Coarse Grained Modeling and Approaches to Protein Folding

Carlo Guardiani[1], Roberto Livi[2] and Fabio Cecconi*[,3]

[1]*Centro Interdipartimentale Studio Dinamiche Complesse (CSDC) Sezione INFN di Firenze, (Italy);* [2]*Dipartimento di Fisica Università di Firenze and Centro Interdipartimentale Studio Dinamiche Complesse (CSDC), Sezione INFN di Firenze e INFM UdR Firenze, Italy;* [3]*INFM-CNR Center for Statistical Mechanics and Complexity (SMC) Istituto dei Sistemi Complessi (ISC-CNR) Via dei Taurini 19, 00185 Rome, Italy*

Abstract: The theoretical prediction of protein structures has become a field of increasing importance in both biology and physics. Reliable prediction methods in fact, would spare time consuming experimental X-ray and NMR techniques and they would represent a challenge for computational protein modeling as well. The well known limitations of all-atom models call for the development of coarse-grained protein descriptions including a minimal number of protein-like features, while being capable of mimicking the essence of protein folding mechanisms. In this paper we review the most important classes of coarse-grained protein models in order of increasing complexity, starting from (over simplified) binary models, to models with one or two reaction centers per residue. We discuss how, despite their simplification, coarse-grained models constitute a viable approach to structure prediction and they shed light on many aspects of protein-folding problem.

Keywords: Protein folding, protein modelling, coarse-graining, united atom, effective potentials.

INTRODUCTION

The word protein derives from the Greek πρωτευω which means "being of primary importance". No better term could have been chosen to describe this class of molecules playing a key role in practically all biological processes [1]. Proteins are involved, for instance, in enzymatic c catalysis, muscle contraction, immunitary defense, transmission of nervous signals, transport of charges and metabolites etc. The possibility for proteins to perform their biological activities depends crucially on their specific tridimensional structure [2]. Therefore, accurate information about a three-dimensional protein structure is required when the action mechanism of that protein is to be studied in detail. The determination of protein 3D-structures is far from trivial and is usually accomplished via X-ray crystal diffraction [3] and nuclear magnetic resonance (NMR) [4]. The X-ray technique provides high-resolution structural information as atom positions can be determined with errors lower than 2Å. Unfortunately, protein crystal preparation is very critical and often slow: for some proteins it may take months to years to get a crystal clean and large enough for a X-ray experiment. The main advantage of NMR technique is that it does not need crystals and can be applied to proteins in solution provided that the concentration is high enough. NMR has got a couple of drawbacks: its resolution power is less than that of X-ray diffraction and it cannot be applied to very large proteins. It follows that such experimental techniques have serious limits in that they cannot be applied to all proteins and the experiments may be time-consuming. The development of computer methods for the prediction of protein structures from the amino acid sequence would therefore make faster and easier the study of the several new proteins that are discovered every day [5]. With this respect, it is worthwhile noticing that, even if bioinformatics techniques, such as homology modeling and threading, are quite effective in structural prediction [6], they do not provide any insight in the folding mechanism and in forces driving it. These problems can be better addressed with physics-based models that will be also applicable to the study of conformational changes, protein-ligand binding and the action mechanisms of macromolecules.

Anfinsen [7] postulated that, the amino-acid sequence contains all the information necessary to determine the native structure of a protein i.e. the conformation with the lowest free energy. Within this framework, the protein folding problem, i.e. the prediction of the three-dimensional structure of a protein by only knowing its amino-acid sequence, can be addressed as a global optimization problem. Two ingredients are therefore necessary: a realistic energy function and an efficient algorithm for the sampling of the conformational space. This review will be mainly concerned with a comparative analysis of protein models with different degrees of simplification.

Quantitative sciences often resort to models to describe and explain natural phenomena. A physical model is a simplified picture of a real system that captures all the essential features and neglects irrelevant details. An *ab- initio* description of the dynamics of the atoms in proteins involves Quantum Mechanics, but this kind of approach for complex molecules is still beyond any available computational resource. Hence, it is customary to use a classical description of molecules in terms of bonds and effective atomic interactions, the only trace left of the electrons being the partial charges on the atoms.

The most straightforward approach is to consider the all-atom representation of the protein molecule together with

*Address correspondence to this author at the INFM-CNR Center for Statistical Mechanics and Complexity (SMC) Istituto dei Sistemi Complessi (ISC-CNR) Via dei Taurini 19, 00185 Rome Italy; Tel: +39 06.4993.7452; Fax: +39 06.4993.7440; E-mail: fabio.cecconi66@gmail.com

1574-8936/10 $55.00+.00

empirical potential functions written as a sum of several contributes [8,9]. The constants that describe molecular geometry and the strength of particular inter-atomic interactions are generally parametrized on empirical structural, spectroscopic and thermodynamic data available from small organic molecules. The potential energy is expressed in the form of atom-centered potentials with the energy of the molecule computed as a sum over all interactions. Therefore the number of additive terms in potential functions is large, leading to extremely long computation times. A further limitation arises in all-atom Molecular Dynamics simulations where the time-step is usually chosen one order of magnitude smaller than the period of the fastest oscillations of the system. Typically, the fastest stretching motion of the C-H bond imposes a time-step of the order of the femto-second. Currently all-atoms simulations sample no more than a few tens of nano-seconds.

This kind of problems can be partially overcome by using united residue simplified models at the price to approximate the folding process. In such models, every aminoacid residue is described by two interaction centers, one representing the side chain (assimilated to a sphere or an ellipsoid), the other representing the peptide group. The Cα atoms are retained to give geometric constraints to the structure, but they are not directly involved in any interaction. Levitt [10] carried out a united-residue simulation on Pancreatic Trypsin Inhibitor reporting a gain of three orders of magnitude in the total simulation time over conventional all-atom methods. The united residue approach was further developed by Scheraga and his group [11-13], who also introduced a cumulant expansion of the free energy, accounting for the multi-body interactions which proved to be of paramount importance for β-sheet and α-helix formation [13,14].

A further level of coarse-graining is represented by the class of one-bead models, where amino-acid residues are represented by beads centered on the position of the α-carbons and they are strung together by virtual bonds. Usually only two or three types of amino acids appear: hydrophobic, hydrophilic and possibly neutral, allowing simplified forms of the interaction between residues to be used. The aim is not providing an algorithm for 3D-structure determination from the amino acid sequence, but rather focusing on particular kinetic and thermodynamical aspects of the folding process. Investigations in this context include: the compact filling of space, the preferential localization of polar and nonpolar aminoacids, the balance of long and short range interactions, the cooperativity of collapse into a compact structure and the discrimination between good and bad folders.

A final, extreme form of coarse-graining is represented by the Ising-like protein models [15,16] where the protein is portrayed as an array of binary variables that may represent residues, peptide bonds or native contacts. Each of these units can exist in two exclusive states: *native-like* and *unfolded*. The most important advantage consists in the significant reduction of the conformational space. For a protein of N residues, the conformation space will only contain 2^N structures so that for suitable N, exhaustive enumeration becomes feasible. Another important feature is that, due to the extreme simplification, Ising-like models are often solvable analytically and amenable to more rigorous treatments [17].

So far, we have classified protein models according to their structural representation of the protein chain. Models however, can also be classified according to whether they are *topology-based* or *sequence-based*. Before discussing key elements of these two classes however, it is necessary to identify which features a good protein model should have. Many of these features stem from the comparison of naturally occuring proteins and random heteropolymers [18]. Natural proteins are characterized by a almost unique, stable native state that can be reached in a reasonably short time-scale (milliseconds to seconds). Moreover, the folding of many natural proteins is often a cooperative process akin to a first order phase transition. This means that intermediate states between the native and the unfolded ones are never significantly populated. This is in sharp contrast with the behavior of random heteropolymers that have no unique native state and may fold in different conformations depending on initial conditions. At variance with proteins, the folding of random heteropolymers is a gradual process with intermediate structures dominating the population at intermediate temperatures. Finally, random heteropolymers fold slowly due to the presence of many kinetic traps. The different properties of natural proteins and heteropolymers are caused by the different topography of their energy landscape [19,20]. Protein energy landscapes, "sculpted" by millions of years of evolution, are funnel-like with rugged walls. The native state is placed at the bottom of the funnel which guarantees thermodynamic stability and fast accessibility. During the folding process, in fact, the protein just climbs down the landscape and the decrease in entropy is accompanied by a decrease in energy so that no significant free energy barriers will arise. At the folding temperature, the unfolded state and native state, will have the same free energy and the protein jumps straight from the unfolded to the native state so that folding will occurr in an *all-or-nothing* fashion. By contrast, the energy landscape of random heteropolymers is glass-like with many degenerate minima, each one representing the end-point of the folding process. Moreover, many metastable minima present in this frustrated landscape act as kinetic traps, slowing down the folding process.

The above discussion suggests that protein models can be assessed according to their ability to reproduce a funnel-shaped energy landscape with a moderate amount of frustration. These requirements are surely satisfied by the topology-based models originally introduced by Go *et al.* [21, 22] in 1981. In the Go-model only residues forming native contacts interact attractively, whereas the others interact through a repulsive, excluded-volume potential. As a decrease in energy can only be attained through an increase in the fraction of native contacts, the Go-model minimizes frustration, assuring a fast folding behavior. A number of experimental evidences underscores the importance of the topology of the native state supporting the use of Go-models. In particular, proteins with similar native states and similar transition states, exhibit similar folding pathways in spite of their lack of sequence homology. Moreover, simple topological parameters such as the contact order were found to correlate well with the folding rates of small globular proteins [23, 24]. The Go-model, producing a perfect funnel with minimal frustration, represents a sort of "ideal limit" for the folding,

and deviations from it are not rare in real proteins. For instance, the classic Go-model cannot discriminate the folding pathways of protein-L and protein-G, that share the same topology, but differ in their level of frustration [25]. Moreover, the folding mechanism proposed by Go-models can result artificial, since in nature, aminoacid residues interact by their physical and chemical properties and not on the basis of a "knowledge" of their neighbors in the native state.

The challenge of physical modeling of protein folding is therefore the development of sequence-based protein models, still capable of producing a funneled energy landscape with a level of frustration higher than Go-models, but much lower than that of random heteropolymers. Sequence-based models must include the most important driving forces of folding [18]. Protein folding is mainly driven by the hydrophobic effect, i.e. the increase in entropy of the water caused by the destruction of the *clathrates* when the protein folds. Clathrates are highly ordered, cage-like structures formed by water molecules surrounding the side-chains of apolar residues. Globular proteins fold in such a way that apolar residues are confined in a solvent-excluded, hydrophobic core, with release of a large number of water molecules. As it is extremely difficult to exactly reproduce this complex mechanism, the hydrophobic effect is usually modeled through an effective Lennard-Jones potential also accounting for van der Waals interactions. Residues effectively behave as if they attracted each other, leading to the collapse of the protein into a compact conformation. The protein folding is also driven by a set of short-range and long-range interactions. Short-range interactions involve atoms and residues close along the primary sequence and include stretching, bending and dihedral potentials. Long-range interactions arise among residues close to each other, even if distant along the primary sequence. They include ionic interactions, disulfide bridges and hydrogen bonds.

In this Review, we will describe the most important families of coarse-grained protein models in order of increasing complexity. In the first Section, we review Ising-like models, in the second one, we discuss one-bead models, and in the third, we describe two-interaction-centers models. Finally, in last section, we draw the conclusions of our discussion.

ISING-LIKE MODELS

Ising-like models are statistical mechanical models where every element of the system can only take on two alternative states. This corresponds to simplifying a protein as an array of residues, peptide bonds or contacts, that can be either "native" or "non-native". The first Ising-like natural approach focused on the helix-coil transition [26], because in a α-helix each residue only interacts with its sequence neighbors. The model correctly showed that the helix-coil transition occurs in two steps: nucleation and elongation.

The application of Ising-like approach to β-sheet structures requires the definition of interaction rules for residues that are not sequence neighbors [27]. This is usually done by using structure-based interactions: residues are allowed to interact only if they establish contacts in the native structure.

In the Galzitskaya Finkelstein (GF) model [16], each residue can be either folded ($s_i = 1$) or unfolded ($s_i = 0$). The protein populates its native state when all binary variables $s_i = 1$, whereas it is a random coil when $s_i = 0$ for all i. Two such residues can interact even when all the other residues in the chain between them are in a coil conformation. The Hamiltonian (actually a free-energy) of the GF model is:

$$H(\mathbf{s}) = \varepsilon \sum_{i<j} \Delta_{ij} s_i s_j - TS(\mathbf{s}) \tag{1}$$

where the conformation entropy S(s) is computed as

$$S(\mathbf{s}) = R\left[q\sum_{i=1}^{L}(1-s_i) + S_{\text{loop}}(\mathbf{s}) \right] \tag{2}$$

with ε, *R*, *T*, *L*, being the energy scale, the gas constant, the absolute temperature and the number of residues, respectively. In Eq.(1), Δ is the contact matrix, whose elements represent the number of heavy-atom contacts in the native state. Non-native contacts do not contribute to the stabilization energy, so that frustration is minimal. The conformational entropy (2) is split in two contributions. The term $Rq\,\Sigma_i(1-s_i)$ is just the sum of the entropies *Rq* of each residue in the unfolded state, while RS_{loop} is the entropy of closing disordered loops computed as,

$$S_{\text{loop}}(\mathbf{s}) = \sum_{i<j} J(r_{ij}) \prod_{k=i+1}^{j-1} (1-s_k) s_i s_j$$

where

$$J(r_{ij}) = -\frac{5}{2}\ln|i-j| - \frac{3}{4}\frac{r_{ij}^2 - d^2}{Ad|i-j|} \tag{3}$$

Expression (3) simply portrays a disordered loop as a random walk with end-to-end distance r_{ij} . The factor 5/2 accounting for the excluded volume effect exerted by the globule surface on loops, replaces the classical factor 3/2 of random walks (see Ref. [16]). The parameters *A* and *d* represent the persistence length and the average distance between consecutive alpha-carbons respectively.

Another interesting model has been proposed by Wako and Saito [28] and later generalized by Muñoz and Eaton (WS-ME) who applied it to a16-residue β-hairpin [15, 29]. It is based on the balance between the destabilization caused by loss of conformational entropy upon folding and the stabilizing effects of native hydrogen bonds and hydrophobic interactions. Binary variables characterizing the protein state are pairs of backbone dihedrals $\psi_i\,\varphi_{i+1}$ assumed to shift between native and non-native values in a coordinated manner. The structural unit of the system is therefore the peptide bond. The free energy function (meant as the Hamiltonian) of WS-ME model can be written as

$$H(\mathbf{s}) = -J\sum_{i<j}\Delta_{ij}\prod_{n=i}^{j} s_n + T\Delta S_{\text{conf}}\sum_{i=1}^{L} s_i \tag{4}$$

where J is the interaction strength, Δ_{ij} is a contact matrix element, s_i is the binary variable describing the status of the i-th peptide bond, T is the temperature and ΔS_{conf} is the loss of conformational entropy per native bond. The product $\Pi_n s_n$ in Eq.(4) constrains peptide bonds i and j to interact only if all the $i+1, i+2, \ldots, j-2, j-1$ bonds are native. As a consequence, the entropic cost of loop closure is very high.

In 1999, Alm and Baker (AB) [30] proposed another native-centric binary model similar to GF and WS-ME approach, whose free-enegy function reads

$$H = -\gamma\Delta(ASA) + KT(\alpha n + \beta\ln(L/L_0)$$

where K is the Boltzmann constant and T the absolute temperature. In the first term native-attractive interactions are proportional to the variation of the accessible surface area of residue pairs forming native contacts (with $\gamma = 16$ cal mol^{-1}Å^{-2}). The term $KT\,\alpha n$ represents the entropic cost of ordering n residues (with $\alpha = 175$ *K*cal mol^{-1}), while $KT\beta \ln(L/L_0)$ indicates the entropic cost of loop closure and only applies to conformations with two consecutive stretches of native residues ($\beta = 1.8$ $L_0 = 0.15$), L being the loop length.

The interaction rules of the above models are based on the concept of contact order [26] which is the average loop length in a protein structure [23]. The known correlation of the contact order with the folding rate suggests that Ising-like models may reliably reproduce the kinetic behavior of folding [31]. The main difference is the entropic cost of loop closure which in the WS-ME is much larger than in GF and AB models.

The discreteness of the conformation space of binary models allows their kinetics and thermodynamics to be easily studied. In fact, at equilibrium, states are sampled with probability:

$$P_{eq}(\mathbf{s}) = \frac{e^{-H(\mathbf{s})/RT}}{\sum_{s'} e^{-H(\mathbf{s}')/RT}},$$

it is therefore possible to monitor the composition of the population by means of a vector **P**(t) of large but finite dimension. The time evolution of the probability vector can be attained from the Master Equation

$$\frac{d\mathbf{P}}{dt} = -\mathbf{M}\,\mathbf{P}$$

with the matrix **M** whose entries are the transition probabilities from conformation **s** to **s'** defined as

$$M_{s,s'} = \begin{cases} \tau_0^{-1} & H(\mathbf{s}') < H(\mathbf{s}) \\ \tau_0^{-1} \exp\{-[H(\mathbf{s}) - H(\mathbf{s}')]/RT\} & H(\mathbf{s}') > H(\mathbf{s}) \end{cases}$$

$1/\tau_0$ being the attempt rate. The transition matrix **M** is such that the sum of the elements of each column is zero. The Master Equation discretized in time, as $\mathbf{P}(t + h) = (1 - h\mathbf{M})\,\mathbf{P}(t)$ allows the kinetics to be simulated. This approach for WS-ME model was applied by Cieplak *et al*. [32] to a shorter 12-residue version of the original β-hairpin in order to sample the Transition State Ensemble exploiting the complete enumerability of conformations in the single sequence approximation [*]. By convention, conformations *1* and *67* were assumed as the native and fully unfolded states respectively. Then, such states can act as probability sinks just setting to zero the first and last column of matrix **M**. In Ref. [32], the flow of probability was studied by starting from a state where only one conformation is populated. The authors discovered that six states had equal probabilities to flow towards the native and the unfolded states, and are therefore located at the boundary between the native and non-native basins. Among these edge conformations, those with the lowest energy identify the transition state ensemble.

The method for studying the Transition State Ensemble developed by Galzitskaya and Finkelstein [16,33] through their model differs from that by Cieplak *et al.* [32] but it still relies upon the enumerability of a subset of structures with a reasonably small number of disordered loops. In this method, the protein is regarded as a chain of U links and each unfolding step is the removal of one chain link. An unfolding pathway can therefore be represented as $P = (S_0 \rightarrow S_1 \rightarrow \cdots \rightarrow S_U)$ where S_0 is the native state, S_U is the fully unfolded state and S_n in general is a microstate with n disordered and $U-n$ ordered loops. The transition state is identified by the maximum of the free energy along the pathway: $F_p^{\dagger} = \max\{F(S_0), F(S_1), \cdots, F(S_U)\}$. The number of unfolding pathways is extremely high since for any number n of disordered loops, there are many possible microstates S_n, and each of them may be crossed by many different pathways. The most efficient unfolding pathway connects the native to the unfolded state through the lowest free-energy barrier:

$$F_p^{\dagger} = \min\{\max\{F(S_0), F(S_1), \cdots, F(S_U)\}\}.$$

This special saddle point that identifies the folding nucleus was found through a systematic exploration of all unfolding pathways by means of a recursive algorithm similar to that of dynamic programming. Since the protein does not necessarily unfold through Transition States of minimal free energy, the strategy suggested in Refs. [16,33] is to identify other not-optimal, but possibly numerous, passages over the free energy barrier. For every microstate S, all unfolding pathways passing through S are considered. If the free energy $F(S)$ is equal to the lowest free energy barrier of all pathways passing through S, then conformation S belongs to the Transition State Ensemble.

As already stated, an interesting feature of binary-models is that they are amenable to analytical or semi-analytical treatments. For instance, Bruscolini *et al*. [34] derived a mean field approach (MFA) for GF model of the Pin1 WW domain. In its variational formulation [35], MFA, for a system of Hamiltonian H and corresponding free-energy F, amounts to minimizing

$$F \le F_0 + \langle H - H_0 \rangle_0 = F_{\text{var}},$$

where H_0 is a solvable trial Hamiltonian and F_0 is the corresponding free-energy, both depending on the variational parameters $\mathbf{x} = \{x_1, \ldots, x_L\}$. Minimization leads to the self consistent equations

$$\left\langle \frac{\partial H_0}{\partial x_l} \right\rangle_0 \langle H - H_0 \rangle_0 = \left\langle (H - H_0) \frac{\partial H_0}{\partial x_l} \right\rangle_0.$$

The standard MFA employs as trial Hamiltonian:

$$H_0(\mathbf{s}) = \sum_{i=1}^{L} x_i s_i \qquad (5)$$

with x_i to be determined by minimizing the variational free-energy [35]

$$F_{\text{var}}(\mathbf{x}, T) = \sum_{i=1}^{L} f_0(x_i, T) + \langle H - H_0 \rangle_0$$

where the first term is the free energy associated to H_0,

$$f_0(x_i, T) = -RT \ln\{1 + \exp(-x_i / RT)\}$$

Thermal averages, performed through the Hamiltonian H_0 factorize $\langle s_i s_j \ldots s_k \rangle_0 = \langle s_i \rangle_0 \langle s_j \rangle_0 \ldots \langle s_k \rangle_0$. The approximate average "site magnetization" $m_i = \langle s_i \rangle_0$ depends only on the field x_i, and is given by

$$m_i = \frac{\partial f_0(x_i, T)}{\partial x_i} = \frac{1}{1 + \exp(x_i / RT)}. \qquad (6)$$

Instead of working with external fields x_i's, it is more intuitive to use the corresponding "magnetizations" m_i's, writing $F_{\rm var}$ as a function of the m_i's. Due to the choice of H_0 [see Eq.(5)] and to expression (6), evaluating the thermal average $\langle H\rangle_0$ amounts to replacing, in the Hamiltonian (1), each variable s_i by its thermal average m_i. Finally one obtains [36]

$$F_{\rm var}(\mathbf{m},T) = \varepsilon\sum_{ij}\Delta_{ij}m_i m_j - TS(\mathbf{m}) + RT\sum_{i=1}^{L} g(m_i)$$

where $g(u) = u\ln(u) + (1-u)\ln(1-u)$.

The MFA and its improvements [34] have been applied to the study of the folding behavior of the β-hairpin fragment from the Immunoglobulin-binding protein (GB1).

The reaction coordinate characterizing the folding is the average magnetization

$$Q = \frac{1}{L}\sum_{i=1}^{L}\langle s_i\rangle \qquad (7)$$

representing th e fraction of native-like residues (s_i =1). In MFA, this recasts to the quantity Q =$\Sigma_i m_i/L$. Fig. (**1**) shows the thermal behavior of Q as computed from exact enumeration simulations and its MFA estimates. The inset reports the hydrophobic cluster (W43−Y 45−F52−V54) population Q_{hyd} as a function of temperature (experimental data from [37]) and its fitting provided by the model. As shown in Fig.(**1**), the experimental data are well reproduced by both exact enumeration and MFA.

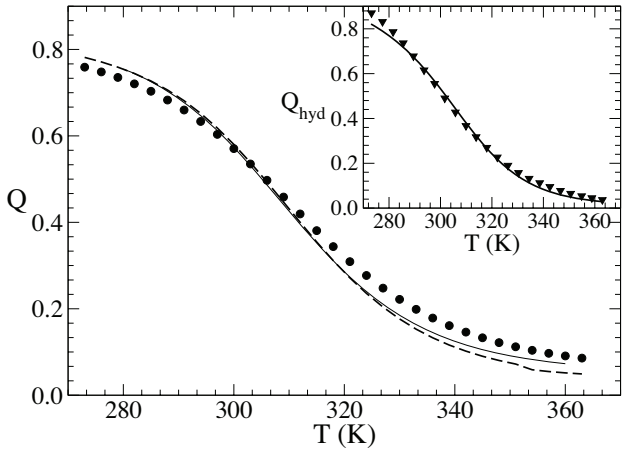

Fig. (1). Fraction of native residues Q (see Eq.(7)) during thermal folding, according to the GF model. Full dots are the exact result obtained by exhaustive enumeration. Dashes and solid lines indicate improved approximations MFA1 and MFA3 respectively (cfr. Ref. [36] for details). Inset: Fit of the hydrophobic cluster (W43−Y45−F52−V54) population Q_{hyd} (solid) to the experimental data from [37] (triangles).

Due to their minimalist character, Ising-like models yield predictions that are often in qualitative but not quantitative agreement with experiments, e.g in the case of Φ-values. Many efforts are currently made [38] for increasing the level of detail and thus the accuracy of Ising-like models. A possibility is to incorporate non-native interactions by increasing the number of configurations for each residue or the letters in the amino-acid alphabet [39].

The Fold-X force-field developed by Serrano and coworkers [38] has the same structure as the GF model [16], but the interaction and entropy terms were derived by means of a statistical analysis of the protein database and are thus more reliable. The free energy function, evaluated on the native structure, reads:

$F = (ACP + HB) + n \times Ent + 2.1\ RT\ \ln(L/0.4)$.

The term *ACP* is a potential for atom-atom and atom-solvent contacts, *HB* is the hydrogen bond term, *Ent* is the local entropy per residue and n is the number of residues. The term $\Delta S_{\rm loop} = -2.1R\ \ln(L/0.4)$ represents the entropy of a disordered loop of length L connecting two native-like segments and it is estimated by fitting the experimental entropy data of three proteins, ROP, SH3 and CI2 [38].

Despite the simplicity, Ising-like models are suitable to the study of proteins with a quasi-linear organization as *repeat proteins* whose sequence composed by tandem repeats of short aminoacid stretches leads to elongated super-helical structures (see Ferreiro *et al*. [40]). Using this approach, Ref. [40] showed that, in repeat proteins, the coupling between stability and cooperativity stems from their common dependence on the inter-repeat interaction energy. Thus assuming, that mutations affect this specific energy contribution, it is possible to infer their experimentally observed simultaneous effects on both stability and cooperativity.

SINGLE BEAD MODELS

HP Model

The first bead models of proteins were introduced by the pioneering work of Dill and coworkers between the end of the 80s and the early 90s [41,42]. The protein is modeled as a linear chain of beads linked by virtual bonds of constant length. In the original version, each conformation was represented as a self-avoiding walk on a two-dimensional square lattice [41,42], (Fig. (**2**)), later versions [43] involved a cubic three-dimensional lattice. The lattice allows a discretization of the conformational space and enforces excluded volume effects.

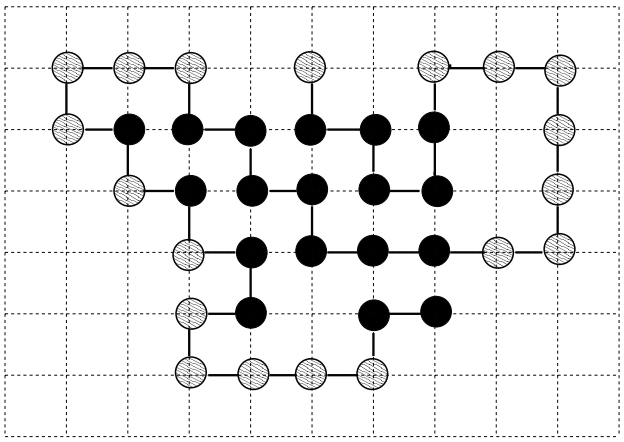

Fig. (2). HP-model in 2-dimension. In this lattice model, the polymer is constrained on a cubic lattice. There are two kinds of beads: hydrophobic (black) and hydrophilic (white). The cartoon shows the propensity of hydrophobic beads to segregate in a hydrophobic core, whereas the hydrophilic ones, are exposed on the surface.

The most important difference between 2D and 3D systems is the lattice coordination number z, *i.e*. the number of nearest neighbors of each site, which is *4* in the square and *6* in the cubic lattice. As a consequence, the cubic lattice allows for more orientations per bond ($z-1=5$ vs $z-1=3$ in the square lattice) and more nearest-neighbor potential binding partners. The 2D approximation allows a significant reduction of computation times, however, it is less limiting than it may seem, as the surface-to-volume ratio of long chains in 3D is the same as that of shorter chains in 2D.

According to the model, just two types of residues compose proteins: hydrophobic (H) and polar (P). To account for the hydrophobic effect, hydrophobic residues interact attractively with each other by negative contact energy $\varepsilon_{HH} < 0$. The other possible interactions do not contribute to stability, $\varepsilon_{HP} = \varepsilon_{PP} = 0$. The extreme simplification reflects in the limited size of both sequence and conformational space that are composed by 2^L and $(z-1)^{L-1}$ elements respectively, for an L-residue protein. As a result, the sequence and conformation space of short peptides can be explored by exhaustive enumeration, whereas for longer sequences the Monte Carlo technique is a more effective choice. Despite the crude approximations, the model reproduces a number of protein-like features. First of all, in the neighborhood of the folding temperature, a high percentage of sequences shows a sharp transition from the unfolded ensemble to the native one [41,42,44]. The native ensemble appears to be composed of compact conformations with a hydrophobic core and, for many sequences, it contains just one or a few structures. Furthermore, 2D compact conformations show the same distribution of secondary structure elements as real proteins [45,46]. The mutational properties of HP sequences are also protein-like, with a large majority of neutral mutations and many instances of sequences attaining the same native structure [42,47].

The model allowed also to clarify and illustrate the different scenarios associated to the energy landscape theory, where the static concept of *folding pathways* is replaced by the statistical notion of energy landscapes and folding funnels.

Stillinger Model

The model proposed by Stillinger [48] can be regarded as an off-lattice evolution of the HP model [41,42]. The protein is still described as a chain of ***L*** beads, hydrophobic and polar such that the sequence can be encoded by a set of binary variables ξ_i, where $\xi = 1$ for hydrophobic and -1 for polar residues. However, the chain lives in a continuous 2D space and each conformation is described by a set of $L-2$ bond angles between pairs of consecutive bond vectors of fixed unit length, (Fig. (**3**)). The potential energy is expressed as follows:

$$V = \sum_{i=2}^{L-1} V_c(\theta_i) + \sum_{i=1}^{L-2} \sum_{j=i+2}^{L} V_{nb}(r_{ij}, \xi_i, \xi_j)$$

Where

$$V_c(\theta_i) = \frac{1}{4}(1 - \cos\theta_i)$$

$$V_{nb}(r_{ij}, \xi_i, \xi_j) = 4\left[\frac{1}{r_{ij}^{12}} - \frac{C(\xi_i, \xi_j)}{r_{ij}^{6}}\right]$$

The term $V_c(\theta_i)$ is related to the curvature of the protein and it favors extended conformations with $\theta_i = 0$, $\forall i = 2,.., L-1$. Conversely, the term $V_{nb}(r_{ij}, \varepsilon_i, \varepsilon_j)$ promoting compact conformations, is the Lennard-Jones potential for non-bonded interactions which depends on the chemical features of the interacting beads through the $C(\varepsilon_i, \varepsilon_j)$ parameter whose values are +1 for HH pairs (strong attraction), +1/2 for PP pairs (weak attraction) and −1 for HP pairs (strong repulsion). The strong attraction between H pairs warrants the propensity of the model to form a hydrophobic nucleus as sketched in Fig. (**3**).

Stillinger model has been widely applied by several authors. For instance, Irbäck *et al.* studied the low-temperature behavior of chains of length 8 and 10 [49]. However, only a minority of sequences, with a very large energy gap between the native state and the lowest-energy decoys exhibit a single folded at the chosen temperature.

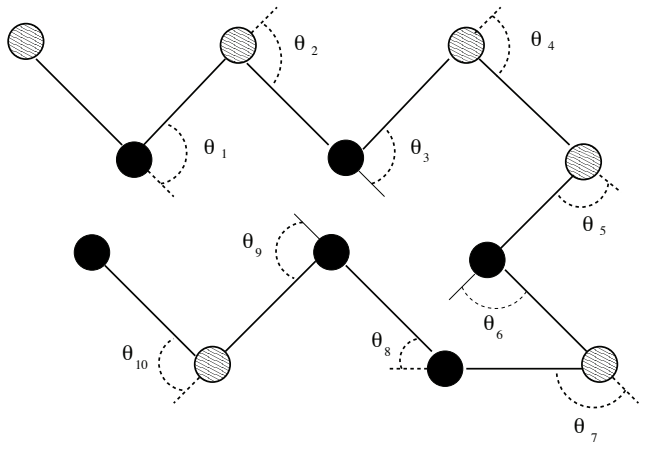

Fig. (3). Chain representation in the off-lattice Stillinger model. Bond vectors of fixed length point from one bead to the next one. The angle between two successive bond vectors is the bending angle θ. Similarly to the HP-model, residues can be either hydrophobic (black) or hydrophilic (white). The attractive interactions between hydrophobic residues generate a hydrophobic core.

In Ref. [50], the Stillinger model was employed to address the problem of the characterization of good and bad folders. These authors mapped the energy landscape in the neighborhood of the native conformation, drawing a graph that shows the connections of the native structure with its first neighboring minima. In the case of bad folders, there are only a few connections between the native state and the neighboring local minima. Thus a protein starting from a random extended conformation is unlikely to find a pathway towards the native conformation. In the case of good folders several routes do exist towards the native valley. Another work [51] described the folding and unfolding dynamics as an activated process whereby the protein jumps from a local minimum to a neighboring one. The transition probabilities were computed using Langer's theory and identifying saddle points via an over-damped dynamics. This computational approach, requiring the identification of a large number of minima and saddles is actually computationally very demanding. However, Livi *et al.* [52] suggested to reduce the number of saddles to be searched by using a bond-angle distance as a metrics in conformation space to identify directly connected pairs of minima. The authors showed that the val-

ues taken by the angular distance between minima of directly connected pairs (pairs of minima separated by one saddle, DCP) are confined to the small value tail of the probability density function of the distance between any pair of minima. As a consequence, the search of DCP is limited to the subset of pairs of minima whose distance is smaller than a threshold value. Since the application of this method could discard DCP belonging to the large value tail of the distance distribution, and since the inspection of the energy landscape showed that the DCP closer to the native structure exhibit the largest separation, the metric criterion was complemented with a systematic search of the DCP belonging to the native valley. This approach provides a general effective strategy also for reconstructing the energy landscape of more realistic models.

Kolinski-Skolnick and Honeycutt Thirumalai Models

It is convenient to describe the Kolinski-Skolnick (KS) [53] and Honeycutt-Thirumalai (HT) [54] models jointly, as the former actually inspired the latter. Both models extend the aminoacid alphabet, beyond hydrophobic (B) and hydrophilic (L) residues, to include also neutral instances (N) providing a more realistic description of sequences than HP [41,42] and Stillinger [48] models.

In the KS model, in particular, the protein lives on a diamond lattice and for each internal dihedral angle (between the two planes identified by 4 consecutive beads) only three discrete conformations are allowed: the *trans* one and the two *gauche* ones. Long-range interactions between hydrophobic residues are rewarded with an $\varepsilon_h < 0$ energy stabilization, whereas interactions involving hydrophilic residues are repulsive ($\varepsilon_W > 0$). Neutral residues thus do not appear in the non-bonded term of KS model. Interestingly, the model also includes a cooperativity term, which tries to account for experimental patterns such as the cooperativity in the formation of α-helices, and, more generally the formation of hydrogen-bond networks. In the KS implementation, two residues *i* and *j* can receive an extra-stabilization depending on the conformation of the neighboring dihedral angles. The energy function of KS model reads:

$$V_{tot} = \sum_{i=2}^{L-2}(1-\delta_{ti}) - \frac{1}{2}\sum_{i=1}^{L}\sum_{|i-j|>1}^{L}\varepsilon_{ij}\Theta[r_{ij}(r_c - r_{ij})]$$
$$-\frac{\varepsilon_c}{2}\sum_{i=2}^{L-1}\sum_{|i-j|>1}^{L}\Theta[r_{ij}(r_c - r_{ij})]\left(\delta_{t,i-1}+\delta_{t,i}\right)\left(\delta_{t,j-1}+\delta_{t,j}\right)$$

where, Θ(s) is the unitary step function, r_c is a distance cut-off, δ_{ti} =1 if the i-th dihedral is in the trans state and zero otherwise; $\varepsilon_{ij} = \varepsilon_h$ if *i* and *j* are both hydrophobic and $\varepsilon_{ij} = \varepsilon_W$, if at least one of the two residues is hydrophilic; finally ε_c sets the energy scale for cooperative effects.

The off-lattice model developed by Honeycutt and Thirumalai [54] can be somehow regarded as the continuous version of KS. Accordingly, the neutral residues (N) mark the bend regions that are necessary to the formation of β-hairpins and related structures. The special role of neutral residues is due to their weak repulsion and weak dihedral forces in neutral stretches. This can be better understood from the examination of the potential energy function

$$V_{tot} = \sum_{k=1}^{L-2}\frac{k_\theta}{2}(\theta_k - \theta_0)^2 + \tag{8}$$
$$\sum_{k=1}^{L-3}[A(1+\cos\phi_k) + B(1+\cos 3\phi_k)] +$$
$$\sum_{i,j\geq i+3}^{L} 4\varepsilon_h S_1\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - S_2\left(\frac{\sigma}{r_{ij}}\right)^{6}\right]$$

The first term in this Hamiltonian is just a harmonic potential allowing only small oscillations of bending angles around their equilibrium value. The third term of the force-field, where ε_h sets the energy scale, is a non-bonded interaction potential that mimics the hydrophobic effect. In fact, the S_1 and S_2 coefficients take on different values according to the nature of the interacting residues: $S_1 = S_2 = 1$ for BB pairs, $S_1 = 2/3$ and $S_2 = -1$ for LL and LB interactions, and $S_1 = 1$, $S_2 = 0$ for all interactions involving N residues. As a consequence, hydrophobic residues interact with each other through an attractive Lennard-Jones-like potential with an equilibrium distance $2^{1/6}\sigma$ where σ is just the virtual bond length between successive residues. On the other hand, LL and LB interactions are long-range and repulsive, while all interactions involving neutral residues are of the excluded-volume type.

The dihedral potential is characterized by three minima corresponding to trans conformation (φ = 0) and two gauche conformations ($\phi = \pm arcos\ [(3B-A)/(12B)]^{1/2}$). In the strand regions the coefficients *A* and *B* are chosen so to favor the trans conformation, whereas in the bend regions the coefficients make the three minima degenerate and separated only by a modest energy barrier. This ensures maximal flexibility to the expected bend regions. Notice that such a dihedral potential is the off-lattice variant of KS dihedral interaction. Moreover HT approach does not include the cooperativity term that was a peculiarity of KS model.

Thirumalai and coworkers used the model to perform several Molecular Dynamics simulations on a 46-residue chain [54,55] that had been designed by Skolnick *et al*. to fold into a 4-stranded β-barrel [53]. In particular, a series of temperature-jump simulations showed the existence of a three-stage kinetics:

Denatured → Compact → Native-like → Native

The folding process therefore first involves a collapse of the protein into a compact conformation that later acquires native-like elements and finally evolves to the native state through activated transitions. The last stage of this folding process is the signature of an energy landscape whose low-energy region is extremely rugged and degenerate. Actually, based on the features of his model, Thirumalai formulated the so called Metastability hypothesis [54] according to which a polypeptide chain can exist in a number of structurally similar but energetically different conformations depending on initial conditions. These features of the landscape, however, are not typical of natural proteins, but are rather reminiscent of glass systems. The energy landscape of HT model has been analyzed by Guo and Brooks [56], Nymeyer, Garcia and Onuchic [57] and Miller and Wales [58]. All these studies pointed out that the landscape contains many traps and long-lived intermediates favoring collapsed states rather than the single native structure.

Sorenson-Head Gordon Model (SHG)

An improvement of HT-model was realized by Head-Gordon and coworkers [59] by changing the dihedral term in potential (8)

$$V_{\phi} = \sum_{k=1}^{L-3}\{A(1+\cos\phi_k) + B(1-\cos\phi_k) + C(1+\cos 3\phi_k) + D[1+\cos(\phi+\pi/4)]\} \quad (9)$$

to describe mixed alpha beta protein structures. Therefore three types of dihedral angles do exist: helical (H), extended (E) and turn (T), corresponding to three sets of dihedral coefficients *A, B, C, D*. It follows that the SHG model requires as input, not only the aminoacid sequence in the three-letter alphabet B, L, N, but also the secondary structure encoded by the E, H and T alphabet. The implementation of this model to realistic cases requires a sequence design technique based on energy gap maximization. The procedure implies first the selection of the native state and a set of low-energy misfolded conformations (called decoys). A search is then performed in sequence space until the mutant maximizes the energy gap between the native state and the next low-lying state. This strategy was tested for the first time on the 46-mer sequence introduced by Skolnick [53]. The designed sequence exhibited a lower collapse temperature and a higher folding temperature resulting in a faster and more cooperative folding transition. Using a more difficult benchmark, the design protocol was then applied to produce sequences that correctly discriminate the differences in folding mechanism experimentally detected for protein-L and protein-G [60]. The model was then improved in 2008 [61], with the introduction of an orientation-dependent hydrogen bonding term in the potential energy function and with the increase of the number of bead flavors from three to four. The new model is also characterized by an increased dependence on the secondary structural information that now affects not only the dihedral but also the bending and hydrogen bonding terms. Moreover, the sets of dihedral parameters have increased from 3 to 6 and four different types of turns are accounted for. This new version retains the ability of the original model [59] to discriminate the folding pathways of protein L and protein G, but it shows greater cooperativity as well as a more funnel-like landscape, allowing the folding temperature to be well above the glass transition temperature.

An Example of off-Lattice Go-Model

Comparative analysis aimed at assessing the folding features of the Head-Gordon model [59] were performed by Cecconi *et al.* [62] using as a benchmark the WW-domain of hPin1 protein whose sequence had already been optimized in a preliminary study by Head-Gordon and coworkers [60]. Ref. [62] compared the performance of the Sorenson Head-Gordon model with that of the Go-like force field proposed by Clementi *et al.* [63]. The energy function of the latter model reads:

$$V_{tot} = \sum_{i=1}^{L-1}\frac{k_h}{2}(r_{i,i+1}-R_{i,i+1})^2 + \sum_{i=1}^{L-2}\frac{k_\theta}{2}(\theta_i-\theta_i^0)^2 + \sum_{i=1}^{L-3}k_\phi^{(1)}[1-\cos(\phi_i-\phi_i^0)] + k_\phi^{(3)}[1-\cos 3(\phi_i-\phi_i^0)] + \sum_{i,j>i+3}^{L}V_{nb}(r_{ij}) \quad (10)$$

where the non bonded interaction potential V_{nb} reads

$$V_{nb}(r_{ij}) = \begin{cases} \varepsilon_{ij}\left[5\left(\frac{R_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{R_{ij}}{r_{ij}}\right)^{10}\right] & i-j \text{ native} \\ \frac{10\varepsilon_r}{3}\left(\frac{\sigma}{r_{ij}}\right)^{12} & i-j \text{ not native} \end{cases}$$

In the above equations r_{ij} is the distance between residue i and j, θ_i is the bending angle identified by the three consecutive Cα's $i-1$, i, $i+1$, ϕ_i is the dihedral angle defined by the two planes formed by four consecutive Cα's $i-2$, $i-1$, i, $i+1$. The symbols with the superscript "0" and R_{ij} are the corresponding quantities in the native conformation. The first two terms of this force-field are harmonic stretching (chain connectivity) and bending potentials. The third term is a dihedral potential introducing a bias towards native secondary structure elements. The non-bonded 12-10 Lennard-Jones potential assigns attractive interactions to residues forming contacts in the native structure. Residues that do not form native contacts repel each other through an excluded-volume potential. In its first application [63], the force-field by Clementi *et al.* proved to correctly identify the transition state ensemble of CI2 and SH3, two small globular proteins folding as two-state folders. The force-field was also able to identify the folding intermediate of barnase, RNAase H and CheY proteins.

In some cases (i.e. small peptides, Ref. [62]), to attain the typical cooperative pattern of two-state folders, the cooperativity of the Clementi *et al.* model can be enhanced by the rescaling technique proposed by Chan [64]. Experimental studies suggest [65,66] that the origin of cooperativity lies in specific interactions appearing only after the assembly of native-like structures. The extra stabilization that the protein receives upon entering the native basin, can be modeled by rescaling the interaction forces according to

$$\mathbf{F}_{conf} = \begin{cases} -\nabla V_{tot} & Q < Q_{th} \\ -\nabla V_h - \rho\nabla(V_{tot}-V_h) & Q \geq Q_{th} \end{cases}$$

where V_h is the stretching potential, Q is the fraction of formed native contacts and $\rho = 2$ is the scaling factor.

The simulations in Ref. [62] showed that potential (10) with rescaling could correctly reproduce the reversible, cooperative, two-state mechanism of folding of hPin1 WW domain. In particular, the cooperativity is indicated by the single narrow peak in the specific heat plot and by a ratio of the van't Hoff to calorimetric enthalpy close to 1. Conversely, the simulation results by the SHG model [59] were rather ambiguous. The thermograms in Fig. (**4**), feature not only a peak, but also a shoulder at lower temperature. This is a signature of a non-cooperative folding involving a collapse into a compact, only partially structured, globule, followed by a rearrangement into the native conformation.

More interestingly, Cecconi and coworkers showed that the native state conformations clustered in two main subsets characterized by non-overlapping distributions of RMSD from the lowest energy, reference conformation. The structures of the two subsets were similar in energy but showed opposite chirality. This situation is an indicator of a partitioning of the native basin that can be visualized in Fig. (**5**) by plotting the potential of mean force as a function of

RMSD. The existence of two distinct clusters of native-like conformations suggests that the SHG model [59] still retains the roughness and degeneration of the energy landscape of HT model from which it was derived. According to Cecconi *et al.*, a possible reason for the degeneracy of the native state relies on the symmetry of the dihedral potential $V(\phi)$, Eq.(9). In fact, the secondary structure of hPin1 WW domain only contains Extended and Turn dihedrals so that $V(\phi)$ is a polynomial in $cos(\phi)$ and it is symmetric for inversion $\phi \rightarrow -\phi$ assigning equal stability to structures with opposite chirality.

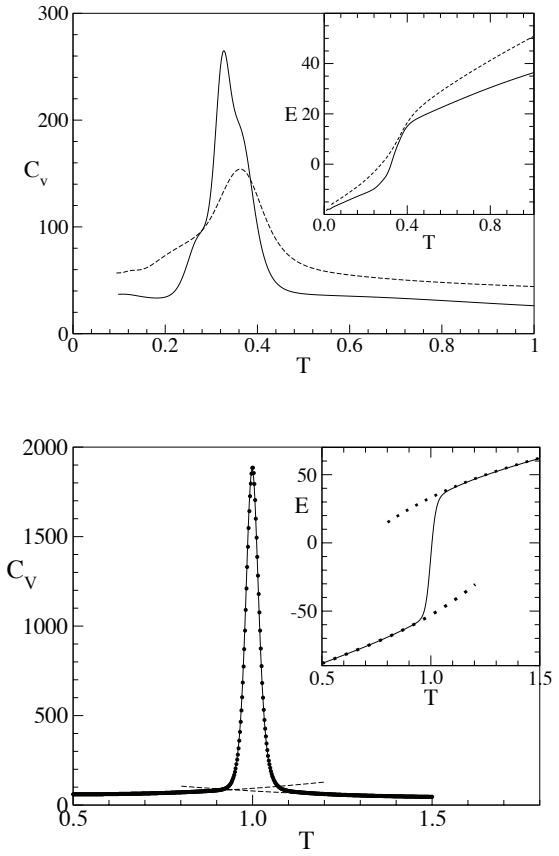

Fig. (4). Comparison of Go and Sorenson-Head-Gordon models. Top panel: discrepancy of the specific heat versus temperature profiles of the folding (solid line) and unfolding (dashed) simulations with the SHG model. A similar mismatch also applies to the thermal behavior of the energy as reported in the inset. Bottom panel: Go-model simulations yield perfect superposition of the specific heat and energy (inset) profiles of folding (solid line) and unfolding (diamonds).

Moreover, a significant overlap has been observed between the energy histograms of native and unfolded state ensembles, suggesting only a partial maximization of the energy gap between the native conformation and the lowest-energy decoy.

Classical Go-models [see Eq.(10)] are completely based on the topology of the native state. They may not be fully adequate when the chemistry of the polypeptide sequence plays a more relevant role than native state topology in stirring the folding process. Several improvements were introduced in the basic Go model to include some chemical and physical information on residues. The first attempt is to take into account the effects of side chain hindrance by the heavy-map approach. In the original model by Clementi *et al.*, two residues are regarded to be in native contact if the distance between their Cα atoms is below a given threshold. This, however, may lead to the erroneous conclusion that two residues with large side-chains, such as Glutamate and Lysine forming a salt bridge, are not in contact because their Cα are very far from each other, while side-chain atoms are actually very close. This difficulty is readily overcome by considering the heavy-map approach where two residues are in native contact whenever at least a pair of side-chain heavy-atoms are within a distance cutoff.

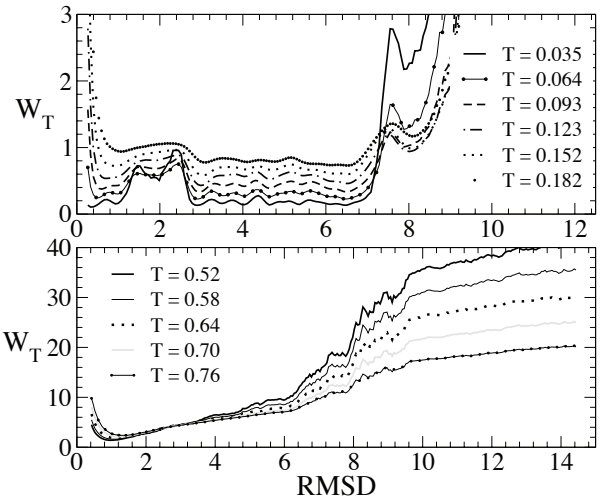

Fig. (5). Upper panel: low-temperature free energy profiles of the SHG model as a function of the RMSD from the reference (native) conformation. The native valley appears to be partitioned in two main sub-basins separated by a barrier. The sub-basin corresponding to RMSD range [0.25-1.00] is populated by conformations with the same chirality as the PDB structure, whereas the sub-valley in the range [2.80-6.50], corresponds to the opposite chirality. Lower panel: low-temperature free energy profiles of the Go-model as a function of the RMSD from the native conformation (pdb-id=1NMV). The native valley shows a single basin as opposed to the partitioning in two sub-valleys typical of the Go-model. For further details see [62].

Moreover, Guardiani *et al.* [67-69], in order to perform a mutational analysis of domain C5 of Myosin Binding Protein C (MyBPC), found more appropriate to use heterogeneous energy couplings. In particular, contact energies were rescaled according to the number of atomic contacts using the equation:

$$\varepsilon_{ij} = \varepsilon_0 \left(1 + \frac{N_{ij}}{N_{\max}} \right)$$

where N_{ij} is the number of atomic contacts between residues i and j while N_{max} is the maximum of N_{ij} over all pairs of residues in native contact of the protein under examination. The use of heterogeneous couplings is a key improvement allowing the Go-model to successfully deal with extremely difficult benchmark proteins such as those with similar topologies but different folding mechanisms. For instance, this is the case of the B1 segments of pepto-streptococcal proteins

G and L. These share an identical fold with a central α-helix packed against a four-stranded β-sheet composed by two hairpins. Experimental evidences [70,71] show that in protein L the N-terminal hairpins form ahead of the C-terminal one, whereas in protein G the folding order of the two hairpins is reversed. Classical Go models such as the one used by Koga and Takada [25] failed to discriminate the folding mechanisms of these two proteins. The task, instead, was successfully accomplished by the models introduced by Karanicolas and Brooks [72] and Sutto *et al.* [73] that included sequence effects via heterogeneous couplings. The model by Karanicolas and Brooks will be discussed in more detail later on in this section. As a further example, we can mention a work by Matysiak *et al.* [74] where a heterogeneous Go model reproduced available experimental data on free energy differences upon single mutations of S6 ribosomal protein and its circular permutants. Energetic heterogeneity has been crucial also to discriminate pathogenic mutations on domain C5 of MyBPC (Guardiani *et al*. [67-69]). In fact, within the framework of the Go approach, a mutation is modeled by removing all the native contacts involving the mutated residue. However, if two residues form the same number of residue-residue contacts, the role of their mutation can be discriminated only by weighting the contact energies via the number of atomic contacts. Fig. (**6**), shows the sensitivity of specific-heat plots of C5-domain from MyBPC to mutations on Arg14, Arg28 and Asn115 and to deletion of first seven residues, Δ 1-7.

This sensitivity is enhanced by the use of heavy-map Go-model with heterogeneous couplings. Another limit of the basic Go-models is that they do not account for desolvation effects. For instance, if we consider the Go-model described by Eq. (10), the solvent is implicitly simulated through a careful tuning of the ε_{ij} parameters. The influence of the solvent on kinetics is also captured performing Molecular Dynamics with the Langevin equation:

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i - \gamma \frac{d\mathbf{r}_i}{dt} + \mathbf{R}_i$$

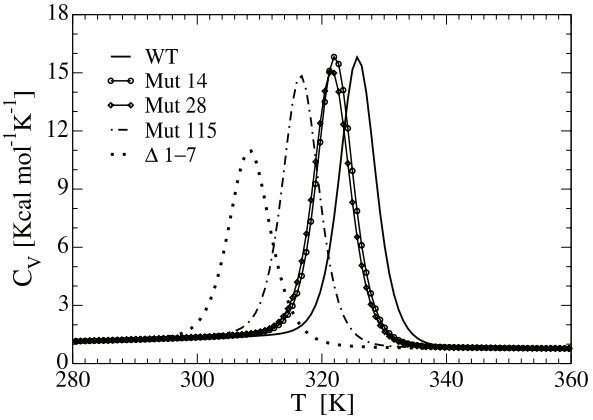

Fig. (6). Effect of pathogenic mutations on C5 domain from MyBPC: the thermal plot of the specific heat shifts toward lower and lower temperatures in agreement with the clinical severity of the mutation. Heat capacity of the Wild Type C5-domain (WT) is compared to those of the missense mutants deprived of the native contacts of Arg14, Arg28 and Asn115 (Mut14, Mut28 and Mut115), as well as, to the one of a deletion mutant (Δ1−7). For further details see Ref. [67].

Accordingly, the force acting on residue *i* is a sum of three contributions: a conformational force $\mathbf{F}_i$ due to the interactions of residue *i* with all the other residues of the protein, a friction force $-\gamma d\mathbf{r}_i/dt$ due to the viscosity of the solvent, and a random force $\mathbf{R}_i$ modeling the random collisions of solvent molecules on the residue. This approach, however, does not consider the particle nature of water, responsible for the desolvation effect that needs appropriate treatment as discussed in the following.

Go-Model with Desolvation

It is known [75] that the potential of mean force modeling the interaction between two methane-like molecules in water exhibits two minima, (Fig. (**7**)): the first minimum corresponds to the interaction of the two particles at a distance equal to the sum of their van der Waals radii; the second minimum refers to the two methane particles separated by a single water molecule. The energy barrier between these two minima corresponds to high energy arrangements where the water molecule has been expelled but the methane particles are not yet close enough to strongly interact with each other. Accordingly, a free energy penalty must be associated with the desolvation of the hydrophobic core of a protein.

An improved Go-model with desolvation effects has been developed by Cheung *et al*. [76]. In this model the native contact LJ potential is corrected such that two minima appear. With reference to Fig. (**7**), the first one corresponds to two residues in contact (at equilibrium distance $r = r'$), the second one refers to a state where two attracting beads are separated by a single water molecule (at equilibrium distance $r = r''$).

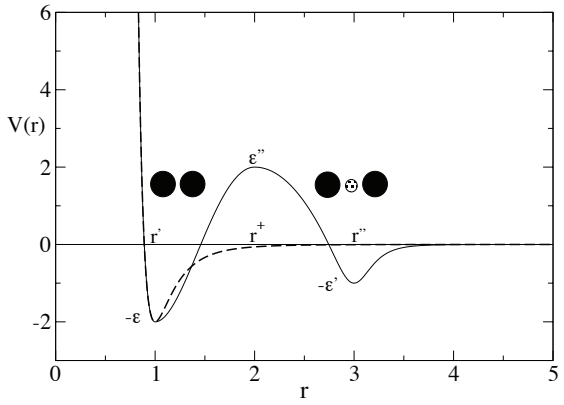

Fig. (7). Native-contact potential (reduced units) with desolvation correction as a function of inter-particle distance (Å). When two residues are at distance *r'* they are in direct contact, whereas at distance *r"*, they are separated by a single water molecule. The expulsion of this water molecule generates an energy barrier. (Adapted from Fig.1 of Ref. [76]).

Between these two minima there is a desolvation barrier located at $r = r^+$. The depth of the energy wells at *r'* and *r"* are labeled $-\varepsilon$ and $-\varepsilon'$ respectively, while ε'' is the height of the desolvation barrier. These parameters are related to each other in such a way that they can be easily derived from the corresponding Lennard-Jones parameters ε and *r'*. In fact, based on the findings reported by Hummer [77,78] and by Hillson [79], one has $(\varepsilon''-\varepsilon')/(\varepsilon'-\varepsilon)=1.33$ and $\varepsilon'/\varepsilon=1/3$. These expressions allow to derive both ε' and ε'' from ε.

Furthermore, as the width of the desolvation barrier corresponds to the diameter of a water molecule (~3Å), then $r'' = r'+3$Å. Finally, the barrier position is assumed to occur at $r^{\dagger} = (r'+ r'')/2$.

The model was applied to the folding of protein SH3 [76] characterized by a native state where a hydrophobic core is enclosed by β-sheets. The simulations showed that the collapse to a native-like structure is followed by a further step where water is expelled from the hydrophobic core. This pattern is supposed to be relevant for the biological activity of the protein. It is important to consider that the computation of energy as a sum of pairwise contributions, as in the Cheung *et al*. approach, is often criticized as being unsuitable to represent the many-body nature of the hydrophobic effect. For example, Czaplewski [75] showed that the three-body term accounts for 10% of the total hydrophobic association energy. Despite this limitation, the sharper profile of the specific heat of SH3 obtained in Ref. [76] as compared to standard Lennard-Jones simulations clearly suggests an improvement in the cooperativity of the folding process.

Karanicolas-Brooks Model

The sophisticated Go-model developed by Karanicolas and Brooks [72] (KB) includes all of the improvements discussed so far and it also accounts for non-native interactions responsible for the energetic frustration of real proteins. As in most Go-models, a stretching and a bending potential impose chain connectivity and elasticity respectively, while residues not forming native contacts repel each other through an excluded-volume potential.

The interaction term for residues forming native contacts has the following form:

$$V_{ij} = \varepsilon_{ij}\left[13\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

This potential differs from the 12-10 Lennard-Jones, for the presence of an additional repulsive term r^{-6}, which imposes a desolvation penalty that every pair of residues must pay before reaching the equilibrium distance. In KB model there are two types of native interactions: hydrogen bonds involving backbone atoms and contacts between side-chains. Hydrogen bonds are identified using the criterion of Kabsch and Sander [80], based on an electrostatic energy threshold. For any pair of hydrogen-bonded residues, ε_{ij} is set to unity and σ_{ij} to the native distance of the Cα's. In the same spirit of the model by Kolinski and Skolnick [53] described above, the cooperativity effects of hydrogen bonds are also accounted for. In particular, when residues *i* and *j* interact via two hydrogen bonds or a hydrogen bond and a side chain contact, the energy of one of the hydrogen bonds is not applied to the original (*i*,*j*) couple only, but it is distributed among four hydrogen bonds formed by *i* and *j* with the neighboring residues, according to the scheme sketched in Fig. (**8**). More explicitly, hydrogen bonds are also created for the pairs (*i*,*j*−*1*), (*i*,*j*+*1*), (*i*−*1*,*j*) and (*i*+*1*,*j*) with ε_{ij} set to *1/4* and σ_{ij} set to the α-carbon separation of each pair in the native state. The side-chain-side-chain interactions reflect the different chemical and physical properties of the 20 natural amino-acids and the ε_{ij}'s are thus scaled in proportion to the contact energies reported by Miyazawa and Jernigan [81]. The model thus implements heterogeneous energy couplings similar to the model by Guardiani *et al.*, discussed above [67-69]. Another interesting feature of the KB model is the use of a dihedral potential, that, unlike the one in Eq.(10), is not native-centric but sequence-based.

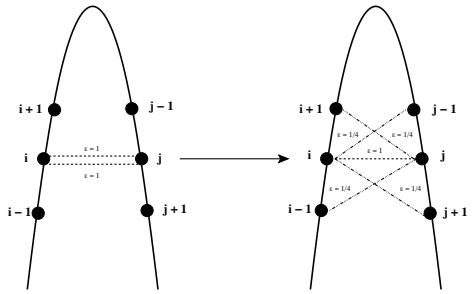

Fig. (8). Karanicolas-Brooks model: redistribution of contact energy among neighboring residues to enhance cooperativity. Whenever two residues i and j are linked by two hydrogen bonds or by a hydrogen bond and a side chain-side chain contact, the coupling parameter $\varepsilon = 1$ of one of these contacts is redistributed among the 4 pairs $(i, j-1)$, $(i, j+1)$, $(i-1, j)$ and $(i+1,j)$, each receiving an energy coefficient $\varepsilon = 1/4$.

This dihedral term introduces frustration in the energy funnel, somewhat reproducing the non-native interactions that influence the folding of natural proteins. The dihedral potential is computed as $V(\phi) \propto \ln P(\phi)$, where $P(\phi)$ is the distribution of the virtual dihedral angle ϕ derived from a survey of Protein Data Bank (PDB). The virtual dihedral angle $\phi = \angle(C_\alpha^{i-1}, C_\alpha^{i}, C_\alpha^{i+1}, C_\alpha^{i+2})$ only depends on the backbone dihedrals of the two middle residues, according to equation

$$\phi = \pi + \varphi_{i-1} + \psi_i + \frac{\pi}{9}(\sin\varphi_i + \sin\psi_{i+1}) \qquad (11)$$

(where $\varphi_i = \angle(C_{i-1}', N_i, C_i^{\alpha}, C_i')$ and $\psi_i = \angle(N_i, C_i^{\alpha}, C_i', N_{i+1})$. The authors considered 400 $P(\phi)$ distributions, for all of the possible amino-acid pairs. As already mentioned, the KB model was successful in discriminating the folding mechanisms of protein L and protein G despite their identical topology [72].

Statistical Potentials: Miyazawa-Jernigan

The development of the dihedral potential in the KB force-field [72] is an example of derivation of a *potential of mean force* exploiting the wealth of information contained in the PDB. Such a method is so general that it was widely used both in topological and sequence-based models. This is the reason why its description closes the section of one-bead models. This approach, in principle, is rather straightforward and is based on the inversion of the Boltzmann weight $P_{AB} \propto \exp(-E_{AB}/RT)$, ($R$ being the gas constant and T the absolute temperature). Accordingly, the potential of mean force between residues A and B located at distance $r \pm \Delta r$ is computed by Miyazawa-Jernigan [81] as

$$w_{AB}(r;\Delta r) = -RT\ln\frac{P_{AB}(r \pm \Delta r)}{P_{XX}(r \pm \Delta r)} \qquad (12)$$

where $P_{AB}(r \pm \Delta r)$ is the probability to find A, B at a distance $r \pm \Delta r$, $P_{XX}(r \pm \Delta r)$ is the corresponding probability in the reference state. With in the framework of the quasi-chemical approximation [82], the reference state is a random mixture where the number of contacts between species A and B is

proportional to their concentration. The potential of mean force developed in Ref. [83] accounts for desolvation effects through the reaction:

$$A \cdot S + B \cdot S \rightleftharpoons A \cdot B + S \cdot S$$

where *S* represents a solvent molecule and *A·S* and *B·S* indicate solvated state of the residues. An estimate of the contact energy of *AB* pair including also desolvation effects is thus:

$$e_{AB} = w_{AB} + w_{SS} - w_{AS} - w_{BS},$$

and requires to apply Eq.(12) four times.

The arbitrariness of w_{AB} stems from the choice of the way of characterizing the reference state (a non interacting mixture of aminoacids). Zhou *et al.* [84] argued that the quasi-chemical approximation deviates from a homogeneous mixture of aminoacids in the very common situation of unbalancing among attractive and repulsive pair interaction in the protein database. According to Ref. [84], this inaccuracy can be corrected by the extension of the ideal gas reference state to finite systems (Distance-scaled Finite Ideal-gas REference state, D.F.I.RE). This approach was applied to develop both an all-atom and a coarse grained statistical potential (including only backbone and C_β-atoms) that proved effective in the computation of the Z-score of 32 multiple decoy set.

It is interesting to notice that, even if the energy of a protein is computed as a sum of pairwise contributions, the potential of mean force accounts for multi-body effects as it automatically incorporates details of chemical neighborhood of the two interacting residues. This feature makes the mean-force potential extremely reliable when native-like conformations must be identified, as for instance in threading computations. For the same reason, the potential of mean force may be less accurate in protein folding simulations, where it is necessary to compute the energy not only of native-like but also of unfolded conformations. It has been noticed, however, that the introduction of a more accurate distance dependence may relieve this kind of shortcomings as testified by the results of Park and Levitt [85] and Wallqvist and Ullner [86]. As a final remark, it should be noticed that the potential of mean force allows the amino-acid alphabet to be extended from the simplified two- and three-letter formulations (e.g. HP-model [41], Stillinger [48] model and Thirumalai model [54]) to a more realistic 20-letter version. This is extremely important because the letter number has an influence on the heterogeneity of interactions, that, in turn, determines the entity of the energy gap [39].

The MJ potentials have been applied to a wide range of cases such as the evaluation of different sequences threaded onto known structures [87], the selection of native-like conformations from large sets of structures [88] and the assessment of the impact of amino-acid mutations on protein stability [89]. The MJ contact energies have also been employed in several simulation studies. Hinds and Levitt [90] used a dynamic programming algorithm for the exhaustive enumeration of protein conformations on a tetrahedral lattice. Covell, on the other hand, preferred to use a dynamic Monte-Carlo scheme with constraints on size, surface area and total number of contacts [91]. The work by Covell shows that even a simple lattice model with effective inter-residue contact energies may have predictive power. Finally, Kolinski and Skolnick [92] used a more detailed lattice representation and a more sophisticated force-field (including non-local terms and cooperativity) to simulate the folding of the B domain of staphylococcal protein A, ROP and Crambin.

TWO BEAD MODELS

In the previous section we discussed some approaches to protein modeling where aminoacids are assimilated to single interacting centers (single bead). In this section we briefly review two-bead models where each residue of a polypeptide chain is represented by two interaction centers, one representing the backbone part (the peptide bond), and the other one the side chain.

Levitt Model

One of the first models considering two reaction centers per residue was developed by Levitt [10]. The peptide group is simplified by combining the C', N and H atoms into an effective N' atom and replacing the O by an effective O' atom, see Fig. (**9**). The side chain is assimilated to a single effective "atom" located at the centroid of the side chain, it is thus completely rigid and sticks out from the backbone.

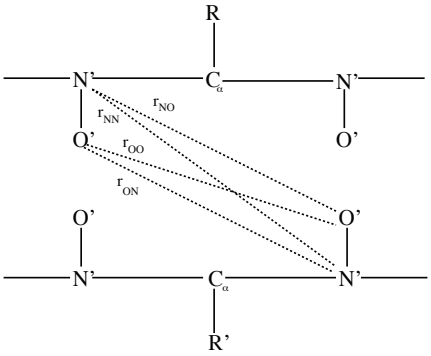

Fig. (9). Cartoon of mutual distances involved in the computation of hydrogen bond energy in Levitt model. In this model hydrogen bonds are only allowed between united peptide groups. They are modeled through a Lennard-Jones like term only involving the pairs of oppositely charged atoms *NO* (at distance R_{NO}) and *ON* (at distance R_{ON}), and by a Coulomb term applying to all possible couples (*NN, OO, NO, ON*).

In this model, bond lengths and bond angles are constant so they do not contribute to the energy function. Energy contributions due to interaction of groups of atoms are computed as effective potentials. The first term is the van der Waals interaction between side chains. All the atoms of a side chain, due to termal motion, are assumed to uniformly fluctuate onto a sphere centered on the side chain centroid and with radius equal to the average radius of gyration of the particular group. The effective potential between two identical side chains can then be calculated at various distances apart using:

$$V(\mathbf{R}) = \sum_{i>j} \int_{S_1} \int_{S_2} d^3u_i d^3u_j \left(\frac{A}{|\mathbf{R} - \mathbf{u}_i + \mathbf{u}_j|^9} - \frac{B}{|\mathbf{R} - \mathbf{u}_i + \mathbf{u}_j|^6} \right)$$

where the pairwise atomic van der Waals potential is integrated over positions, $\mathbf{r}_i = \mathbf{u}_i$, $\mathbf{r}_j = \mathbf{R} + \mathbf{u}_j$, of atom *i* anywhere in sphere S_1 and atom j anywhere in the sphere S_2, the inter-sphere distance is *R* and $d^3\mathbf{u}_i = dx_i dy_i dz_i$. The position and depth of the minimum of the plot of $V(|\mathbf{R}|)$ against $R=|\mathbf{R}|$ are

termed R^{O}_{ij} and ε_{ii} respectively. In the case of non identical side chains of type *i* and *j*, it is assumed that $\varepsilon_{ij} = (\varepsilon_i \varepsilon_j)^{1/2}$ and $r^{O}_{ij} = (r^{O}_{ii}\, r^{O}_{jj})^{1/2}$.

Another important contribution to the potential energy is the side-chain-solvent interaction characterized by the hydrophobic parameters, denoted by s_i, taken by the experimental free energy of transfer from water to ethanol [93]. Each parameter s_i can therefore be considered as the free energy change of a side chain that has lost all the water molecules of the hydration shell. This implies that s_i is negative for hydrophobic side chains and positive for hydrophilic ones; a side chain retaining all the water molecules is characterized by s_i=0. If residue *j* at distance r_{ij} from *i* produces a displacement of a fraction $g(r_{ij})$ of water molecules, the hydration energy becomes $s_i\, g(r_{ij})$. The function $g(r_{ij})$is approximated by a sigmoid:

$$g(r_{ij}) = \begin{cases} 1 - \frac{1}{2}(7x^2 - 9x^4 + 5x^6 - x^8) & \text{if } x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

where $x = r_{ij}/r_{max}$, with $r_{max} \sim 9$Å. It must be considered that the energy term related to loss of solvent due to the close approach of residues *i* and *j* is $(s_i + s_j)\, g(r_{ij})$, because in such a collision water is displaced from both *i* and *j*. In this model the dihedral angle between atoms $C^{\alpha}_{i-1}\ C^{\alpha}_{i}\ C^{\alpha}_{i+1}\ C^{\alpha}_{i+2}$, is roughly proportional to $\psi_i + \varphi_{i+1}$ according to formula (11). The effective torsion potential is expressed as a Fourier expansion of the form:

$$V(\phi) = k_a \sum_{k=1}^{6} \left\{ A_k \cos[(k-1)\phi] + B_k \sin[(k-1)\phi] \right\}$$

where k_a is the scale factor normally taken as 2. The coefficients A_k and B_k only depend on the chemical identity of the third residue defining the dihedral angle. In fact, $\phi_i \propto \psi_i + \varphi_{i+1}$, but as side chains have a greater influence on φ than on ψ angles [see Fig. (**10**)], the side chain of the third residue (the one linked to C^{α}_{i+1} affecting the φ_{i+1} dihedral) will determine most of the torsional energy of angle ϕ_i. The third amino acids can be divided into 3 classes according to their effect on ϕ -torsion energy: the first group includes Gly, Asp and Asn found to favor reverse turns; the second group includes only Pro; finally, the third group includes Ala and all the other aminoacids not belonging to the first two groups. Adding together the different energy contributions gives the complete energy function:

$$V_{tot} = \sum_{ij} \varepsilon_{ij} \left[3\left(\frac{r^{o}_{ij}}{r_{ij}}\right)^{8} - 4\left(\frac{r^{o}_{ij}}{r_{ij}}\right)^{6} \right] +$$

$$\sum_{ij} (s_i + s_j) g(r_{ij}) + \sum_{SS} K_{SS} (r^{SS}_{ij} - r^{SS}_{0})^2 +$$

$$\sum_{ij} \varepsilon_p \left[\left(\frac{r^{o}_{p}}{r_{NO}}\right)^{12} - 2\left(\frac{r^{o}_{p}}{r_{NO}}\right)^{6} + \left(\frac{r^{o}_{p}}{r_{ON}}\right)^{12} - 2\left(\frac{r^{o}_{p}}{r_{ON}}\right)^{6} \right] +$$

$$332 \sum_{ij} q^2_p \left(\frac{1}{r_{NN}} + \frac{1}{r_{OO}} + \frac{1}{r_{NO}} + \frac{1}{r_{ON}} \right) +$$

$$\sum_{i} \left\{ 2 \sum_{k=1}^{6} A^{(i)}_{k} \cos[(k-1)\phi_i] + B^{(i)}_{k} \cos[(k-1)\phi_i] \right\}$$

The third term accounts for disulfide bonds which must be known a priori and are given as elements of the primary structure. The disulfide bond harmonically oscillates around an equilibrium r^{SS}_{o} distance.

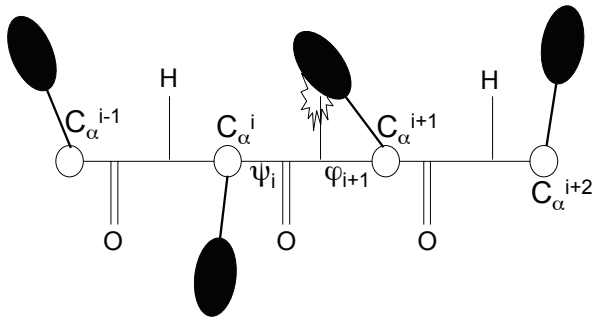

Fig. (10). Geometric constructon showing why the side chain linked to C^{α}_{i+1} mainly influence the diedral φ_i.

The fourth and fifth terms, together give the peptide hydrogen bonds. The q_p parameters are the partial charges on N'and O' united atoms, while ε_p and r_p represent the position and the depth of energy minimum of a 12-6 Lennard-Jones potential between two pairs of N' and O' atoms. The r_{NN}, r_{OO}, r_{NO} and r_{ON} are the distances of the pairs N ··· N, O ··· O, N ··· O and O ··· N respectively, according to Fig. (**9**).

This simplified model was extensively tested under a variety of different conditions. Bovine Pancreatic Trypsin Inhibitor (BPTI) was chosen as the benchmark protein because, in the early 70s, it was the only small protein of known conformation comprising a single subunit and no prosthetic group. The simulations starting either from a fully extended conformation or an extended conformation with a pre-formed C-terminal helix, were performed through alternating cycles of energy minimization and normal-mode thermalization. They showed that 70% of the runs with the pre-set α-helix ended up with a compact conformation having the size, shape and β-sheet structure of native BPTI. No near-native minimum closer than 2.5 Å RMSD to the native structure, however, could be found. Levitt verified that this limitation was due to the intrinsic lack of side-chain detail: the spherical shape of the side-chains, in fact, prevented them from reaching a compact packing arrangement. The runs starting from a fully extended chain are somewhat less successful in that the protein gets trapped in a metastable minimum and a pushing potential must be introduced to attain native-like conformations. This poor performance is possibly due to poor parametrization of hydrogen bonds preventing the spontaneous formation of α-helices. Another limitation is that it accelerates the folding process thus possibly preventing the correct formation of secondary structure elements, leading to structural artifacts.

Despite such drawbacks, Levitt model introduced a number of seminal ideas in the field of coarse-grained modeling. For instance, it showed that folding could be favored by a combination of chain stiffness and flexibility at special turn-promoting points. As already discussed (see Sect. KS and HT model), this idea was later borrowed by Honeycutt and Thirumalai who marked turn points of proteins with neutral residues whose dihedral coefficients favor the interconversion between the *trans* and *gauche* conformations.

The experience gained with the 1975 model, also led Levitt to make some general remarks on the applicability of coarse-grained models. He observed that the averaged forces computed in simplified models lead the protein chain to a rapid collapse in a compact, globular conformation. The protein can now move between various approximately isoenergetic neighboring minima about 6Å RMSD away from the native structure. At this stage, however, due to the side-chain packing, detailed atom-atom interactions would come into effect and the coarse-grained description of the protein becomes inadequate. He therefore, suggested to shift from a coarse-grained to an all-atom representation when the protein enters the molten globule state, thus anticipating the introduction of *multi-scale models*. The research on coarse-grained models is still ongoing in Levitt's laboratory. The new 2002 model [94], however, is more similar to Irbäck model (see Sect. Irbäck C-model) than to the original 1975 release. In order to attain a better description of hydrogen bonds, the backbone is modeled at the all-atom level (with the only exclusion of H_α atoms). The model also includes C_β atoms while the remainder of the side-chain reduces to a single virtual atom. With respect to the 1975 model, the force field now includes much fewer terms, namely an hydrogen bonding, an hydrophobic burial and a residue-residue interaction contribution. The burial and the pairwise interaction potentials are expressed as linear combinations of Chebyshev polynomials and the model comprises a total of 755 parameters that were tuned either through Z-score optimization [94] or through a funnel sculpting approach [95]. The latter method provided promising results in that it yielded parameters allowing the folding of five unrelated sequences at once.

Kolinski-Skolnick 2-Bead Model

Another approach considering two reaction centers per residue was proposed by Kolinski and Skolnick in [96] (KS2). The protein backbone is modeled as a chain of beads centered in the position of the α-Carbons and it is constrained onto a cubic lattice with edges of unit length. In each conformation, the backbone can be ideally built by adding one bond at a time. In the coarser lattice introduced by the authors, each new bond can be chosen among 56 vectors that are generated by circular permutation of the *x*, *y* and *z* coordinates of fundamental vectors (2,1,1), (2,1,0) and (1,1,1), and considering all possible combinations of signs. A finer lattice was also created by permutations of vectors (3,1,1), (3,1,0), (3,0,0), (2,2,1), and (2,2,0). The purpose of this procedure was the creation of protein conformations as close as possible to the known geometrical features of natural proteins.

Side chains are modeled as single united atoms and occupy off-lattice positions. Each side chain can occupy a number of alternative rotamers depending on the chemical nature of the aminoacid and on the local backbone conformation. Each side group is characterized by a strongly repulsive

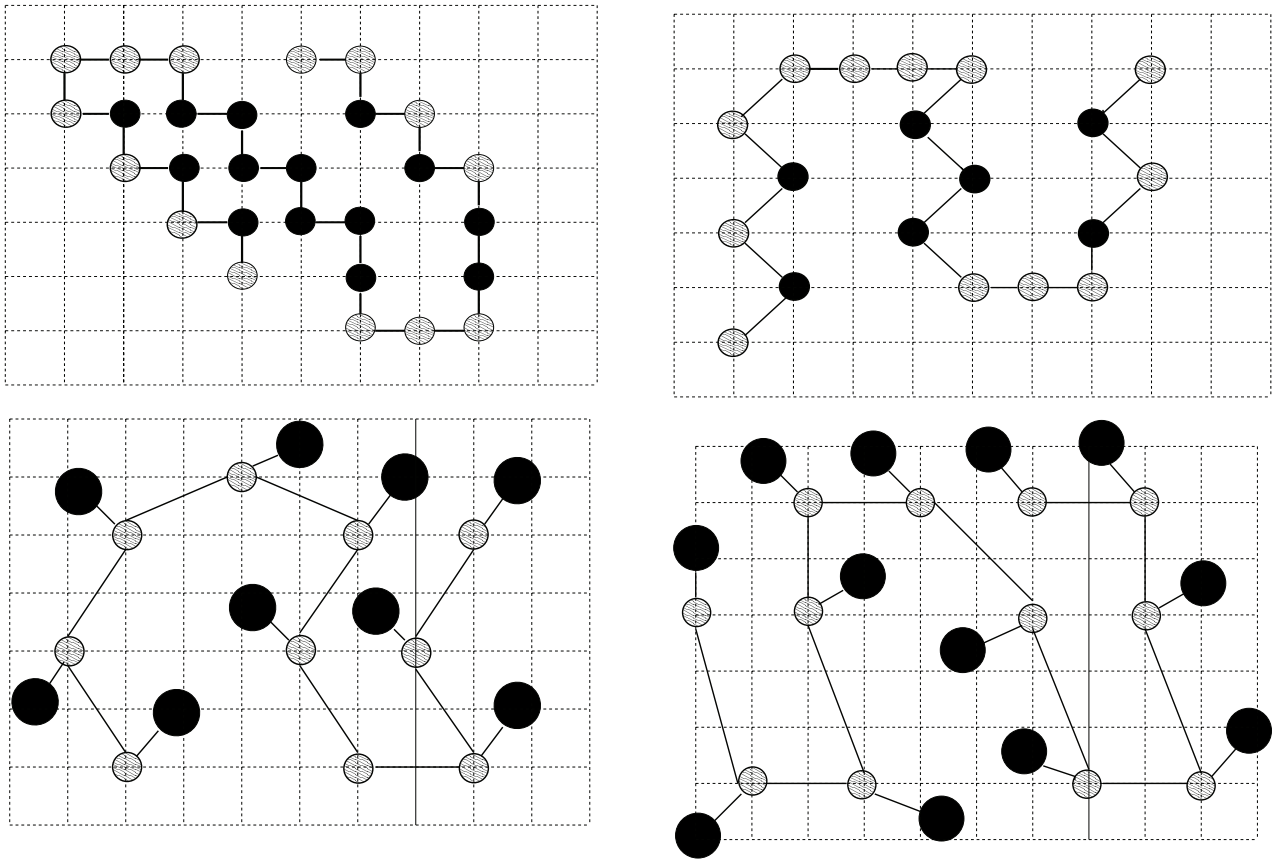

Fig. (11). Kolinski and Skolnick lattice protein models [97]. (**A**) Cubic lattice: black and white beads represent hydrophobic and hydrophilic residues respectively. (**B**) Face centered cubic lattice with two types of residues. (**C**) The coarse-grained lattice model introduced by Kolinski and Skolnick. The backbone units are white while side-chains, occupying off-lattice positions, are black. (**D**) The finer lattice model. Side-chains are once again off-lattice (adapted from Fig. 1 of Ref. [97]).

square well, surrounded by an interaction sphere portrayed as a square well too. The geometric features of the coarse-grained and fine-lattice protein representations introduced KS2 model are sketched in Fig. (**11**).

Dynamic Monte Carlo simulations [96], thanks to the existence of a coarser and a finer lattice representation, allowed a hierarchical approach, whereby a fast simulation on the coarser lattice yield conformations with a loose side chain packing that can be refined by a new run on the finer lattice, leading to structures with a correct hydrogen-bonding pattern and protein-like chain packing. The latter conformations enable the identification of a number of secondary structure and tertiary contact constraints that can be used in a final, all-atom simulation used as a final refinement step.

The Hamiltonian of KS2 model is the sum of a number of potentials of mean force derived from a statistical analysis of the Protein Data Bank and basically reads:

$$E = E_{\mathrm{Ca-trace}} + E_{HB} + E_{\mathrm{Rot}} + E_{\mathrm{SC-corr}} + E_{\mathrm{Hyd}} + E_{\mathrm{Pair}} + E_{\mathrm{MB}}$$

The term $E_{Ca\text{-}trace}$ is a sequence-independent constraint for the chain to remain in the protein-like region of the conformational space. In other words, it increases the frequency of visits to the minima of the Ramachandran plot. This term consists of a set of energy parameters that were obtained from the frequency distribution of the distances between the α-Carbons *i*-th and (*i+3*)-*th* in the structures of the protein database and then reversing the Boltzmann expression.

Energy E_{HB} is a cooperative hydrogen bonding interaction; E_{Rot} accounts for the conformational energy of side-chains; $E_{\mathrm{SC-corr}}$ depends on the positional correlation of the side-groups; E_{Hyd} models the hydrophobic effect rewarding hydrophobic residues located in the core of the protein; E_{Pair} is a pairwise interaction term and finally E_{MB} is a multibody interaction term accounting for the cooperativity in side-chain packing. The terms $E_{\mathrm{SC-corr}}$ and E_{MB} were introduced for the first time in Ref. [96].

In this model, H-bonds link Cα atoms whose distance is within a given range and they must satisfy specific geometric constraints: $|(\mathbf{b}_{i-1}-\mathbf{b}_i)\cdot \mathbf{r}_{ij}| \le a_{max}$ and $|(\mathbf{b}_{j-1}-\mathbf{b}_j)\cdot r_{ij}| \le a_{\max}$ which means that the difference of the bond vectors $(\mathbf{b}_{i-1}-\mathbf{b}_i)$ and $(\mathbf{b}_{j-1}-\mathbf{b}_j)$, that are perpendicular to the direction of the backbone, must also be orthogonal to the r_{ij} vector linking the H-bonded residues, [see Fig. (**12**)].

Hydrogen-bond formation in this model is highly cooperative: whenever two H-bonds insist on consecutive residues, the protein receives an additional stabilization, which favors the formation of α-helices and β-sheets

$$E_{\mathrm{HB}} = \sum_{i,j} E^{H}\Delta(i,j) + \sum_{i,j} E^{HH}\Delta(i,j)\ \ \Delta(i\pm 1, j\pm 1)$$

Here E^{H} is the energy of a single H-bond, E^{HH} is the cooperativity prize and $\Delta(i,j) = 1$, if residues *i* and *j* satisfy the geometric requirements for H-bonding and *0* otherwise. The terms E_{rot} and $E_{SC-corr}$ both model the short-range interactions. In particular, $E_{SC-corr}$ stems from the observation that the correlation between the orientation of side chains has a large influence on the local conformation. The $E_{SC-corr}$ term was derived as mean field potential of $\cos(\Theta_{i,i+k})$ by using as a reference state a random population with a uniform distribution in all bins. The angle $\Theta_{i,i+k}$, Fig. (**13**), is between the vectors connecting α-Carbons *i* and *i+k* with the respective side-chains. A similar approach was followed to derive E_{rot} from the frequency distribution of rotamers for each amino-acid type.

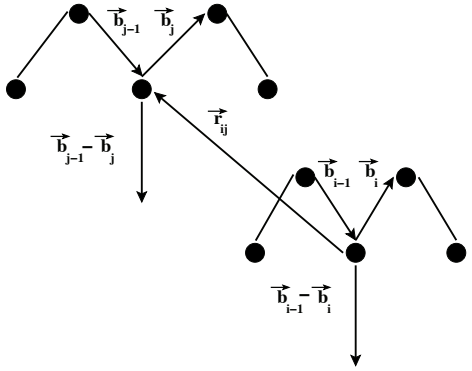

Fig. (12). Geometric elements necessary to compute the structural constraints for hydrogen bond formation in KS2 model (adapted from Figure 5 of Ref. [97]). The difference between two successive bond vectors such as $\mathbf{b}_{i-1}$ - $\mathbf{b}_i$ and $\mathbf{b}_{j-1}$ - $\mathbf{b}_j$, is approximately perpendicular to the direction of the polymer chain. In order to establish a hydrogen bond between residues *i* and *j*, $\mathbf{b}_{i-1}$ - $\mathbf{b}_i$ and $\mathbf{b}_{j-1}$ - $\mathbf{b}_j$ must also be orthogonal to vector $\mathbf{r}_{ij}$. In other words, this constraint favors the establishment of hydrogen bonds between parallel stretches of polymer chain.

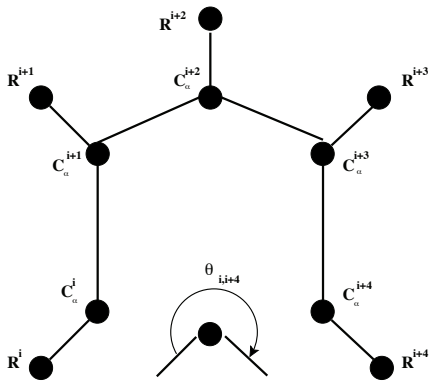

Fig. (13). Sketch of the angular correlation in the side-chain orientation of KS2-model. The angle Θ appearing in Eq.(13) is between two vectors each connecting the C_α of a residue with its side-chain. These two vectors must refer to units no more than 4 residues apart. (Adapted from Figure 6 of Ref. [97]).

As a conclusion, the short-range interaction energy is computed as:

$$E_{\mathrm{rot}} + E_{\mathrm{SC-corr}} = \sum_i \left\{ E_{\mathrm{rot}}(a_i) + \sum_{k=1}^{4} E_k(\Theta_{i,i+k}, a_i, a_{i+k}) \right\} \qquad (13)$$

where a_i is the aminoacid on the site *i*.

There are three contributions to the long-range interactions: E_{hydrop}, E_{pair} and E_{multib}. The first simulates the hydrophobic effect and can be derived as a potential of mean force from the distribution of the distance (normalized to the radius of gyration) of each amino-acid from the center of mass of the protein. The pairwise interaction energy is expressed as:

$$\begin{cases} E^{rep} & \text{for} \quad r_{ij} < R_{ij}^{rep} \\ \varepsilon_{ij} & \text{for} \quad R_{ij}^{rep} < r_{ij} < R_{ij} \quad \text{and for} \quad \varepsilon_{ij} \geq 0 \\ f\varepsilon_{ij} & \text{for} \quad R_{ij}^{rep} < r_{ij} < R_{ij} \quad \text{and for} \quad \varepsilon_{ij} < 0 \end{cases}$$

where E^{rep} is a penalty for the side-chain overlap, ε_{ij} is a pair-dependent potential of mean force derived from statistical analysis of the protein database, and the *f* factor favors the attraction of residues located on secondary structure elements forming small angles. In order to attain compact side-chain packing and a cooperative transition from the molten globule to the native conformation, Kolinski and Skolnick introduced a cooperative term also for side-chain interaction. This contribution is a four-body term of the following form:

$$E_{multib} = \sum_{i,j} \sum_{k,n} (\varepsilon_{i,j} + \varepsilon_{i+k,j+n}) C_{i,j} C_{i+k,j+n}$$

where |k| = |n| and n = ±3, ±4; $\varepsilon_{i+k,j+n}$ is the energy prize that rewards interactions between side chains of residues close in the primary sequence, while $C_{i,j}$ =1 if side chains *i* and *j* are in contact, and zero otherwise. The KS2 model was tested on three proteins: the B-domain of staphylococcal protein A, a monomeric version of ROP dimer and Crambin [92]. To avoid a possible bias, the analyzed proteins were not included in the database used for deriving the statistical potentials. The simulations with protein A and ROP were quite successful: protein A has a three-helix bundle topology that could be recovered in 2/3 of the runs in the low resolution lattice. The structures so obtained have a molten globule conformation with correctly formed secondary structure, but poorly defined tertiary contacts. The refinement on higher resolution lattice yielded conformations with an average RMSD of 2.25Å from the native structure. The simulations with ROP have been even more successful predicting the correct four-helix bundle structure in 11 runs out of 12, many years before, this fold was experimentally resolved by Kresse *et al.* [98]. The simulations on Crambin were much less satisfactory suggesting that the model might be biased toward highly regular helical structures. The antiparallel β sheet which in Crambin is packed against a helical hairpin reduces to 10% the success rate of the runs. Such a limitation was removed in the improved version named CABS [99]. The acronym stems from the four interaction centers per residue: Cα, Cβ, the center of the side chain and the united peptide bond as in the UNRES model (following section). This allowed a better description of the hydrogen-bonds that is essential for the correct formation of the secondary motifs. With respect to the original KS2 model, CABS is characterized by a more careful design of a set of sequence independent potentials that enforce a protein-like geometry compensating for the structural inaccuracy due to the simplified lattice representation. For instance, the latter environment artificially induces a gaussian distribution of the end-to-end distance of four-bond stretches in contrast with the experimental bimodal distribution. In a similar way, the lattice structure makes "crumpled" conformations with very close U-turns, more frequent than in real proteins. The right handedness of the α-helices and the up-and-down geometry of β-sheets is also imposed through sequence independent potentials. The model is completed by short and long range sequence dependent orientational interactions. The performance of CABS in the CASP6 competion achieved the second best score among 200 participants. A significant test was the study by Kmiecik and Kolinski [100] on the Chymotrypsin Inhibitor 2 (CI2) and Barnase. In agreement with experimental data the authors showed that residual structure elements in the unfolded ensemble act as nucleation sites where the folding begins. Moreover the computed Φ-values of CI2 matched the experimental ones. As CI2 and Barnase fold via a two-state and a multi-state kineticts respectively, CABS is expected to reliably reproduce a wide range of folding mechanisms. Finally, as both proteins belong to the $\alpha+\beta$ fold the bias of KS2 for α-helical conformations appears to be overcome.

UNRES Model

The *UNRES model*, originally developed by Scheraga's group in 1993 [13], generalizes Levitt's model [10]. It is similar as far as the geometry of the polypeptide chain is concerned, but it has a more complex force-field including terms that confer mobility to side-chains, and multibody terms that are relevant to the formation of secondary structures. The UNRES model was successfully used in conjunction with many optimization techniques such as the Monte Carlo with energy minimization (MCM), the electrostatically driven Monte Carlo (EDMC), and the conformational simulated annealing [13,101]. Recently UNRES was also adapted to a Molecular Dynamics simulation scheme [102].

The UNRES simplified protein chain, (Fig. (**14**)), is described as a sequence of α-carbons linked by virtual bonds. In the middle of each virtual bond, there is a united peptide group, and each C^{α} is also linked (through the SC_i vector) with a united side chain represented as a sphere or as an ellipsoid. The united side chains and the united peptide groups are the only interaction centers of the molecule while the Cα atoms play a role only in the definition of the geometry. All the virtual bond lengths are fixed and the geometrical variables of the model are: i) the virtual bond angle ϑ defined by 3 consecutive C^{α} atoms C^{α}_{i-1} C^{α}_{i} C^{α}_{i+1}, ii) the virtual dihedral angle γ defined by 4 consecutive Cα atoms: C^{α}_{i-1} C^{α}_{i} C^{α}_{i+1} C^{α}_{i+2}, iii) the side chain bond angle α_{SC_i} formed by SC_i and the bisector of the angle defined by C^{α}_{i-1} C^{α}_{i} C^{α}_{i+1} and iv) the side chain dihedral angle β_{SC_i} of rotation about the bisector of C^{α}_{i-1} C^{α}_{i} C^{α}_{i+1} angle, (Fig. (**13**)).

The energy function of the UNRES model is expressed by:

$$U = \sum_{i,j>i} U_{SC_i,SC_j} + \sum_{j\neq i} U_{SC_i p_j} + w_{el} \sum_{i,j>i+1} U_{p_i p_j} + w_{corr} U_{corr}$$
$$+ w_{tor} \sum_i U_{tor}(\gamma_i) + w_{loc} \sum_i \left[U_b(\vartheta_i) + U_{rot}(\alpha_{SC_i}, \beta_{SC_i}) \right]$$

where $U_{SC_i SC_j}$ accounts for the interactions between united side chains and is determined mainly by hydrophobic/hydrophilic interactions, the $U_{SC_i p_j}$ term is an excluded volume potential preventing the collapse of side chains onto the peptide groups, the $U_{p_i p_j}$ term represents the average electrostatic interaction between the centers of the peptide groups and accounts for hydrogen bonding within the backbone, $U_{tor}(\gamma_i)$ denotes the energy of variation of the virtual bond dihedral angle γ_i, $U_b(\vartheta_i)$ is the bending energy of the virtual bond ϑ_i, $U_{rot}(a_{SC_i}, \beta_{SC_i})$is the local energy of side chain *i* and U_{corr} includes the cooperative terms.

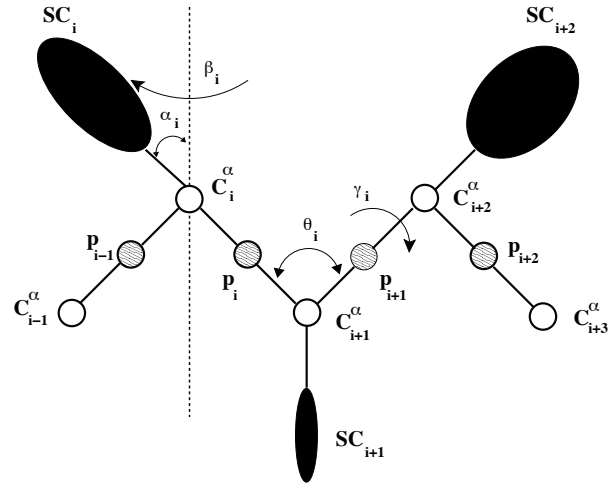

Fig. (14). Representation of a stretch of polypeptide chain in UNRES. The white beads represent the α Carbons whose role is to enforce the correct chain geometry. Halfway between successive Cα's lie the united peptide groups represented as dashed circles. The united side chains are black ellipsoids and are linked to each Cα through a fixed bond vector. The degrees of freedom of the model are the virtual bond (θ) and dihedral (γ) angles, as well as angles α and β that describe the mobility of side chains (adapted from figure 1 of Ref. [11]).

The *w* values represent the relative weights of the respective energy terms. The general form of the side chain interaction is given by:

$$U_{ij} = 4[|\varepsilon_{ij}| x_{ij}^{12} - \varepsilon_{ij} x_{ij}^{6}] \tag{14}$$

with ε_{ij} the pair specific van der Waals well depth, positive for hydrophobic-hydrophobic interactions and negative for hydrophobic-hydrophilic and hydrophilic-hydrophilic interactions. The particular choice of the sign of ε_{ij} favors close contacts between hydrophobic side chains to yield close packing of hydrophobic groups within the core of the protein. The x_{ij} parameter can take on different functional forms. In the case of the two radial potentials employed in the UNRES model:

$$x_{ij} = \frac{\sigma_{ij}^0}{r_{ij}} \qquad x_{ij} = \frac{r_{ij}^0}{r_{ij} + r_{ij}^0 - \sigma_{ij}^0} \tag{15}$$

The constant ${\sigma_{ij}}^0$, in this case represents the distance between side chains *i* and *j*, such that $U(r_{ij} = {\sigma_{ij}}^0) = 0$. The second functional form of x_{ij} shown in Eq.(15) corresponds to the shifted Lennard-Jones potential proposed by Kihara (LJK). In this case, the quantities ${r_{ij}}^0 - {\sigma_{ij}}^0$ and ${r_{ij}}^0$ can be identified with the dimensions of the hard and soft core respectively. In this function $U(r) = 0$ for $r = \sigma$ and $U_{min} = -\varepsilon$ as in the LJ potential, but in the LJK potential the energy barrier is at $r = \sigma - r_o$, when the hard core spheres of the two interacting side chains begin to inter-penetrate. In the 1998 model, three potentials with angular dependence were also tested, namely: the Berne-Pechukas potential (BP) [103], the Gay-Berne potential (GB) [104] and the Gay-Berne-Vorobjev potential (GBV) [105]. These are still described by Eq.(14), but the parameters ε_{ij} and x_{ij} depend not only on the distance but also on the relative orientation of the two interacting side chains. In the 1993 version of the model [13], the van der Waals radii were taken from set *C* of Levitt [10], while the ε's were calculated from the inter-residue contact energies [81]. In the 1997 model [11], conversely, the ε_{ij} parameters, like all other parameters of the $U_{SC_iSC_j}$ potential were chosen so that the theoretical correlation function fit best to the correlation functions determined from protein crystal data.

The side chain-peptide group interaction is modeled as an excluded volume potential: $U_{SC,p} = \varepsilon_{SCp}(r_{SC,p}/r_{ij})^6$. This term clearly acts as a penalty function forbidding too close contacts of the side chain of one residue with the backbone of another one. The peptide-group peptide-group interaction is treated as the interaction between two permanent dipoles placed in the middle of the $C^\alpha - C^\alpha$ virtual bonds. Two permanent dipoles with moment vectors $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ placed at distance r_{ij} interact as:

$$U_{ij} = \frac{\mathbf{p}_i \cdot \mathbf{p}_j - 3(\mathbf{p}_i \cdot \mathbf{e}_{ij})(\mathbf{p}_j \cdot \mathbf{e}_{ij})}{\varepsilon r_{ij}^3},$$

where $\mathbf{e}_{ij}$ is the unit vector pointing from the center of peptide group *i* to the center of peptide group *j* and ε is the dielectric constant. The orientation of $\mathbf{p}_i$ and $\mathbf{p}_j$, located in the middle of two consecutive C^α, is determined according to a method proposed in Ref. [106]. For the evaluation of the dihedral energy, three classes of amino acids were considered [12]: Gly, Ala and Pro (the Ala class includes all the amino acids apart for Gly and Pro).

The dihedral potential

$$U_{tor}^{XY}(\gamma) = A_0 + \sum_{k=1}^{n}\left\{A_k^{XY}[1+\cos(k\gamma)] + B_k^{XY}[1+\sin(k\gamma)]\right\} \tag{16}$$

is expressed as a Fourier's series whose coefficients are tuned on experimental distribution of the γ angles in protein crystals. In Eq.(16), *n* takes on values from 3 to 6, depending on the type of the two central residues *X* and *Y* of the virtual dihedral angle.

The bending potential $U_b(\theta)$ was also derived as a potential of mean force from a crystallographic frequency distribution of θ angles that turned out to be the sum of two Gaussians, one centered at $\theta_0 = 90^o$ and the other one centered at a residue specific value θ_c. A statistical analysis of the PDB structures showed that the angles α_{SC} and β_{SC} defining the orientation of a side chain centroid with respect to the C^α frame vary considerably and tend to cluster into different rotamers. Thus, the simplest analytical form of the distribution of side-chain rotamers would be a sum of two-dimensional Gaussians in α_{SC} and β_{SC}. However, a correlation was found to exist between the virtual bond angle ϑ and the angles α_{SC} and β_{SC} belonging to the same residue. Therefore, a correct expression for rotamer distribution has to include three-dimensional Gaussians in ϑ, α_{SC} and β_{SC}. Such information was useful to compute the orientational potential of side chains as a potential of mean force.

The first version [13] of the UNRES model assumed that virtual bond angles ϑ had a constant value of 90^o and that α_{SC} and β_{SC} angles had constant values too. It could successfully predict the three dimensional structures of simple helical proteins such as the avian pancreatic polypeptide (APP) [13] and Galanin [107]. However, one of the reasons for the success of this first generation force-field was that the

C^{α}–C^{α}–C^{α} angles were fixed, thus introducing a bias towards regular helical structures. Therefore, although the model was able to reproduce some α-helical proteins, there were problems with β-sheet structures.

In the 1997 version of UNRES [11], the angles ϑ, α_{SC} and β_{SC} were kept variable, but the structural prediction of the β-class proteins could be achieved only via the introduction of multibody terms accounting for the cooperativity in backbone hydrogen bonding and correlation between local and electrostatic interactions [108]. The multibody terms arise naturally as follows. The energy $E(\mathbf{x},\mathbf{y})$ of a physical system can be considered as a function of two kinds of variables: the primary (or important) variables $\mathbf{x} = (x_1,..., x_m)^T$ whose variation leads to major changes in the conformation of the system, and the *secondary* variables $\mathbf{y} = (y_1,..., y_n)^T$ that only weakly affect the conformation. A straightforward strategy to develop a simplified model of the system is therefore to average the energy over the secondary variables, yielding a mean-force potential

$$U(\mathbf{x}) = F(\mathbf{x}) = -RT\ln\left\{\frac{1}{V_y}\int_{V_y} d^n y\ \exp[-E(\mathbf{x},\mathbf{y})/RT]\right\} \quad (17)$$

where $V_y = \int d^n y$ and $F(\boldsymbol{x})$ is the average free energy of the system corresponding to a fixed value $\mathbf{x}$ of the important variables. In UNRES the primary variables are the virtual bond angle ϑ, the virtual bond dihedral angle γ and angles α_{SC} and β_{SC} that define the orientation of the side chains. The secondary variables are the dihedral angles χ that describe the conformation of the side chains, the angles λ of rotation of the peptide groups around the $C\alpha-C\alpha$ virtual bonds and the degrees of freedom of water. It should be noted that the average free energy as expressed by Eq.(17), can be evaluated only numerically, but in the case of proteins, this computation is very time consuming due to the large number of degrees of freedom. This problem can be relieved using a reliable approximation of the average free energy. This was computed by Scheraga and coworkers via the *cluster cumulant expansion* method developed by Kubo [109]. Basically, the potential of mean-force is expanded as a sum

$$F(\mathbf{x}) = \sum_i f_i^{(1)}(\mathbf{x}) + \sum_{i<j} f_{ij}^{(2)}(\mathbf{x}) + \sum_{i<j<k} f_{ijk}^{(3)}(\mathbf{x}) + \ldots\ldots + \sum_{i_1<i_2<\ldots<i_n} f_{i_1,i_2,\ldots,i_n}^{(n)}(\mathbf{x}) + \ldots\ldots$$

where, as an example, the first three factors can be expressed as:

$$f_i^{(1)}(\mathbf{x}) = F_i^{(1)}(\mathbf{x})$$

$$f_{ij}^{(2)}(\mathbf{x}) = F_{ij}^{(2)}(\mathbf{x}) - F_i^{(1)}(\mathbf{x}) - F_j^{(1)}(\mathbf{x})$$

$$f_{ijk}^{(3)}(\mathbf{x}) = F_{ijk}^{(3)}(\mathbf{x}) - F_{ij}^{(2)}(\mathbf{x}) - F_{ik}^{(2)}(\mathbf{x}) - F_{jk}^{(2)}(\mathbf{x}) + F_i^{(1)}(\mathbf{x}) + F_j^{(1)}(\mathbf{x}) + F_k^{(1)}(\mathbf{x})$$

The first order factor $f_i^{(1)}$ is just the average free energy of component interaction i. The terms U_{SCSC},U_{SCp}, U_{pp} as well as U_b and U_{rot} are typical examples of the $f_i^{(1)}$ factors. If the average free energy of each component interaction is independent of the others, the $F(\mathbf{x})$ function will be a sum of the $f_i^{(1)}$ terms only and this case is regarded as the independent sites approximation. The factors $f_{ij}^{(2)}$ contain the average free energy of pairs of component interactions minus the sum of the average free energies of single component interactions. Therefore $f_{ij}^{(2)}$includes correlation effects due to the coupling between the secondary degrees of freedom i and j. Likewise, the third order factors reflect the coupling between three component interactions which cannot be decomposed in coupling between pairs.

The analytic expression of the multibody term is extremely complicated [14,108], however the relevant contributions to the average free energy function $F(\mathbf{x})$ are provided by the $U_{MB}(p\text{-}p)$ *and* $U_{MB}(el\text{-}loc)$ terms. The $U_{MB}(p\text{-}p)$ energy is a four-body term corresponding to the correlation of the electrostatic interactions between the pairs of neighboring dipoles associated to peptide groups. This term is very important because accounts for the well-known phenomenon of cooperativity in α-helix formation that was also studied by Kolinski and Skolnick [110,111].

The term $U_{MB}(el\text{-}loc)$, on the other hand, is necessary to attain correct conformational predictions of proteins belonging to the β-structural class. This accounts for correlations between local interactions and peptide group-peptide group interactions. The progressive improvements of UNRES approach can be followed by the results attained in different editions of CASP (Critical Assessment of Structural Prediction) competitions. In CASP3 the applicability of UNRES was limited to short fragments of α structural family. The inclusion of cooperativity terms based on cumulant expansion improved the performance to correct structural prediction of some proteins of the $\alpha+\beta$ family (CASP4) [112]. UNRES performance was improved through a complete reparametrization of the Force-Field via a funnel sculpting (hierarchical optimization) procedure that was trained on four proteins of different α, β content. The resulting P4 Force-Field correctly predicted the structure of 50-79 residue fragments of the α, β, and $\alpha+\beta$ classes, even if with some significant topological errors. A further optimization of P4 enabled an accurate prediction of the structure of two α and three $\alpha+\beta$ proteins in CASP6. In particular the correct prediction of the three-helix bundle of target T0215, whose non-homologous template was available, showed that a physics based CG-model could outperform knowledge-based methods [113]. Finally, it must be remarked, that in the last UNRES versions the conformational search is carried on through the multiplexing replica exchange MD method, which replaces the traditional conformational space annealing based on genetic algorithm. The new algorithm allows exploring time scales of the order of the microsecond and milliseconds for small proteins [114].

Irbäck C_{β}-Model

Another interesting model still describing the side chain as a single interaction center was developed by Irbäck and coworkers [115-117]. As shown in Fig. (**15**), every residue retains the backbone atoms Cα, C' and N as well as the O and H of the backbone units. The side chain is represented by a C_{β} only and can be hydrophobic, hydrophilic or absent (Glycine residues). Bond lengths and bond angles are kept fixed so that the internal coordinates reduce to the backbone dihedrals φ and ψ. The energy function of this model is simply given by the sum:

$$E = E_{dih} + E_{excl} + E_{hb} + E_{hp}$$

of a dihedral term, an excluded-volume term, a hydrogen bonding energy and a potential of interaction between hy-

drophobic residues. The dihedral potential embodies the standard three-fold symmetry:

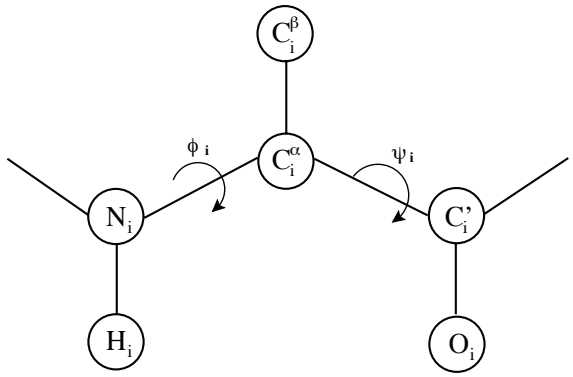

Fig. (15). Structural representation of a protein residue within the Irbäck model. Adapted for figure 1 of Ref. [115]).

$$E_{dih} = \frac{\varepsilon_\phi}{2}\sum_i (1+\cos 3\phi) + \frac{\varepsilon_\psi}{2}\sum (1+\cos 3\psi)$$

The excluded-volume E_{excl} energy also has the standard $(\sigma_{ij}/r_{ij})^{12}$ distance dependence: it involves pairs of hydrophobic C_βs and $\sigma_{ij} = \sigma_i + \sigma_j + \Delta\sigma_{ij}$. The radii of the atoms hard-core σ, in the latter expression, have been determined mostly by trial and error and values of $\Delta\sigma_{ij}$ greater than zero prevent a possible clash of C_β atoms with backbone atoms.

The hydrogen bonding term is the product of a distance-dependent Lennard-Jones-like and an orientation-dependent factor designed to favor the Donor-Hydrogen-Acceptor alignment:

$$E_{hb} = \varepsilon_{hb}\sum_{ij} u(r_{ij})v(\alpha_{ij},\beta_{ij})$$

where:

$$u(r) = 5\left(\frac{\sigma_{hb}}{r}\right)^{12} - 6\left(\frac{\sigma_{hb}}{r}\right)^{12}$$

$$v(\alpha,\beta) = \begin{cases} \cos^2\alpha\cos^2\beta & \text{if } \alpha,\beta > \pi/2 \\ 0 & \text{elsewhere} \end{cases}$$

The above equations refer to hydrogen bonds between the N−H and C=O groups of the backbone and α_{ij} and β_{ij} designate respectively angles NHO and HOC'. The force-field is completed by a hydrophobicity term in the classical Lennard-Jones form:

$$E_{hp} = \sum_{ij>i}\left[\left(\frac{\sigma_{hp}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{hp}}{r_{ij}}\right)^{10}\right].$$

The most striking feature of this model is its remarkable simplicity, especially if compared with the sophisticated formulations of the other models of the same family reviewed in this section. However, it must be considered that this model was parametrized by trial and error so as to reproduce the features of a specific set of proteins under investigation. Application to other proteins may therefore require a re-adjustment of the parameters. The model was successfully applied to a fragment of the four-helix bundle protein designed by Regan and De-Grado [118]. The three-helix bundle topology was correctly reproduced and the two-state folding mechanism as well [115,117].

CONCLUSIONS

We surveyed several coarse-grained (CG) protein models without any claim to be exhaustive, rather we preferred to select models that, in our opinion, marked the historical evolution of the field.

The principle of coarse-grained modeling amounts to grouping together atoms of side-chains and even a complete residue into simple units (virtual atoms) which absorb most of the molecular details. This approach stems from the necessity to link atomistic and mesoscopic space and time scales, the choice of a certain CG model being generally guided by the scales of interest and the problem addressed.

If the objective of the research is the structural prediction of a protein, then knowledge-based methods such as *threading* and *homology modeling*, constitute the best strategy. In fact, the availability of templates with more than 50% sequence identity yields predictions with about 1Å-RMS error for main chain atoms, which is comparable to the accuracy of a medium-resolution NMR experiment [6]. However, when no template with sequence identity higher than 30% is available, the accuracy of knowledge-based methods drops dramatically due to alignment errors [6], requiring the resort to *de novo* physics-based models.

It has been estimated [119] that no fewer than 16000 structures should be deposited in the PDB to provide templates for 90% of the protein structural families. Moreover, there exist proteins with similar structures but no sequence homology, which represent difficult targets for knowledge-based methods [120,121].

However, the application of physics-based models is not restricted just to structural prediction as they aim at a deeper understanding of the basic interactions and dynamics that govern protein properties. Physics-based models in fact, allow the exploration of processes not pertaining to bioinformatics such as: conformational changes (e.g. allosteric transitions and induced fit), the action mechanism of proteins (e.g. enzyme catalysis) and the dependence of protein structure and function on environmental conditions (temperature, pH, salts, denaturants). For examples, Ising-like (binary) models were successfully used to study α-helix and β-hairpin formation [37,122]. In these simple protein systems, binary models allowed a straightforward calculation of physical properties such as folding rates, sometimes bridging the gap between theory and experiment. The discovery of a correlation between the contact order and the experimental folding rates [23] led to the suggestion that Ising-like models could be generally applicable to any protein fold. The contact order, in fact, is just the mean loop length of a protein, and since loops are similar to β-hairpins whose folding could be correctly reproduced by binary models, it was deduced that Ising-like models could be used to study complex globular proteins. A deeper analysis, however [123], revealed that Ising-like models systematically impose the *nucleation-condensation* folding mechanism i.e. folding starts with the formation of a nucleus composed by residues far away along the protein sequence, and then the formation of secondary and tertiary structure occurs almost simultaneously. This mechanism, however, cannot be applied to all proteins. Another common mechanism, in fact, is described by the *diffusion-collision* process according to which secondary struc-

tural elements form first and then they self-assemble into the correct tertiary structure. This mechanism cannot be reproduced by Ising-like models because their formalism does not allow the representation of intermediates where secondary structures are formed but tertiary ones are not. Assume that a protein is formed by two α-helices packed against each other and connected by a loop. According to the diffusion-collision model, there will be an intermediate with two helices connected by a disordered loop. This structure has no representation in the binary formalism since the string 111100001111 corresponds to a situation where the helices are natively packed. The limits of Ising-like models are even more evident in the light of the recently proposed *zipping and assembly* (ZA) mechanism [124]. The basic idea is that a global optimization problem can be solved by combining a set of locally optimal solutions. A polymer chain can then be divided into small fragments that rapidly fold into metastable structures such as α-helical turns, β-turns or small loops. These structures can be stabilized through interactions with other structured fragments. The process is then iterated with the formation of larger and larger complexes until the protein is completely folded. Ising-like models can then be used when there are reasons to think that the protein will fold through a nucleation-condensation mechanism, and in any case, many physical observables such as Φ-values will only be reproduced qualitatively.

When the topology of native states strongly influences the folding process, Go models and more in general topology-based models turn out to be more useful in predicting folding mechanisms.

Experimental support to the applicability of such approaches is well summarized in the review by [125]: i) mutations not affecting the overall native state topology of a protein, have weak influence on its folding rate; ii) the transition states of proteins with similar structures are similar too, almost irrespective to sequence variation; iii) folding rates of small proteins correlate with simple topological indicators, such as contact order.

Go models, by rewarding only native contacts, posit the existence of a minimally frustrated energy funnel that warrants the stability and fast foldability of natural proteins. Go models have been successfully used for the sampling of the Transition State Ensemble in view of the Φ-value computation, from which, on turn, folding rates and mechanisms can be explored [62,63, 67,126]. Go models have been employed in a wide range of applications from the study of mechanical unfolding [127,128] to biomolecular machines [129] and from the analysis of influence on folding of macromolecular crowding and confinement [130,131] to the determination of the action mechanism of some enzymes [132]. The minimal frustration principle, however, is a zero-th order approximation, and recent studies point out the spread and relevance of energetic frustration in real proteins. For instance, the occurrence of ϕ-values larger than one usually indicates the presence of native contacts that stabilize the transition state without being present in the native one. This experimental evidence is supported by all-atom simulations with the CHARMM force field, according to which non-native contacts account for 20-25% of the transition state energy [133]. The ideal funnel scenario implied by Go-model approach may sometimes conflict with the evolutionary pressure to optimize biological function [134]. Indeed a statistical survey of the Protein Data Bank using a quantitative parameter for localizing frustration, showed that 15% of contacts are highly frustrated, and they cluster near ligand binding sites so that a functional role may be suggested. Moreover, as suggested by Plotkin and Clementi a moderate amount of frustration reduces the free energy barriers and increases folding rates [135]. This apparently counterintuitive mathematical result, was confirmed by numerical simulations on the src SH3 protein and by the experimental finding that a strengthening of the non-specific hydrophobic stabilization of α-*spectrin* SH3 domain sped up the folding process [136]. In agreement with these observations, recent implementations of the Go-model, such as those by Karanicolas and Brooks [72] and by Das, Matysiak and Clementi [74], make use of moderately frustrated funneled landscapes allowing more realistic simulations.

Another challenge for Go models is the inclusion of sequence effects. In Go models the folding is driven by the topology of the native state, but it is the sequence that determines the topology. In this context, the precise role of the sequence in folding remains to be understood. A common strategy to account for sequence effects is to use heterogeneous contact energies that may be chosen using different criteria. Karanicolas and Brooks [72] used the statistical potentials derived by Miyazawa and Jernigan [81], while Dokholyan [137] rescaled the energy couplings based on a set of critical contacts identified through an all-atom calculation. All-atom Go models also implicitly adopt heterogeneous couplings since larger residues will establish a larger number of atom-atom contacts [127,138]. The introduction of heterogeneous energy couplings in Go models is equivalent to the increase in the number of flavours in sequence-based models. The key idea of these models is that the hydrophobic effect is the main driving force of folding so that the attraction between hydrophobic beads and the repulsion between hydrophobic and polar ones is a basic ingredient. The energy landscape generated by these force fields however, is not funnel-like but rather rugged, with many degenerate global energy minima.

As a consequence, with these models, proteins do not fold to a unique native conformation, but rather to a large set of ground state structures. In two dimensions, short-chain, exhaustive simulations with the HP model showed that the fraction of sequences that have a unique ground state is only 2.5% [139], and maximally compact 3D 48-mer designed sequences exhibited more than 10^3 structurally heterogeneous global minima [140]. This problem was still present in the model proposed by Honeycutt and Thirumalai [54] despite the off-lattice geometry and the increase of the number of flavours from two to three. According to Shakhnovich the uniqueness of the native state stems from the existence of a sufficiently large energy gap between the ground state and the lowest-energy decoys [39]. The size of the gap depends on the total number of chain conformations and on the diversity of interactions that depends on the diversity of the amino acid alphabet. The soundness of this remark seems to be confirmed by the historical evolution of sequence-based models from a lattice to an off-lattice geometry and from the two flavors of the HP model [41] to the three flavors of the HT model [54] and the four flavors of the latest release of the Sorenson/Head-Gordon model [61]. The latter model in par-

ticular, seems to have significantly reduced the level of glassiness with respect to the previous versions, yielding for Protein G a folding temperature well above the glass transition temperature (T_f/T_g~2.3). This good performance, however, was attained at the expense of further strengthening the reliance on the secondary structural information, placing the model somewhere between Go-models and sequence-based models. It must also be noted that an excessive bias towards native secondary structures may artificially enforce a diffusion-collision reaction mechanism. From this point of view, thus, the SHG model can have a problem opposite to the one highlighted for Ising-like models that impose a nucleation-condensation folding mechanism [123]. Moreover, the efficient application of the SHG model requires a sequence design procedure that, according to its position in the protein structure, might assign different flavors to the same residue. This reflects the difficulty of including in just a few parameters the effects of the residue size and geometry that lead to highly anisotropic, multimodal interactions critically influenced by the biochemical environment. These limits therefore suggest the opportunity to increase the number of beads per residue to increase the specificity of the interactions.

The two-bead models, initially introduced by Michael Levitt and later developed by Scheraga [11] and Kolinski and Skolnick [96], had a common difficulty in reproducing secondary structures. Levitt model could correctly fold BPTI only starting from a conformation with a pre-formed α-helix [10]. By contrast, the KS2 model was successful in folding of α-helical proteins such as Protein A and ROP, but it performed very poorly on the Crambin benchmark due to the presence of a β-sheet [92]. This limitation motivated a trend towards a more detailed representation of the backbone resulting in an improved modeling of hydrogen bonds. The detailed backbone representation, however, needs to be combined with a careful parametrization, as testified by Irbäck model whose parametrization by trial and error limits its applicability to helical proteins [117]. The history of the UNRES model is somehow different. The original version of the model was incapable of promoting the formation of β-sheets [13], but instead of modifying the representation of the backbone, the authors chose to introduce cooperativity through a complex cumulant expansion technique [108]. Also in this case, however, parametrization played a key role as testified by the good achievements in the CASP contests only after the parameters were fine-tuned through a funnel sculpting approach [113]. One of the most effective multi-bead models is currently represented by the CABS model [99,100] i.e. the evolution of KS2 model [96]. CABS is based on a compromise between the computational efficiency offered by the lattice geometry and an accurate modeling through statistical potentials. The price paid for computing speed, however, is a non protein-like geometry that must be corrected through a number of non physical aspecific potentials.

In extreme synthesis, if the problem at hand is just structure prediction, knowledge-based methods represent the optimal tool. On the other hand, Go models, possibly including moderate frustration and heterogeneous couplings are the best tool for the study of folding mechanism. Sequence-based single-bead models are currently not particularly reliable while the extreme complication of multi-bead schemes is disproportionately high with respect to their performance. This by no means implies that this research line has to be abandoned: the folding problem will be solved only with the development of sequence-based models yielding a funnel-like, minimally frustrated landscape. Finally, the application of Ising-like models is recommended for simple proteins folding through a nucleation-condensation mechanism. However often the agreement with experimental data remains only qualitative.

After critically discussing the advantages and drawbacks of the simplified protein models currently available, we believe it is essential to underscore in general the merits of coarse-grained models with respect to atomistic ones. The simplifying approach is at the heart of modern physics and it dates back to Galileo who laid the foundations of modern mechanics deliberately neglecting friction although this force plays an essential role in almost every aspect of our everyday experience. Physicists were perfectly aware that they were studying an idealized world as it emerges from the words of Evangelista Torricelli (1608-1647):

"Qui studieremo il moto di quelli oggetti soggetti alla forza di gravità trascurando l'attrito; e se le vere palle di cannone non seguono queste leggi, loro danno: vorrà dire che non parleremo diesse".

(We will study the motion of those bodies subject to gravity, neglecting friction; and if real cannonballs do not follow these laws, so much the worse for them: we will not talk about them).

The best known advantage of coarse-grained models is that they accelerate the simulations, making relevant biological phenomena accessible. However as the CG dynamics is considerably faster than that of atomistic models, we cannot *a priori* exclude that inference and reconstruction of the true biological mechanism could be partially altered.

CG schemes however, are not just computational tricks to by-pass the limitations of current computational resources. In fact, their ability to correctly reproduce some experimental patterns shows that not all the molecular degrees of freedom are equally important. In other words, the coarse-graining approach allows to single out the relevant driving for interaction among the multitude of chemical details of macromolecules.

REFERENCES

[1] Creighton TE. Proteins: structures and molecular properties. New York, US: Freeman WH **1992**.
[2] Fersht A. Structure and mechanism in protein science: A Guide to enzyme catalysis and protein folding. New York, US: Freeman WH **1989**.
[3] Drenth J. Principles of protein X-ray crystallography. Heidelberg: Springer-Verlag **1999**.
[4] Wuthrich A. NMR of proteins and nucleic acids. New York: US, Wiley **1986**.
[5] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE. The protein data bank. *Nucl Acids Res* **2000**; 28: 235-42.
[6] Baker D, Sali A. Protein structure prediction and structural genomics. *Science* **2001**; 294: 93-96.
[7] Anfinsen CB. Principles that govern the folding of protein chains. *Science* **1973**; 20: 223-30.
[8] Shlick T. Molecular modeling and simulation: An interdisciplinary guide. New York: Springer-Verlag **2002**.
[9] Ponder JW. Case DA force fields for protein simulations adv. *Protein Chem* **2003**; 66: 27-85.

[10] Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* **1976**; 104: 59-107.
[11] Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comp Chem* **1997**; 18: 849-73.
[12] Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. A united-residue force field for off-lattice protein structure simulations.ii. parametrization of short-range interactions and determination of weights of energy terms by z-score optimization. *J Comp Chem* **1997**; 18: 874-87.
[13] Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. Prediction of protein conformation on the basis of a search for compact structures: test on avian pancreatic polypeptide. *Protein Sci* **1993**; 2: 1715-31.
[14] Liwo A, Czaplewski C, Pillardy J, Scheraga HA. Cumulant-based expressions for the multi-body terms for the correlation between local and electrostatic interactions in the united-residue force field. *J Chem Phys* **2001**; 115: 2323-47.
[15] Muñoz V, Eaton WA. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc Natl Acad Sci USA* **1999**; 96: 11311-16.
[16] Galzitskaya OV, Finkelstein AV. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc Natl Acad Sci USA* **1999**; 96: 11299-304.
[17] Bruscolini P, Pelizzola A. Exact solution of the Mu˜noz-Eaton model for protein folding. *Phys Rev Lett* **2002**; 88: 258101-4.
[18] Finkelstein AV, Ptitsyn OB. Protein physics. A course of lectures. Amsterdam NL: Academic Press **2002**.
[19] Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* **1997**; 48: 545-600.
[20] Plotkin SS, Onuchic JN. Understanding protein folding with energy landscape theory. Part I: basic concepts. *Q Rev Biophys* **2002**; 35: 111-67.
[21] Ueda Y, Taketomi H, Go N. Studies on protein folding, unfolding and fluctuations by computer simulation. I the effects of specific amino acid sequence represented by specific interunit interactions. *Int J Peptide Res* **1975**; 7: 445-59.
[22] Go N, Abe H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I formulation. *Biopolymers* **1981**; 20: 991-1011.
[23] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Bio* **1998**; 277: 985-94.
[24] Chiti F, Taddei N, White PM, *et al*. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct Biol* **1999**; 6: 1005-9.
[25] Koga N, Takada S. Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J Mol Biol* **2001**; 313: 171-80.
[26] Zimm BH, Bragg JK. Theory of the phase transition between helix and random coil in polypeptide chains. *J Chem Phys* **1959**; 31: 526-33.
[27] Muñoz V. What can we learn about protein folding from Ising-like models? *Curr Opin Struct Biol* **2001**; 11: 212-16.
[28] Wako H, Saito N. Statistical mechanical theory of the protein conformation I. General considerations and the application to homopolymers. *J Phys Soc Jpn* **1978**; 44: 1931-38.
[29] Muñoz V, Henry ER, Hofrichter J, Eaton WA. A statistical mechanical model for β-hairpin kinetics. *Proc Natl Acad Sci USA* **1998**; 95: 5872-9.
[30] Alm E, Baker D. Prediction of protein folding mechanisms from free-energy landscapes derived from native structures *Proc Natl Acad Sci USA* **1999**; 96: 11305-310.
[31] Weikl TR, Palassini M, Dill KA. Cooperativity in two-state protein folding kinetics. *Protein Sci* **2004**; 13: 822-29.
[32] Cieplak M, Banavar JR, Maritan A. What one can learn from experiments about the elusive transition state? *Protein Sci* **2004**; 13: 2446-57.
[33] Garbuzynskiy SO, Finkelstein AV, Galzitskaya OV. Outlining folding nuclei in globular proteins. *J Mol Biol* **2004**, 336: 509-25.
[34] Bruscolini P, Cecconi F. Analysis of Pin1 WW-domain through a simple statistical mechanics model of protein. *Biophys Chem* **2005**; 115: 153-8.
[35] Plischke M, Bergersen B. Equilibrium statistical physics. Singapore: World Scientific **1989**.
[36] Bruscolini P, Cecconi F. Mean-field approach for a statistical mechanical model of proteins. *J Chem Phys* **2003**; 119: 1248-56.
[37] Muñoz V, Thompson PA, Hofrichter J, Eaton WA. Folding dynamics and mechanism of β-hairpin formation. *Nature* **1997**; 390: 196-9.
[38] Guerois R, Serrano L. The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J Mol Biol* **2000**; 304: 967-82.
[39] Shakhnovich EI. Proteins with selected sequences fold in to unique native conformation. *Phys Rev Lett* **1994**; 72: 3907-10.
[40] Ferreiro DU, Walczak AN, Komives EA, Wolynes PG. The energy landscapes of repeat-containing proteins: topology, cooperativity and the folding funnels of one-dimensional architectures. *PLoS Comput Biol* **2008**; 4: e1000070.
[41] Lau KF, Dill KA. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **1989**; 22: 3986-97.
[42] Lau KF, Dill KA. Theory of protein mutability and biogenesis. *Proc Natl Acad Sci USA* **1990**; 87: 638-42.
[43] Yue K, Dill KA. Sequence-structure relationships in proteins and copolymers. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **1993**; 48: 2267-78.
[44] Chan HS, Dill KA. Sequence space soup of proteins and copolymers. *J Chem Phys* **1991**; 95: 3775-87.
[45] Chan HS, Dill KA. Compact polymers. *Macromolecules* **1989**; 22: 4559-73.
[46] Chan HS, Dill KA. Origins of structure in globular proteins. *Proc Natl Acad Sci USA* **1990**; 87: 6388-92.
[47] Shortle D, Chan HS, Dill KA. Modeling the effects of mutations on the denatured states of proteins. *Protein Sci* **1992**; 1: 201-15.
[48] Stillinger FH, Head-Gordon T, Hirshfeld CL. Toy model for protein folding. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **1993**; 48: 1469-77.
[49] Irback A, Potthast F. Studies of an off-lattice model for protein folding: sequence dependence and improved sampling at finite temperature. *J Chem Phys* **1995**; 103: 10298-305.
[50] Torcini A, Livi R, Politi A. A dynamical approach to protein folding. *J Biol Phys* **2001**; 27: 181-203.
[51] Bongini L, Livi R, Politi A, Torcini A. Thermally activated processes in polymer dynamics. *Phys Rev E Stat Nonlin Soft Matter Phys* **2003**; 68: 061111.
[52] Bongini L, Livi R, Politi A, Torcini A. Exploring the energy landscape of model proteins: a metric criterion for the determination of dynamical connectivity. *Phys Rev E Stat Nonlin Soft Matter Phys* **2005**; 72: 051929.
[53] Skolnick J, Kolinski A. Monte Carlo studies on equilibrium globular protein folding. II β-barrel globular protein models. *Biopolymers* **1989**; 28: 1059-95.
[54] Honeycutt JD, Thirumalai D. The nature of the folded states of globular proteins. *Biopolymers* **1992**; 32: 695-709.
[55] Guo Z, Thirumalai D. Kinetics of protein folding: nucleation mechanism, time scales and pathways. *Biopolymers* **1995**; 36: 83-102.
[56] Guo Z, Brooks-III CL. Thermodynamics of protein folding: a statistical mechanical study of a small all-β protein. *Biopolymers* **1997**; 42: 745-57.
[57] Nymeyer H, Garcia AE, Onuchic JN. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc Natl Acad Sci USA* **1998**; 95: 5921-8.
[58] Miller MA, Wales DJ. Energy landscape of a model protein. *J Chem Phys* **1999**; 111: 6610-6.
[59] Sorenson JM, Head-Gordon T. Re-designing the hydrophobic core of a model β-sheet protein: destabilizing traps through a threading approach. *Proteins: Struc Func Genet* **1999**; 37: 582-91.
[60] Brown S, Fawzi NJ, Head-Gordon T. Coarse-grained sequences for protein folding and design. *Proc Natl Acad Sci USA* **2003**; 100: 10712-7.
[61] Yap EH, Fawzi NL, Head-Gordon T. A coarse-grained α-Carbon protein model with anisotropic hydrogen bonding. *Proteins Struct Func Bioinf* **2008**; 70: 626-638.
[62] Cecconi F, Guardiani C, Livi R. Testing simplified protein models of the hPin1 WW domain. *Biophys J* **2006**; 91: 694-704.
[63] Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state

ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* **2000**; 298: 937-53.

[64] Kaya H, Liu Z, Chan HS. Chevron behaviour and isostable enthalpic barriers in protein folding: successes and limitation of simple Go-like modeling. *Biophys J* **2005**; 89: 520-35.

[65] Northey JGB, Di-Nardo AA, Davidson AR. Hydrophobic core packing in the SH3 domain folding transition state. *Nat Struct Biol* **2002**; 9: 126-30.

[66] Northey JGB, Korzhnev DM, Stogios PJ, Zarrine-Afsar A, Kay LE, Davidson AR. Dramatic acceleration of protein folding by stabilization of nonnative backbone conformation. *Proc Natl Acad Sci USA* **2004**; 101: 7954-9.

[67] Cecconi F, Guardiani C, Livi R. Stability and kinetic properties of C5-domain from myosin binding protein c and its mutants. *Biophys J* **2008**; 94: 1403-11.

[68] Guardiani C, Cecconi F, Livi R. Computational analysis of folding and mutation properties of c5 domain from Myosin Binding Protein C. *Proteins: Struct Func Bioinf* **2008**; 70: 1313-22.

[69] Cecconi F, Guardiani C, Livi R. Analyzing pathogenic mutations of C5 domain from cardiac myosin binding protein c through md simulations. *Eur Biophys J* **2008**; 37: 683-91.

[70] Gu H, Kim D, Baker D. Contrasting roles for symmetrically disposed β-turns in the folding of a small protein. *J Mol Biol* **1997**; 274: 588-96.

[71] Kuszewski J, Clore GM, Gronenborn AM. Fast folding of a prototypic polypeptide: the immunoglobulin binding domain of streptococcal protein g. *Protein Sci* **1994**; 3: 1945-52.

[72] Karanicolas J, Brooks-III CL. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci* **2002**; 11: 2351-61.

[73] Sutto L, Tiana G, Broglia RA. Sequence of events in folding mechanism: beyond the Gøo-model. *Protein Sci* **2006**; 15: 1638-52.

[74] Matysiak S, Clementi C. Optimal combination of theory and experiment for the characterization of the protein folding landcsape of S6: how far can a minimalist model go? *J Mol Biol* **2004**; 343: 235-48.

[75] Czaplewski C, Rodziewicz S, Liwo A, Ripoll DL, Wawak RJ, Scheraga HA. Molecular simulation study of cooperativity in hydrophobic association. *Protein Sci* **2000**; 9: 1235-45.

[76] Cheung MS, Garcia AE, Onuchic JN. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc Natl Acad Sci USA* **2002**; 99: 685-90.

[77] Hummer G, Garde S, Garcia S, Paulaitis AE, Pratt LR. The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *Proc Natl Acad Sci USA* **1998**; 95: 1552-55.

[78] Hummer G, Garde S, Garcia S, Pohorille E, Pratt LR. An information theory model of hydrophobic interactions. *Proc Natl Acad Sci USA* **1996**; 93: 8951-55.

[79] Hillson N, Onuchic JN, Garcia AE. Pressure-induced protein folding/unfolding kinetics. *Proc Natl Acad Sci USA* **1999**; 96: 14848-53.

[80] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**; 22: 2577-637.

[81] Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* **1985**; 18: 534-52.

[82] Hill L. Statistical Mechanics: Principles and Applications. New York, USA: Dover Publications **1987**.

[83] Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. *Curr Opin Struc Biol* **1996**; 6: 195-209.

[84] Zhou H, Zhou Y. Distance-Scaled finite ideal-gas reference state improves structure derived potentials of mean-force for structure selection and stability prediction. *Protien Sci* **2002**; 11: 2714-26.

[85] Park B, Levitt M. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J Mol Biol* **1996**; 258: 367-92.

[86] Wallqvist A, Ullner M.A simplified amino acid potential for use in structure predictions of proteins. *Proteins: Struct Func Genet* **1994**; 18: 267-80.

[87] Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **1996**; 256: 623-44.

[88] Covell DG, Jernigan RL. Conformations of folded proteins in restricted spaces. *Biochemistry* **1990**; 29: 3287-94.

[89] Miyazawa S, Jernigan RL. Protein stability for single substitution mutants and the extent of local compactness in the denatured state. *Protein Eng* **1994**; 7: 1209-20.

[90] Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol* **1994**; 234: 668-82.

[91] Covell DG. Folding protein α-carbon chains into compact forms by monte carlo methods. *Proteins: Struct Func Genet* **1992**; 14: 409-20.

[92] Kolinski A, Skolnick J. Monte Carlo simulation of protein folding. II. Application to protein A, ROP and crambin. *Proteins: Struct Func Genet* **1994**; 18: 353-66.

[93] Nozaki Y, Tanford C. The solubility of amino acids and related compounds in aqueous ethylene glycol solution. *J Biol Chem* **1965**; 240: 3568-73.

[95] Fain B, Xia Y, Levitt N. Design of an optimal Chebyshev-expanded discrimination function for globular proteins. *Protein Sci* **2002**; 11: 2010-21.

[96] Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins: Struc Func Genet* **1994**; 18: 338-52.

[97] Kolinski A, Skolnick J. Reduced models of proteins and their applications. *Polymer* **2004**; 45: 511-24.

[98] Kresse HP, Czubayko M, Nyakatura G, Vriend G, Sander C, Bloecker H. Four-helix bundle topology re-engineered: monomeric Rop protein variants with different loop arrangements. *Protein Eng* **2001**; 14: 897-901.

[99] Kolinski A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* **2004**; 51: 349-371.

[100] Kmiecik S, Kolinski A. Characterization of protein-folding pathways by reduced-space modeling. *Proc Natl Acad Sci USA* **2007**; 104: 12330-35.

[101] Lee J, Ripoll DR, Czaplewski C, Pillardy J, Wedemeyer WJ, Scheraga HA. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. *J Phys Chem B* **2001**; 105: 7291-8.

[102] Liwo A, Khalili M, Scheraga HA. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc Natl Acad Sci USA* **2005**; 102: 2362-7.

[103] Berne BJ, Pechukas P. Gaussian model potentials for molecular interaction. *J Chem Phys* **1972**; 56: 4213.

[104] Gay JG, Berne BJ. Modification of the overlap potential to mimic a linear site-site potential. *J Chem Phys* **1981**; 74: 3316.

[105] Vorobiev YN. Block-units method for conformational calculations of large nucleic acid chains. I. block-units approximation of atomic structure and conformational energy of poly-nucleotides. *Biopolymers* **1990**; 29: 1503-18.

[106] Nishikawa K, Momany FA, Scheraga HA. Low-energy structures of two dipeptides and their relationship to bend conformations. *Macromolecules* **1974**; 7: 797-806.

[107] Liwo A, Oldziej S, Ciarkowski J, Kupryszewski G, Pincus MR, Scheraga HA. Prediction of conformation of rat galanin in the presence and absence of water with the use of monte carlo methods and the ECEPP/3 force field. *Protein Chem* **1994**; 13: 375-80.

[108] Liwo A, Kazmierkiewicz R, Czaplewski C, *et al*. United-residue force field for off-lattice protein-structure simulations: Iii. origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *J Comp Chem* **1998**; 19: 259-76.

[109] Kubo R. Generalized cumulant expansion method. *J Phys Soc Jpn* **1962**; 17: 1100-20.

[110] Godzick A, Kolinski A, Skolnick J. *De novo* and inverse folding predictions of protein structure and dynamics. *J Comput-Aid Mol Des* **1993**; 7: 397-438.

[111] Skolnick J, Kolinski A, Ortiz AR. Monsster: a method for folding globular proteins with a small number of distance restraints. *J Mol Biol* **1997**; 265: 217-41.

[112] Pillardy J, Czaplewski C, Liwo A, *et al*. Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc Natl Acad Sci USA* **2001**; 98: 2329-33.

[113] Oldziej S, Czaplewski C, Liwo A, *et al.* Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proc Natl Acad Sci USA* **2005**; 102: 7547-52.

[114] Czaplewski C, Kalinowski S, Liwo A, Scheraga HA. Application of multiplexed replica exchange molocular dynamics to the UNRES force field: tests with α and α+β proteins. *J Chem Theor Comput* **2009**; 5: 627-40.

[115] Irback A, Sjunnesson F, Wallin S. Three-helix bundle protein in a Ramachandran model. *Proc Natl Acad Sci USA* **2000**; 97: 13614-18.

[116] Irback A, Sjunnesson F, Wallin S. Hydrogen bonds, hydrophobicity forces and the character of the folding transition. *J Biol Phys* **2001**; 27: 169-79.

[117] Favrin G, Irback A, Sjunnesson F. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Proteins* **2002**; 47: 99-105.

[118] Regan L, De-Grado WF. Characterization of a helical protein designed from first principles. *Science* **1988**; 241: 976-8.

[119] Vitkup D, Melamud E, Sander C. Completeness in structural genomics. *Nat Struct Biol* **2001**; 559-66.

[120] Gonzalez C, Langdon GM, Bruix M, *et al*. Bacteriocin AS-48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin. *Proc Natl Acad Sci USA* **2000**; 97: 11221-26

[121] Obmolova G, Ban C, Hsieh P, Yang W. Crystal structures of mismatch repair protein MutS and its complex with a substrate. *DNA Nat* **2000**; 407: 703-710.

[122] Muñoz V, Serrano L. Elucidating the folding problem of helical peptides using empirical parameters. *Nat Struct Biol* **1994**; 1: 399-409.

[123] Karanicolas J, Brooks III CL. The importance of explicit chain representation in protein folding models: an examination of Ising-like models. *Prot Struct Funct Genet* **2003**; 53: 740-747.

[124] Dill KA, Fiebig KM, Chan HS. Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci USA* **1993**; 90: 1942-46.

[125] Baker D. A surprising simplicity to protein folding. *Nature* **2000**; 405: 39-42.

[126] Hills RD, Brooks III CL. Subdomain competition, cooperativity, and topological frustration in the folding of CheY. *J Mol Biol* **2008**; 382: 485-95.

[127] Kleiner A, Shakhnovich E. The mechanical unfolding of ubiquitin through all-atom Monte Carlo simulation with a Go-type potential. *Biophys J* **2007**; 92: 2054-61.

[128] Li MS, Kouza M, Hu CK Refolding upon force quench and pathways of mechanical and thermal unfolding of ubiquitin. *Biophys J* **2007**; 92: 547-61.

[129] Hyeon C, Onuchic JN. Internal strain regulates the nucleotide binding site of the kinesin leading head. *Proc Natl Acad Sci USA* **2007**; 104: 2175-80.

[130] Homouz D, Perham M, Samiotakis A, Cheung MS, Wittung-Stafshede P. Crowded, cell-like environment induces shape changes in aspherical protein. *Proc Natl Acad Sci USA* **2008**; 105: 11754-59.

[131] Griffin MA, Friedel M, Shea JE. Effects of frustration, confinement and surface interactions on the dimerization of an off-lattice beta-barrel protein. *J Chem Phys* **2005**; 123: 174707.

[132] Neri M, Baaden M, Carnevale V, Anselmi C, Maritan A, Carloni P. Microseconds dynamics simulations of the outer-membrane protease T. *Biophys J* **2008**; 94: 71-78.

[133] Paci E, Vendruscolo M, Karplus M. Validity of Go models: comparison with a solvent-shielded empirical energy decomposition. *Biophys J* **2002**; 83: 3032-38.

[134] Ferreiro DU, Hegler JA, Komives EA, Wolynes PG. Localizing frustration in native proteins and protein assemblies. *Proc Natl Acad Sci USA* **2007**; 104: 19819-24.

[135] Clementi C, Plotkin SS. The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci* **2004**; 13: 1750-66.

[136] Viguera AR, Vega C, Serrano L Unspecific hydrophobic stabilization of folding transition states. *Proc Natl Acad Sci USA* **2002**; 99: 5349-54.

[137] Khare SD, Ding F, Dokholyan NV. Folding of Cu-Zn superoxide dismutase and familial amyotrophic lateral sclerosis. *J Mol Biol* **2003**; 334: 515-25.

[138] Luo ZL, Ding JD, Zhou YQ. Folding mechanism of individual beta-hairpins in a Go model of Pin1 WW domain by all-atom molecular dynamics simulations. *J Chem Phys* **2008**; 128: 225103.

[139] Chan HS, Dill KA. Transition states and folding dynamics of proteins and heteropolymers. *J Chem Phys* **1994**; 100: 9238-58.

[140] Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA* **1995**; 92: 325-329.

[*] The single sequence approximation prescribes to consider only conformations with a contiguous stretch of native bonds, thus significantly narrowing the conformational space. However, this approximation has been shown to underestimate free energy barriers in Ref. [15] and Ref. [17].

Received: June 03, 2009 Revised: November 24, 2009 Accepted: February 08, 2010