possible conformations with different shapes can be obtained. Note that, by taking into account the amino acid volumes, we introduce in the model a distinction among the points x_i, that are considered equal in the original tube model.

amino acid	volume	amino acid	volume
ALA	88.6	LEU	166.7
ARG	173.4	LYS	168.7
ASP	111.1	MET	162.9
ASN	114.1	PHE	189.9
CYS	108.5	PRO	112.7
GLU	138.4	SER	89.0
GLN	143.8	THR	116.1
GLY	60.1	TRP	227.8
HIS	153.2	TYR	193.6
ILE	166.7	VAL	140.0

Table 2. The volumes of the 20 amino acids, in $\mathring{A}^3$.

To determine the constants c_1 and c_2, the mean value of the Euclidean distances of all pairs of consecutive α-carbons has been computed for each protein of the PDBSELECT set (the corresponding frequency distribution is shown in Figure 9). However, since the algorithm applied to problem (1)-(9) in our numerical experiments does not change these distances (see Section 10.1), we set c_1 and c_2 both equal to the most frequent mean Euclidean distance, i.e. $c_1 = c_2 = 3.81$ Å.

The remaining constants c_3, c_4 and c_5 have been chosen by observing the perfect helix. In this helix, the Euclidean distance between two α-carbons x_i and x_{i+2} is 5.43 Å, hence we set $c_4 = 5.0$ and $c_5 = 6.0$. Similar observations on the minimum distance between two generic α-carbons led to the choice $c_3 = c_4$. Actually, these choices of the c_i constants have been supported by numerical experiments.

10 Computational Experiments

Computational experiments based on the modified tube model have been carried out to simulate α-helices and an all-α protein, using a Metropolis Monte Carlo Simulated Annealing algorithm to search the conformational space. This algorithm has been implemented in Fortran 77 and in C and the software has been run on a personal computer with a 2 GHz Athlon processor and a 516 MBytes RAM, under the Linux operating system.

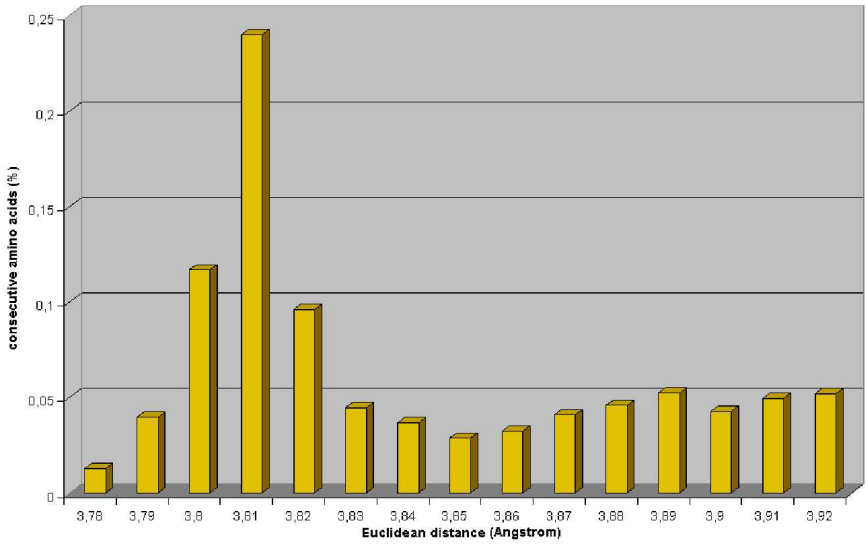

Fig. 9. Frequency distribution of the mean Euclidean distances of the pairs of consecutive amino acids in the proteins of the PDBSELECT set.

A short description of the Simulated Annealing algorithm and a discussion on the results of the computational experiments follow.

10.1 The Metropolis Monte Carlo Simulated Annealing Algorithm

As observed in Section 5, Simulated Annealing (SA) algorithms [40, 52] are based on an analogy with the annealing physical process, in which the temperature of a given system is decreased slowly, in order to obtain a crystalline structure. The structure of a SA algorithm can be described by two nested loops. The inner one generates at each iteration a new candidate approximation to the solution, by applying Monte Carlo perturbations to the previous one. The new approximation is accepted or rejected, by using a random mechanism based on the evaluation of the so-called Metropolis acceptance function, whose value depends on a parameter called temperature. The lower is the temperature, the smaller is the number of accepted approximations. The outer loop controls the decrease of the temperature parameter, i.e. defines the so-called cooling schedule.

From the above description it results that SA algorithms are built up from three basic components: next candidate generation, acceptance strategy and cooling schedule.

To generate the next candidate approximation to the solution, we use operations called *Monte Carlo moves* [88]. In particular, we consider the *pivot*, *multipivot* and *crankshaft* moves. The pivot move randomly selects a pivot

point x_i, with $1 < i < n$ and two coordinate axes ξ and η, and then rotates each point x_k, with $i < k \leq n$, of a random angle with respect to the axis through x_i and orthogonal to ξ and η. The multipivot move is obtained by performing a sequence of pivot moves. In our case, $n/10$ points x_i, with $1 < i < n$, are randomly selected and used as pivots. Finally, the crankshaft move randomly selects two points x_i and x_j, with $1 \leq i < j - 1 < n$, and then rotates the points x_k, with $i < k < j$, of a random angle around the axis passing through x_i and x_j.

The acceptance strategy used in our experiments is based on the well-known *Metropolis acceptance function* [55]. If $X^{(k)}$ is the approximation of the solution at a step k and $\bar{X}$ is a candidate approximation obtained by a Monte Carlo move, then $\bar{X}$ is accepted if

$$A(X^{(k)}, \bar{X}, t^{(k)}) = \min \left\{ 1, e^{\frac{F(\bar{X}) - F(X^{(k)})}{t^{(k)}}} \right\} > p,$$

where F is the objective function to be maximized (see (1)), $t^{(k)}$ is the temperature value at step k and p is a random number from the uniform distribution in $(0, 1)$. The candidate approximation can be accepted even if it does not increase the value of F, depending on $t^{(k)}$ and p. At high temperatures, many candidate approximations can be accepted, but, as the temperature decreases, the number of candidate approximations decreases, in analogy with the physical process of annealing.

The cooling strategy has an important role in SA. The temperature must be decreased very slowly to avoid trapping into local optima that are far from the global one. This reflects the behaviour of the physical annealing, in which a fast temperature decrease leads to a polycrystalline or amorphous state. In our experiments, a fixed number *nsteps* of Metropolis Monte Carlo iterations is performed at constant temperature and then the temperature value is decreased by a fixed factor $\gamma < 1$. The values of *nsteps* and γ have been experimentally set to $10^3 n$ and 0.99, respectively.

Our algorithm terminates when the value of the objective function F has not been changed for ten outer iterations, or a maximum number of outer iterations, *maxout*, is achieved. We set $maxout = 300$, but this value was never reached in our experiments. A sketch of the whole algorithm is provided in Figure 10.1.

We note that the cost of evaluating the term f in the objective function F (see (1) and (2)-(3)) is usually lower than $O(n^3)$. Indeed, if two points have a Euclidean distance greater than $2(\bar{D} + \epsilon)$, then all the triplets containing these points have a three-body radius greater than $\bar{D} + \epsilon$ (in a circle, a chord is smaller than the diameter) and hence they do not give any contribution to f. Therefore, once the Euclidean distances of all the pairs of points are computed, as required by the constraints (8), the three-body radii are computed only for triplets such that the Euclidean distance of all the pairs in the triplet is not greater than $2(\bar{D} + \epsilon)$.

```
t = t_0
X = random conformation satisfying the constraints
nout = 0

{outer loop}
while ( F(X) not settled down and nout ≤ maxout )
    nout = nout + 1

    {inner loop}
    for k = 1, nsteps
        X^(k) = random MC move on X
        if ( X^(k) satisfies the constraints ) then
            p = uniform random number in (0,1)
            if ( A(X, X^(k), t)) > p ) then
                X = X^(k)
            endif
        endif
    endfor

    t = γ · t
endwhile
```

Fig. 10. Metropolis Monte Carlo Simulated Annealing algorithm.

10.2 Simulation of α-helices

First experiments have been performed with very short amino acid chains and with the objective function of the original tube model, i.e. without considering the compactness term $g(X)$ in the objective function $F(X)$ (see (1)).

Many simulations have been carried out with $n = 10$ amino acids, starting from different initial conformations. All the computed optimal conformations are clock-wise rotated helices with about 3.6 points per helix turn, as in the real α-helices. About 60% of these conformations differ each other by a RMSD value of about 0.5 $\mathring{\text{A}}$; a maximum RMSD of 2.0 $\mathring{\text{A}}$ has been observed. The value of the objective function at the solution is always equal to 22 and is due to all the triplets (x_i, x_{i+1}, x_{i+2}), (x_i, x_{i+1}, x_{i+3}) and (x_i, x_{i+2}, x_{i+3}) (8, 7 and 7 triplets, respectively), which have a tree-body radius ranging between $\bar{D} - \epsilon$ and $\bar{D} + \epsilon$, where $\bar{D} = 2.70$ $\mathring{\text{A}}$ and $\epsilon = 0.20$ $\mathring{\text{A}}$, as discussed in Section 9. Each simulation was completed in about 7 seconds. An example of computed optimal conformation is shown in Figure 10.2.

Other experiments have been performed by changing the value of $\bar{D}$, but keeping $\epsilon = 0.20$. In this case, the computed conformations are unrealistic helices, with less than 3.6 points per turn if $\bar{D} < 2.70$ and more than 3.6 if

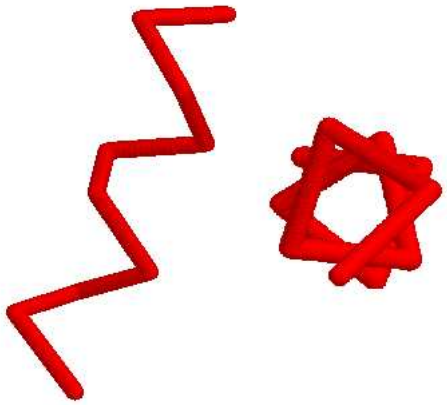

Fig. 11. Two views of a computed optimal conformation with $n = 10$ points $\left(\bar{D} = 2.70 \text{ Å}\right)$.

$\bar{D} > 2.70$. These results support the choice $\bar{D} = 2.70$. Some conformations obtained with different values of $\bar{D}$ are shown in Figure 10.2.

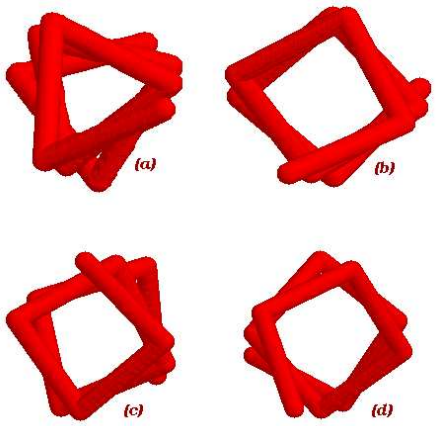

Fig. 12. Conformations obtained with $n = 10$ and different values of $\bar{D}$ ((a) $\bar{D} = 2.60$, (b) $\bar{D} = 2.80$, (c) $\bar{D} = 2.90$, (d) $\bar{D} = 3.20$).

Further experiments with $n > 10$ led to similar results. When $\bar{D} = 2.70$, conformations very close to real α-helices are obtained, while unrealistic he-

lices are generated for $\bar{D} \neq 2.70$. Furthermore, for $n > 30$, single long helices are computed which do not exist in nature, hence the need of introducing a compactness term into the problem objective function.

10.3 Simulation of All-α Proteins

Some experiments have been devoted to generate all-α protein conformations. A globular protein composed of 153 amino acid residues, the sperm whale myoglobin (PDB code `1mbn`), has been chosen as reference protein. Obviously, we did not expect to generate conformations very close to the myoglobin one, since the information contained in the considered model is too poor for an accurate fold prediction. On the other hand, we wished to analyze the reliability and accuracy provided by such a simplified model.

The lengths of the ellipsoid semiaxes a, b and c have been computed using the amino acid volumes of the selected protein, as explained in Section 9. According to the whole myoglobin shape, the following lengths have been considered: $a = b = 1.15s$ and $c = 0.76s$, where s is radius of the sphere with volume equal to the sum of the amino acid volumes, increased by 3.8% (see (11)), i.e. $s = 17.32$ Å. A few experiments with different semiaxis lengths have been also performed to analyze the weight of the compactness term $g(X)$ with respect to the thickness term $f(X)$ in the objective function $F(X)$. Sixty simulations have been performed until now, each requiring an execution time of about two hours. Better simulated conformations could be obtained by running a larger number of experiments.

The results obtained so far show that, as a, b and c get closer, the value of the term $f(X)$ at the solution decreases. A minimum value of 300 has been achieved with $a = b = c$. On the other hand, as the difference between two semiaxes increases, and hence the formation of longer helices is allowed, the value of $f(X)$ at the solution usually increases; $f(X) = 360$ has been obtained for $a = b = 1.2s$ and $c = 0.7s$. Conformations with values of $g(X)$ varying between 100 and 153 have been obtained, where larger values of $g(X)$ correspond to smaller values of $f(X)$.

Like the all-α proteins, the computed conformations are globular objects with secondary structures that are very close to real α-helices. For $a = b = 1.15s$ and $c = 0.76s$, i.e. for semiaxis lengths corresponding to the myoglobin shape, we obtained two conformations that have 66.7 and 59.5 identity percentages of secondary structures with respect to the reference protein. If we consider only the α-helices, the identity percentages are 67.8 and 51.6, respectively. This is shown in Figure 10.3. The corresponding three-dimensional representations are given in Figure 10.3. On the other hand, while having a certain similarity, the real protein and the computed conformations can have different numbers of helices, with different lengths and orientations, thus indicating that more information must be included in the model to perform more accurate simulations.

Fig. 13. A comparison of the sperm whale myoglobin with two simulated conformations. First row: amino acid sequence of the sperm whale myoglobin; second row: myoglobin amino acid residues that are contained into α-helices in the original conformation; third and fourth row: myoglobin amino acid residues that are contained into helices in the two simulated conformations. The identity percentage is 66.7 for the conformation called `test1` and 59.5 for the one called `test2`.

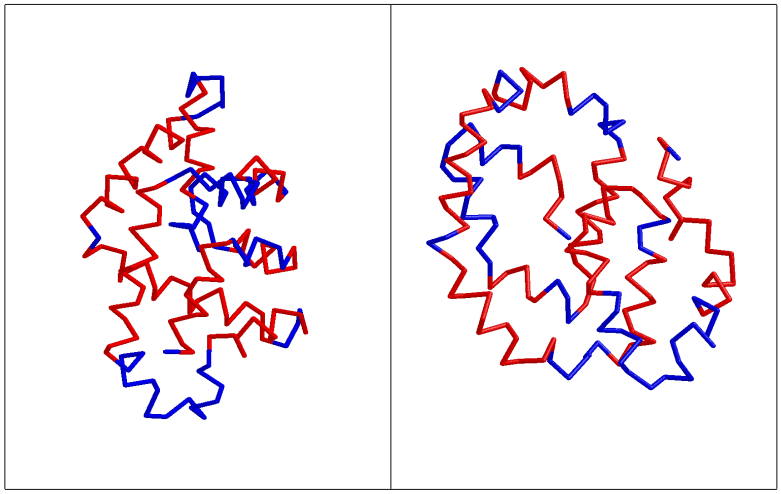

Fig. 14. Conformations obtained from the sperm whale myoglobin protein chain ($a = b = 1.15s$ and $c = 0.76s$). The α-helices are lighter, all the other structures are darker.

11 Conclusions

The great interest in the solution of the protein folding problem strongly pushes the research activity in this area. However, despite the many efforts performed so far, this problem is still considered a big challenge in science.

In this chapter we focused on ab-initio computational methods for protein fold predictions that are potentially able to discover unknown native state conformations. In this context, we analyzed an interesting topological approach, that takes into account geometrical rather than physicochemical protein features. This approach is based on a very simplified model that represents the polymer chain as a non-intersecting tube of nonzero thickness, by explicitly considering only the C_α trace of the protein and describing the amino acid interactions through the use of a suitable metric that measures the "distance" among any three C_α atoms. This model leads to the formulation of a global constrained optimization problem.

To enhance compactness and globularity in the computed conformations, we introduced a modification into the above model, and presented a methodology for choosing the values of characterstic parameters. The results of computational experiments devoted to simulating α-helices and all-α proteins can be considered "promising", especially if we take into account the great simplicity and the relatively low computational cost of the model. Indeed, simulations performed using the sperm whale myoglobin as target protein, generated a conformation with a percentage identity equal to 66.7. Hence, we expect that the model can be significantly improved by adding some physicochemical features to the geometrical ones currently considered. The introduction of the amino acid hydrophobicity into the model and the definition of ad hoc constraints and suitable parameter values for the simulation of β-strands and β-sheets are currently under investigation.

Acknowledgements

We wish to thank Davide Marenduzzo from University of Oxford, Department of Physics, for providing us an implementation of the Metropolis Monte Carlo Simulated Annealing algorithm, that we used as a basis for our implementation, and for some helpful discussions. This work has been partially supported by the MIUR FIRB projects "Large Scale Nonlinear Optimization" (grant no. RBNE01WBBB) and "Identification and functional analysis of gene and molecular modifications of hormone-responsive breast cancer" (grant no. RBNE0157EH_03).

References

1. N.L. Allinger. MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms. *Journal of the American Chemical Society*, 99(25): 8127-8134, 1977.
2. N.L. Allinger, Y.H. Yuh, and J.-H. Lii. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. *Journal of the American Chemical Society*, 111(23): 8551-8565, 1989.
3. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215: 403-410, 1990.
4. S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25 (17): 3389-402, 1997.
5. C.B. Anfinsen, E. Haber, M. Sela, and F.H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences*, 47: 1309-1314, 1961.
6. I.P. Androulakis, C.D. Maranas, and C.A. Floudas. αBB: A Global Optimization Method for General Constrained Nonconvex Problems. *Journal of Global Optimization*, 7(4): 337-363, 1995.
7. J.R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan, and A. Trovato. Geometry of Compact Tubes and Protein Structures. *ComPlexUs*, 13: 1-4, 2003.
8. J.R. Banavar, O. Gonzalez, J.H. Maddocks, and A. Maritan, Self-interactions of strands and sheets. *Journal of Statistical Physics*, 110: 35-50, 2003.
9. J.R. Banavar, A. Maritan, C. Micheletti, and F. Seno. *Geometrical aspects of protein folding*, Lectures held at the "Enrico Fermi Summer School", Varenna, Italy, 2001.
10. J.R. Banavar, A. Maritan, C. Micheletti, and A. Trovato. Geometry and Physics of Protein. *Proteins*, 47(3): 315-322, 2002.
11. D. Baker. A surprising simplicity to protein folding. *Nature*, 405: 39-42, 2000.
12. R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C.E.M. Strauss, and D. Baker. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins: Structure, Function and Genetics Supplement*, 5: 119-126, 2001.
13. P. Bradley, D. Chivian, J. Meiler, K.M.S. Misura, C.A. Rohl, W.R. Schielf, W.J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C.E.M. Strauss, and D. Baker. Rosetta Predictions in CASP5: Successes, Failures, and Prospects for Complete Automation. *Proteins: Structure, Function and Genetics Supplement*, 53: 457-468, 2003.
14. C. Caporale, A. Facchiano, L. Bestini, L. Leopardi, G. Chiosi, V. Buonocore, and C. Caruso. Comparing the modelled structures of PR-4 proteins from wheat. *Journal of Molecular Modeling*, 9: 9-15, 2003.
15. CHARMM Home Page, `http://www.charmm.org/`.
16. EMBnet Home Page, `http://www.ch.embnet.org/MD_tutorial/`.

17. A.M. Facchiano, P. Stiuso, M.L. Chiusano, M. Caraglia, G. Giuberti, M. Marra, A. Abruzzese, and G. Colonna. Homology modelling of the human eukaryotic initiation factor 5A (eIF-5A). *Protein Engineering*, 14: 881-890, 2001.
18. J.S. Fetrow, A. Giammona, A. Kolinski, and J. Skolnick. The protein folding problem: A Biophysical Enigma. *Current Pharmaceutical Biotecnology*, 3: 329-347, 2002.
19. G.S. Fishman, editor. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, 1996.
20. C.A. Floudas. *Deterministic global optimization: theory, methods and applications*. Kluwer Academic Publishers, 2000.
21. C.A. Floudas, J.L. Klepeis, and P.M. Pardalos. Global Optimization Approaches in Protein Folding and Peptide Docking. In M. Farach, F.S. Roberts, M. Vingron, and M. Waterman, editors, *Mathematical Support for Molecular Biology*, pages 141–171. DIMACS Series, Volume 47, American Mathematical Society, Providence, RI, 1999.
22. I. Friedberg, T. Kaplan, and H. Margalit. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Science*, 9: 2278-2284, 2000.
23. O. Gonzalez and J.H. Maddocks. Global curvature, Thickness and the Ideal Shapes of Knots. *Proceedings of the National Academy of Sciences*, 96: 4769-4773, 1999.
24. T.X. Hoang, M. Cieplak, J. Banavar, and A. Maritan. Prediction of Protein Secondary Structures from Conformational Biases. *Proteins: Structure, Function and Genetics*, 48: 558-565, 2002.
25. D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292: 195-202, 1999.
26. D.T. Jones. Predicting novel protein folds by using FRAGFOLD. *Proteins: Structure, Function and Genetics Supplement*, 5: 127-132, 2001.
27. D.T. Jones. Critically assessing the state-of-art in protein structure prediction. *The Pharmacogenomics Journal*, 1(2): 126-134, 2001.
28. D.T. Jones and L.J. McGuffin. Assembling novel protein folds from super-secondary structural fragments. *Proteins: Structure, Function and Genetics*, 53: 480-485, 2003.
29. JUFO Home Page, `http://www.jens-meiler.de/jufo.html`.
30. W. Kabsch and C. Saunder. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22: 2577-2637, 1983.
31. M. Karplus and J.N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14: 325-332, 1981.
32. S.-Y. Kim, S.J. Lee and J. Lee. Conformational space annealing and an off-lattice frustrated model protein. *Journal of Chemical Physics*, 119: 10274 - 10279, 2003.
33. J.L. Klepeis and C.A. Floudas. Deterministic Global Optimization for Protein Structure Prediction. In C. Caratheodory, N. Hadjisavvas and P.M. Pardalos, editors, *Advances in Convex Analysis and Global Optimization*, pages 31-74, Kluwer, 2001.
34. J.L. Klepeis and C.A. Floudas. ASTRO-FOLD: Ab Initio Secondary and Tertiary Structure Prediction in Protein Folding. In J. van Schijndel, ed-

itor, *European Symposium on Computer Aided Process Engineering*, Volume 12, Elsevier Applied Science, 2002.
35. J.L. Klepeis and C.A. Floudas. Ab initio tertiary structure prediction of proteins. *Journal of Global Optimization*, 25: 113-140, 2003.
36. J.L. Klepeis and C.A. Floudas. ASTRO-FOLD: a combinatorial and global optimization framwork for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical Journal*, 85: 1-28, 2003.
37. J.L. Klepeis and C.A. Floudas. Prediction of *beta*-Sheet Topology and Disulfide Bridges in Polypeptides. *Journal of Computational Chemistry*, 24: 191-208, 2003.
38. J.L. Klepeis and C.A. Floudas. Analysis and Prediction of Loop Segments in Protein Structures. *Computers & Chemical Engineering*, 29: 423-436, 2005.
39. J.L. Klepeis, Y. Wei, M.H. Hecht and, C.A. Floudas. Ab initio Prediction of the 3-Dimensional Structure of a De novo Designed Protein: A Double Blind Case Study. *Proteins*, 58: 560-570, 2005.
40. S. Kirkpatrick, C.D. Gelatt Jr., and M.P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598): 671-680, 1983.
41. P. Koehl and M. Levitt. Improved recognition of native-like protein structures using a family of designed sequences. *Proceedings of the National Academy of Sciences*, 99(2): 691-696, 2002.
42. N. Koga and S. Takada. Roles of Native Topology and Chain-length Scaling in Protein Folding: A Simulation Study with a Go-like Model. *Journal of Molecular Biology*, 313: 171-180, 2001.
43. A. Kolinski, P. Rotkiewicz, B. Ilkowski and J. Skolnick. Protein Folding: Flexible Lattice Models. *Progress of Theoretical Physics*, 138: 292-300, 2000.
44. A. Kolinski, P. Rotkiewicz and J. Skolnick. Structure of proteins: New Approach to Molecular Modeling. *Polish Journal of Chemistry*, 75: 587-599, 2001.
45. A. Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, 51(2): 349-371, 2004.
46. M. Kuhn, J. Meiler and D. Baker. Strand-loop-strand motifs: prediction of hairpins and diverging turn in proteins. *Proteins: Structure, Function and Bioinformatics*, 54: 282-288, 2004.
47. J. Lee, H. A. Scheraga and S. Rackovsky. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. *Journal of Computational Chemistry*, 18: 1222-1232, 1997.
48. C. Levinthal. Are there pathways for protein folding? *Chemical Physics*, 65: 44-45, 1968.
49. M. Levitt and A. Hinds. A lattice model for protein structure prediction at low resolution. *Proceedings of the National Academy of Sciences*, 89: 2536-2540, 1992.
50. A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky, and H.A. Scheraga. Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment. *Protein Science*, 2: 1697-1714, 1993.
51. A. Liwo, J. Lee, D.R. Ripoll, J. Pillardy, and H.A. Scheraga, Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences*, 96: 5482-5485, 1999.

52. M. Locatelli. Simulated Annealing Algorithms for Continuous Global Optimization. In P.M. Pardalos and H.E. Romeijn, editors, *Handbook of Global Optimization*, Volume 2, pages 179-229. Kluwer Academic Publishers, 2002.
53. C.D. Maranas, L.P. Androulakis, and C.A. Floudas. A Deterministic Global Optimization Approach for the Protein Folding Problem. In P. M. Pardalos, D. Shalloway, and G. Xue, editors, *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, pages 133-150. DIMACS Series, Volume 23, American Mathematical Society, Providence, RI, 1996.
54. D. Marenduzzo, A. Flammini, A. Trovato, J.R. Banavar, and A. Maritan. Physics of thick polymers. *Journal of Polymer Science, Part B: Polymer Physics*, 43: 650679, 2005.
55. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth. A.H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21: 1087-1092, 1953.
56. C. Micheletti. Prediction of Folding rates and Transition-State Placement From Native-State Geometry. *Proteins: Structure, Function and Genetics*, 51: 74-84, 2003.
57. F.A. Momany, L.M. Carruthers, R.F. McGuire, and H.A. Scheraga. Intermolecular potentials from crystal data. III. *Journal of Physical Chemistry*, 78: 1595-1620, 1974.
58. F.A. Momany, L.M. Carruthers, and H.A. Scheraga. Intermolecular potentials from crystal data. IV. *Journal of Physical Chemistry*, 78: 1621-1630, 1974.
59. G. Némety, M.S. Pottle, and H.A. Scheraga. Energy Parameters in Polypeptides. 9. *Journal of Physical Chemistry*, 87: 1883-1887, 1983.
60. G. Némety, K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H.A. Scheraga. Energy Parameters in Polypeptides. 10. *Journal of Physical Chemistry*, 96: 6472-6484, 1992.
61. D.J. Osguthorpe. Ab initio protein folding. *Current Opinion in Structural Biology*, 10: 146-152, 2000.
62. P.M. Pardalos and H.E. Romeijn, editors. *Handbook of Global Optimization*, Volume 2. Kluwer Academic Publishers, 2002.
63. P.M. Pardalos and G. Xue, editors. *Advances in Computational Chemistry and Protein Folding. Journal of Global Optimization*, Special Issue, 4(2), 1994.
64. P.M. Pardalos, D. Shalloway, and G. Xue, editors. *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*. DIMACS Series, Volume 23, American Mathematical Society, Providence, RI, 1996.
65. PDBSELECT Home page `http://www.cmbi.kun.nl/swift/pdbsel/`.
66. S. Petit-Zeman. Treating protein folding diseases. *Nature*, 2002, available at `http://www.nature.com/horizon/proteinfolding/background/-treating.html`
67. J. Pietzsch. The importance of Protein Folding. *Nature*, 2002, available at `http://www.nature.com/horizon/proteinfolding/background/-importance.html`

68. J. Pietzsch. Protein Folding diseases. *Nature*, 2002, available at `http://www.nature.com/horizon/proteinfolding/-background/disease.html`.
69. K.W. Plaxco, K.T. Simons and D. Baker. Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins. *Journal of Molecular Biology*, 277: 985-994, 1998.
70. Protein Structure Prediction Center Home Page, `http://predictioncenter.llnl.gov`.
71. PROSPECTOR Home Page, `http://www.bioinformatics.buffalo.edu/-new_buffalo/services/threading.html`.
72. D.R. Ripoll, A. Liwo, and H.A. Scheraga. New Developments of the Electrostatically Driven Monte Carlo Method: Test on the Membrane-Bound Portion of Melittin. *Biopolymers*, 46: 117, 1998.
73. D.R. Ripoll and H.A. Scheraga. On the multiple-minima problem in conformational analysis of polypeptides. IV. *Biopolymers*, 30: 165-176, 1990.
74. D.R. Ripoll, M.J. Vàsquez, and H.A. Scheraga. The electrostatically driven Monte Carlo method - Application to conformational analysis of decaglycine. *Biopolymers*, 31: 319-330, 1991.
75. K. Rother, R. Preissner, A. Goede, and C. Frommel. Inhomogeneous molecular density: reference packing densities and distribution of cavities within proteins. *Bioinformatics*, 19(16): 2112-2121, 2003.
76. I. Ruczinski, C. Kooperberg, R. Bonneau and D. Baker. Distributions of Beta Sheets in Proteins With Application to Structure Prediction. *Proteins: Structure, Function and Genetics*, 48: 85-97, 2002.
77. R. Samudrala, Y. Xia, E. Huang, and M. Levitt. Ab initio Protein Structure Prediction Using a Combined Hierarchical Approach. *Proteins: Structure, Function and Genetics Supplement*, 3: 194-198, 1999.
78. J.A. Saunders, K.D. Gibson, and H.A. Scheraga. Ab initio folding of multiple-chain proteins. *Pacific Symposium on Biocomputing*, 7: 601-612, 2002.
79. G. Scapigliati, S. Costantini, G. Colonna, A. Facchiano, F. Buonocore, P. Boss, J.W. Holland, and C.J. Secombes. Modelling of fish interleukin 1 and its receptor. *Developmental and Comparative Immunology*, 28: 429-41, 2004.
80. G. Settanni, A. Cattaneo, and A. Maritan. Role of Native-State Topology in the Stabilitazion of Intracellular Antibodies. *Biophysical Journal*, 81: 2935-2945, 2001.
81. K.T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Function. *Journal of Molecular Biology*, 268: 209-225, 1997.
82. K.T. Simons, R. Bonneau, I. Ruczinski, and D. Baker. Ab Initio Protein Structure Predictions of CASP III Targets Using ROSETTA. *Proteins: Structure, Function and Genetics Supplement*, 3: 171-176, 1999.
83. J. Skolnick, A. Kolinski, D. Kihara, M. Betancourt, P. Rotkiewicz, and M. Boniecki. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins: Structure, Function and Genetics Supplement*, 5: 149-156, 2001.

84. J. Skolnick, D. Kihara, H. Lu, and A. Kolinski. TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proceedings of the National Academy of Sciences*, 98(18): 10125-10130, 2001.
85. J. Skolnick, D. Kihara, Y. Zhang, H. Lu, and A. Kolinski. Ab initio protein structure prediction on a genomic scale: Application to the Mycoplasma genitalium genome. *Proceedings of the National Academy of Sciences*, 99(9), 5993-5998, 2002.
86. J. Skolnick, Y. Zhang, A.K. Arakaki, A. Kolinski, M. Boniecki, A. Szilágyi, and D. Kihara. TOUCHSTONE: A Unified Approach to Protein Structure Prediction. *Proteins: Structure, Function and Genetics*, 53: 469-479, 2003.
87. J.E. Smith. Genetic Algorithms. In P.M. Pardalos and H.E. Romeijn, editors, *Handbook of Global Optimization*, Volume 2, pages 275-362. Kluwer Academic Publishers, 2002.
88. A.D. Sokal. Monte Carlo methods for the self-avoiding walk. *Nuclear Physics B (Proceedings Supplements)*, 47: 172-179, 1996.
89. R. Srinivasan and G.D. Rose. LINUS: a hierarchic procedure to describe the fold of a protein. *Proteins*, 22: 81-99, 1995.
90. R. Srinivasan and G.D. Rose. A physical basis for protein secondary structure. *Proceedings of the National Academy of Sciences*, 96(25): 14258-14263, 1999.
91. R.H. Swendsen and J.S. Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters*, 57: 2607-2609, 1986.
92. A. Trovato. *A Geometric Perspective on Protein Structures and Heteropolymer Models*. PhD Thesis, SISSA, Trieste, 2000.
93. S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A New Force Field for Molecular Mechenical Simulation of Nucleic Acids and Proteins. *Journal of the American Chemical Society*, 106: 765-784, 1984.
94. Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala. Ab Initio Construction of Protein Tertiary Structures Using a Hierarchical Approach. *Journal of Molecular Biology*, 300: 171-185, 2000.
95. A. Zemla, C. Venclovas, K. Fidelis, and B. Rost. A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment. *Proteins: Structure, Function and Genetics*, 34: 220-223, 1999.
96. Y. Zhang, D. Kihara, and J. Skolnick. Local Energy Landscape Flattering: Parallel Hyperbolic Monte Carlo Sampling of Protein Folding. *Proteins: Structure, Function and Genetics*, 48: 192-201, 2002.
97. Y. Zhang, A. Kolinski, and J. Skolnick. TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction. *Biophysical Journal*, 85: 1145-1164, 2003.