Available online at www.sciencedirect.com

ScienceDirect

Physics Procedia 8 (2010) 63–72

Physics Procedia

www.elsevier.com/locate/procedia

VI Encuentro Franco-Español de Química y Física del Estado Sólido
VI[ème] Rencontre Franco.Espagnole sur la Chinie et la Physique de l'État Solide

New Set of 2D/3D Thermodynamic Indices for Proteins. A Formalism Based on "*Molten Globule*" Theory

Ruiz-Blanco Yasser B.[1], García Y.[2*], Sotomayor-Torres C.M.[2,3] and Marrero-Ponce Yovani [1,4]

1.Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.
2.Institut Català de Nanotecnologia - Centre d'Investigació en Nanociència i Nanotecnologia (ICN-CSIC), Campus UAB, Edifici CM3, 08193 Barcelona, Spain
3.Catalan Institute for Research and Advanced Studies, ICREA, 08010 Barcelona, Spain
4.Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain.

Abstract

We define eight new macromolecular indices, and several related descriptors for proteins. The coarse grained methodology used for its deduction ensures its fast execution and becomes a powerful potential tool to explore large databases of protein structures. The indices are intended for stability studies, predicting Φ-values, predicting folding rate constants, protein QSAR/QSPR as well as protein alignment studies. Also, these indices could be used as scoring function in protein-protein docking or 3D protein structure prediction algorithms and any others applications which need a numerical code for proteins and/or residues from 2D or 3D format.

Keywords: protein folding descriptor, protein stability, protein indices, FPI, folding degree.

1. Introduction

Proteins are the building blocks of life [1,2]. Catalysis of biochemical reactions, transport, and recognition - among other biological functions - are impossible without the active participation of proteins. The 3D structure of proteins determines all these functions [2]. This fact has been called the second genetic code [2,3]. Practical applications of this knowledge to vast fields, such as biotechnology or pharmaceutical sciences, have produced a

* Corresponding author. Tel.: +34 93 586 8312; fax: +34 93 586 8313.
E-mail address: ygarcia@icn.cat.

1875-3892
doi:10.1016/j.phpro.2010.10.013

variety of computational methods along with significant improvement of the potentials to be minimized in order to reach the structure of the proteins and also learn more about the protein folding problem [4-7]. Many comprehensive review articles are available in the literature [8-17].

The rugged landscape topography of the potential energy surface and its high dimensionality are responsible for the underlying difficulties in solving the native structure of proteins. A large variety of methods have been considered to overcome -at least partially- these difficulties. For instance, homology modeling (sequence alignment) and use of efficient techniques among which, neural networks and genetic algorithms can be mentioned [18].

The development of a fast and reliable protein force-field is a complex task given the delicate balance between the different energy terms that contribute to protein stability [19,20]. Many different force-fields have been constructed for predicting protein stability changes. These range from force-fields based on pure statistical analysis of structural sequence preferences [9,21-24] and force-fields based on multiple sequence alignments [25-27], to molecular dynamics inspired force-fields [28,29]. As a matter of rigor appeal, force-fields can be divided into three major categories: (i) those using a physical effective energy function [30,31]; (ii) those based on statistical potentials for which energies are derived from the frequency of residue or atom contacts in the protein database [16, 32-36]; and (iii) those using empirical data obtained from experiments run on proteins [37].

In the present report, we propose a coarse grained (smooth) physical effective energy function which we have called Folding Potential Index (FPI) together with several related indices and descriptors.

Among the main features introduced in our indices is the more explicit consideration of the entropy of the first layer of water molecules near the protein. This factor, together with the solvation energy, have been fused and described commonly by an empirical term having the next general form: $G^S = \gamma * A$ [32] where G^S represents the superficial free energy, γ is an empiric factor and A is usually the Accessible Surface Area (ASA) [38,39]. In our opinion, this approach underestimates the effect of the interfacial water molecule's entropy, responsible of the called *hydrophobic effect*, which is of major importance for the folding process's spontaneity [40] and its pathway. The *hydrophobic effect* [40-44] is defined as the entropy increase, due to the minimization of the number of ordered water molecules required to surround the hydrophobic portions of a protein during the folding process, and it is responsible of its thermodynamic justification. Here, our goal is to obtain fast and reliable indices with a clear physical interpretation that improve the actual thermodynamic and kinetic description of the folding process.

2. Theoretical scaffold

The following sections explain the theoretical basis of our indices. First, a series of postulates are established in order to formalize the descriptors associated to every contributing factor to the folding free energy. Then, our indices are built combining the previously defined descriptors. Finally, a global free energy function and other related global indices as well as their local versions are obtained.

2.1 Theory

Nowadays, two main theories have been developed in order to explain protein folding. First, is the classical theory. Basically, this theory supposes that protein's stability is governed by the multiple interactions among residues in the structure [45, 46]. The second is the known as molten globule theory [45, 47]. In addition, this theory considers the effect of the solvent and the hydrophobic interactions in the stability of the folded structure. Molten globule theory establishes that the hydrophobic residues of a protein chain suffer a spontaneous collapse into a compact state which has part of the secondary structure motives already formed, the tertiary structure is not in its native form yet and the hydrophobic core is still partially exposed. This intermediate structure derives in the native state by an optimization process guided mainly by local interactions. Similarly, a theory exposed by Kauzmann (1959) [48] suggests that the interior of a globular protein is similar to an oil-drop. This model confirms that hydrophobic interactions play a crucial role in folding and stability of three dimensional protein structures.

Following this idea we have developed a physical model that allows a mathematical formalization of the molten globule theory. Our model relates the protein folding with colloidal and superficial chemistry's principles, based on two main analogies. On one hand, the hydrophobic residue nucleation and coalescence (Figure 1A) and on the other hand, the distribution of polar / non polar residues in proteins and formation of micelles (Figure 1B).

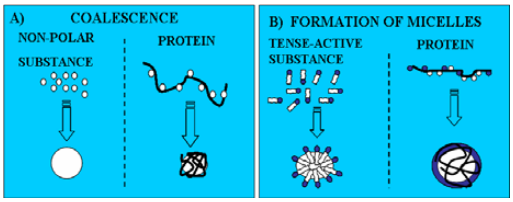

Figure 1. A) Schematic representation of coalescence and hydrophobic residue nucleation. B) Representation of the formation of micelles and the polar / non polar residue distribution in proteins. In white are represented the hydrophobic zones and in dark blue the hydrophilic ones.

As is well known, multiple faces of a non polar substance in a polar medium suffer the phenomenon called coalescence, where the diverse faces of an hydrophobic substance fuse into a larger one. This fact is justified by a reduction of the interfacial area between the polar and non polar substances. In consequence, a minimization of the superficial free energy, $\Delta G^S = \sigma A$, is achieved. Here, σ and A represent the superficial tension and the superficial area, respectively. In proteins, as shown in Figure 1A, we can see a similar effect when most of the hydrophobic residues aggregate into a core isolated from the polar solvent –e.g. water-, producing the corresponding decrease in the solvent accessible surface area.

Also it is widely known that when molecules of a tense-active substance are present in a polar medium, they suffer an aggregation process, forming the structures called micelles [40]. In this particular arrangement, they orient their hydrophobic chains toward the core of the micelle, while its hydrophilic part remains at the surface allowing a decrease in the superficial tension and therefore, in the superficial free energy. In the same way, when globular proteins fold, they selectively leave most of the hydrophilic residues on the surface (toward the water -polar- face), while most hydrophobic residues are addressed into the protein, forming a hydrophobic core (see Figure 1B). As a result, the superficial tension decays, and stabilizes the disperse system consisting on the protein and the aqueous medium.

Notice that taking into account these analogies it is possible to accept that mainly those factors that minimize the interfacial free energy potential of a heterogeneous system (to know *interfacial entropy,* S_{H2O}, and *solvation energy,* H_W) are responsible for folding and for stabilizing 3D-protein structures. *This statement represents the basic hypothesis of our model*. Consequently, during the folding process, folded structures that do not contribute to the gradient of an hypothetical hydrophobic potential are not explored, causing a funnel like effect [49].

However, there are two other factors that we should take into account. These are the *energetic contribution of the disulfide bridges* (H_{DS}) and the *conformational entropy* (S_{conf}). The disulfide bonds represent the unique bonding interactions that appear during the folding process. Therefore, its contribution to the global energetic balance must be significant. The conformational entropic cost of folding a protein is also an important factor due to the evident lost of freedom degrees going from a highly mobile unfolded state to a compact and more rigid native state. This factor has a negative effect in folding. On the other hand, the interaction between superficial residues is neglected in the present approach, because as have been probed by mutagenesis experiments the electrostatic interaction's energy between mobile side-chain on the surface of a protein is offset by the entropic cost of immobilizing or restricting the motion of the interacting partners [47,50]. In the same way we ignore the contribution of the H-bonds because "for every H-bond that is formed in the folded structure one have been broken in the unfolded state between the protein

and the surrounding water molecules" [45]. Other kind of interactions we assume that are weaker enough to be discarded. Despite the important reduction in the free energy potential, our model maintains the principal factors we think that have direct and significant incidence over the protein folding. Following these ideas the folding free energy can be expressed as follows:

$$G = H_{DS} - TS_{conf} + H_W - TS_{H_2O} \quad (1)$$

The mathematical expression for each term is obtained based in some postulates described in the next section. Before presenting the postulates, it is necessary to define several relevant concepts. The first one is referred to as constrained pair of residues. These constrained residues are defined as all pairs of residues for which the quotient between their topological and spatial (between the α-carbons) distances would be larger than the ratio of these values in an unfolded state. We introduce a delta (δ^C) in the formalism which takes value 1 if a pair of residues is considered a constrained pair or 0 if it isn't.

The second one is referred to as the superficial residues. Several methods have been implemented to approach the accessible surface area of Lee and Richards (Lee et al., 1971) [38,39,51], which in general rolls a sphere of average solvent radius over the surface of a macromolecule (see Figure 2). The solvent probe's boundary traces out the molecular surface, while its center traces out the solvent accessible surface. To obtain the solvent accessible surface area and the superficial residues we use the method of Hasel et. al. [52], also implemented by David et. al. 2009 [53]. The sum of the areas of every atom in a residue represents the superficial area of the amino acid in a folded state (A^F). The superficial residues are identified taking a cutoff value for the accessible surface area of an atom. If the superficial area of an amino acid is larger than the product of this cutoff by the number of atom of the corresponding residue, then it is classified as superficial.

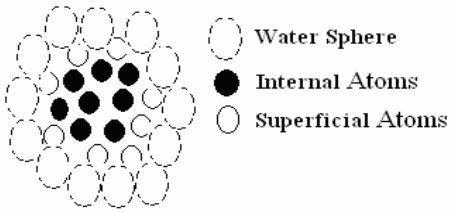

Figure 2. Representation of the algorithm to determine superficial residues.

Another important concept is what we called *vibrational subspaces*. To understand this concept we must think about the whole set of vibrational movements of a protein as a vectorial space with $3N^a$-6 dimensions, being N^a the number of atoms in the protein. A basis set of this vectorial space would be all the normal vibrational modes (NVM) of the protein. Then any combination of one or more different NVM will generate a given subspace. Thus, the obtained subspaces are named *vibrational subspaces*.

Finally, we introduce the concept, *hydrophilic superficial group* (HSG), to group all the residues which obey the next rules:

- The residues that belong to a HSG must be superficial and hydrophilic.
- In the unfolded state the HSGs are taken as sets of hydrophilic residues with no more than one hydrophobic amino acid between two hydrophilic ones.
- In a folded state, the distance between the α-carbons of any residue and another one of the same HSG must be equal or smaller than 7.3 Å.

2.2 Postulates

In order to obtain mathematical expressions for all the terms of the free energy equation (Eq. **3**), we established a series of postulates for the enthalpy, entropy and solvation energy [32].

First claim: Postulate for the free energy contribution of disulfide bridges.

- All the disulfide bridges are considered energetically equivalent.

$$H_{DS} = \sum_{1}^{P} \sum_{i=1}^{N} \delta_i^{DS} \quad (2)$$

Where, δ^{DS} takes value -56.87 (half of the free energy variation of the reaction: 2R-SH + 1/2O_2 = R-S-S-R + H_2O in kJ/mol under standard conditions), if the residue (cysteine) is involved in a disulfide bridge or 0 if it is not. N is the number of residues in a protein chain, and P is number of protein's subunits. Therefore, the contribution of all the disulfide bridges can be expressed as proportional to the total number of these bonds.

Second claim: Postulates for the conformational entropy

This claim is crucial in our model. In proteins, due to the large number of degrees of freedom -in contrast with simple chemical reactions- entropic contributions are comparable, or even larger than energetic ones. Thus, energy alone does not determine the direction or path of the folding process.

- The conformational entropy (S_{conf}) is proportional to the logarithms of the number of *accessible vibrational subspaces* (W_{vib}) of the protein, (in analogy with Bolztmann's definition of entropy):

$$S_{conf} = R \ln W_{vib} \quad (3)$$

- The ratio between the number of *accessible vibrational subspaces* of a folded and the unfolded state follows the Bolztmann distribution's law with the increase of the *constrained pair of residues*:

$$\frac{W_{vib}}{W_{vib}^{0}} = e^{\frac{-\sum_{i=1}^{N}\sum_{j\geq i+l}^{N}\frac{\delta_{ij}^{c}(j-i)}{d_{ij}}}{RT}} \quad (4)$$

- The total number of *accessible vibrational subspaces* of an unconstrained protein chain state will be the number of all the combinations of one or more different NVM:

$$W_{vib}^{0} = \sum_{x=1}^{3N^a-6} \frac{(3N^a-6)!}{(3N^a-6-x)!x!} = 2^{3N^a-6} - 1 \quad (5)$$

- For several protein subunits, the total number of *vibrational subspaces* would be the product of the corresponding W_{vib} of every single protein chain. Finally the conformational entropy is formalized as follow:

$$S_{conf} = R\left[\ln\prod_{1}^{P} W_{vib}^{0} - \frac{\sum_{1}^{P}\sum_{i=1}^{N}\sum_{j\geq i+l}^{N}\frac{\delta_{ij}^{c}(j-i)}{d_{ij}}}{RT}\right] \quad (6)$$

Where,

$$\delta_{ij}^{C} = \begin{cases} 1 & if \ \frac{\sqrt[k]{|j-i|}}{d_{ij}} \geq \frac{\sqrt[k]{l}}{d_0} \\ 0 & otherwise \end{cases} \quad (7)$$

$$\ln\prod_{1}^{P} W_{vib}^{0} = \sum_{1}^{P} \ln(2^{3N^a-6} - 1) \cong 2\sum_{1}^{P}\sum_{i=1}^{N} N_i^a \quad (8)$$

and k represents an integer, R is the gas constant, (j-i) is the topological (sequential) distance between residues i and j, d_{ij} accounts for spatial distances between a pair of residues, d_0 is the cutoff of the spatial distance, l the cutoff of the topological distance, N_i^a the number of atoms of every residue in the protein, and T is the absolute temperature.

When the folding process begins the number of *accessible vibrational subspaces* decrease with the increasing of the number of *constrained pairs of residues* because of the large number of amino acids "forced" to adopt a particular conformation. We considered that the number of *accessible vibrational subspaces* follows an exponential decay due to the cooperative effect during the folding process, which produces a major decay at the beginning than

in the last steps of the folding. This assumption is consistent with the theory of the earlier hydrophobic collapse (molten globule).

Third claim: Solvation energy's postulates.

- The solvation energy is proportional to the number of the superficial residues.
- The contribution of every residue is represented by the product of its hydrophobicity and the corresponding surface area.

In accordance with these two postulates, we define the solvation energy in the following way:

$$H_W = \sum_{1}^{P} \sum_{i=1}^{N} \varphi_i A_i \delta_i^S \quad (9)$$

Where φ, A, and δ^S, are the hydrophobicity of a particular residue, surface area and a delta –named superficiality- that takes value 1 if the residue is superficial or 0 if it is not. There are several scales of hydrophobicity defined in previous studies [54, 58-61]. We chose the Kyte- Doolittle scale (KDS) [60] since it obeys the logical monotony and sign required for our formalism. As a measure of the surface area of the residues in the unfolded state (A^U), we built a new descriptor called Side-Chain Surface Area (SSA). This descriptor is the sum, for each residue, of the Isotropic Surface Area (ISA)[54], and the Polar Surface Area (PSA)[62] of every amino acid.

This term has been widely used in many force fields and indices that adopted an implicit solvation model to account with energetic and entropic phenomena that occur in the water-protein interface[32]. There are several authors that have some discrepancies about employing this kind of solvation model to address the folding problem [55-57]. We consider that in protein systems, the use of a continuum solvent model only fairly describes the energetic effects (solvation energy, H_W) but underestimate the entropic ones. The next is probably the most remarkable term of our indices because it introduces the contribution of the *configurational entropy* of the water molecules on the surface of the protein (S_{H2O}), improving the implicit solvation model without the explicit consideration of water molecules.

Fourth claim: Postulate for the interface water molecule's entropy.

- The water molecules associated to a HSG can be interchanged among all the members of the group, keeping each residue a constant number of linked water molecules. Therefore, it is generated by a given number of combinations of water molecules (W_j).
- The entropic change associated with the water molecules in the interface of the protein is proportional to the logarithm of the product of the number of combinations of the water molecules in all the hydrophilic superficial groups.

Taken into account this consideration the interface water molecule's entropic contribution is defined according to the Bolztmann's equation for the entropy: $S = k\, ln\omega$. Thus, it takes the form that follows:

$$S_{H2O} = R \ln \prod_{j=1}^{n} W_j \quad (10)$$

Where,

$$W_j = \frac{N_j^w!}{\prod_{i=1}^{p} N_{ij}^w!} \quad (11)$$

and,

$$N_j^w = \sum_{i=i}^{p} N_{ij}^w \quad (12)$$

The parameters N^w_j, N^w_{ij}, W_j, n and p are; the reduced number of the water molecules associated to a hydrophilic superficial group, the reduced number of water molecules associated to every residue, the number of combinations of the water molecules of a hydrophilic superficial group, the number of HSGs, and the number of residues in a HSG, respectively.

2.3. Definition of New Indices

Folding Potential Index (FPI): This is the main index proposed in the present study, and it is thought as a measure of the free energy of a particular state of a protein. Then, following Eq. 1, and according with the previous definition of each term in this equation, the **FPI** descriptor can be formalized as follows:

$$FPI = \sum_{1}^{P}\sum_{i=1}^{N}\delta_i^{DS} - TR\left[\ln\prod_{1}^{P}W_{vib}^{0} - \frac{\sum_{1}^{P}\sum_{i=1}^{N}L_i}{RT}\right] + \sum_{1}^{P}\sum_{i=1}^{N}\varphi_i A_i^F \delta_i^S - RT\ln\prod_{j=1}^{n}W_j^F \quad (13)$$

Where,

$$L_i = \sum_{j\geq i+l}^{N}\frac{\delta_{ij}^c(j-i)}{d_{ij}} \quad (14)$$

Unfolded State Index (USI): This index gives a description of the unfolded state of a protein and is obtained considering the definition given for an unfolded state (see end of theory section). Thus, it is defined as:

$$USI = -TS_{conf}^U + H_W - TS_{H2O}^U \quad (15)$$

That is:

$$USI = -TR\ln\prod_{1}^{P}W_{vib}^{0} + \sum_{1}^{P}\sum_{i=1}^{N}\varphi_i A_i^U - T\sum_{1}^{P}\left[0{,}075\left(\sum_{i=1}^{N}N_{ij}^{w}\right)\ln\frac{T}{298} + R\ln\prod_{j=1}^{m}W_j^U\right] \quad (16)$$

Where, m and W_j^U are the number of hydrophilic superficial groups and the number of combinations of the water molecules of a hydrophilic superficial group in an unfolded state, respectively.

The last term, in the present index, contains information about the contributions to the water molecule's entropy by concept of thermal variation and water molecules delocalization over the protein surface. This is a particular important index because it allows the characterization of a protein based only on its sequence.

Global Stability Index (GSI): While the **FPI** takes into account, in an absolutely way, the free energy of any state of a protein (knowing its coordinates); **GSI** provides a characterization of the stability relative to the corresponding unfolded state. Therefore, **GSI** allows comparison for the stability of different proteins:

$$GSI = \sum_{1}^{P}\sum_{i=1}^{N}\delta_i^{DS} + \sum_{1}^{P}\sum_{i=1}^{N}\sum_{j\geq i+l}^{N}\frac{\delta_{ij}^c(j-i)}{d_{ij}} + \sum_{1}^{P}\sum_{i=1}^{N}\varphi_i(\delta_i^S A_i^F - A_i^U) - T\left[R\ln\left[\frac{\prod_{j=1}^{n}W_j^F}{\prod_{1}^{P}\prod_{j=1}^{m}W_j^U}\right] - \sum_{1}^{P}0{,}075\left(\sum_{i=1}^{N}N_{ij}^{w}\right)\ln\frac{T}{298}\right] \quad (17)$$

Amino acid Contribution to the Folding Potential Index (ACFPI): This index provides the contribution to the **FPI** of every amino acid in a folded state according its nature and environment. It is obtained approaching each term of the **FPI** as a sum over all the residues. It has the following definition:

$$ACFPI_i = \delta_i^{DS} - RT\left[2N_i^a - \frac{\frac{1}{2}\sum_{j=1;\,|j-i|\geq l}^{N}\frac{\delta_{ij}^c|j-i|}{d_{ij}}}{RT}\right] + \delta_i^S\varphi_i A_i^F - \left[\frac{RT\delta_i^S N_{ij}^w\ln\left[\left(\sum_{i=1}^{p_j^F}N_{ij}^w\right)!\right]}{\sum_{i=1}^{p_j^F}N_{ij}^w} - RT\delta_i^S\ln(N_{ij}^w!)\right] \quad (18)$$

Amino Acid Contribution to the Unfolded State Index (ACUSI): This index was obtained in the same way of **ACFPI,** expressing **USI** as a sum over all the residues, the resulting expression takes the following form:

$$ACUSI_i = -2RTN_i^a + \varphi_i A_i^U - T\left(\frac{RN_{ij}^w \ln\left[\left(\sum_{i=1}^{p_j^U} N_{ij}^w\right)!\right]}{\sum_{i=1}^{p_j^U} N_{ij}^w} - R\ln(N_{ij}^w!) + 0{,}075\mathrm{N}_{ij}^w \ln\frac{\mathrm{T}}{298} \right) \quad (19)$$

Local Stability Index (LSI): The subtraction of the indices **ACFPI** and **ACUSI** results in the **LSI** index:

$$LSI_i = \delta_i^{DS} + \frac{1}{2}\sum_{\substack{j=1; \\ |j-i|\geq l}}^{N} \frac{\delta_{ij}^c |j-i|}{d_{ij}} + \varphi_i(\delta_i^S A_i^F - A_i^U) - T\left(\frac{R\delta_i^S N_{ij}^w \ln\left[\left(\sum_{i=1}^{p_j^F} N_{ij}^w\right)!\right]}{\sum_{i=1}^{p_j^F} N_{ij}^w} + R\delta_i^I \ln(N_{ij}^w!) - \frac{RN_{ij}^w \ln\left[\left(\sum_{i=1}^{p_j^U} N_{ij}^w\right)!\right]}{\sum_{i=1}^{p_j^U} N_{ij}^w} - 0.075\mathrm{N}_{ij}^w \ln\frac{T}{298} \right) \quad (20)$$

Where, δ^I takes value 1 if the residue is internal or zero if it is not ($\delta^I = 1 - \delta^S$).

This index also obeys the same dependence, with its corresponding global index (**GSI**), shown for **ACFPI** and **ACUSI.** Thus, the **GSI** can be obtained as the sum of all the **LSI** for each amino acid in the structure. Finally, in addition to global and single amino acid-level indices computed for each residue in the protein, local-fragment formalism can be developed. In this sense, group and residue-type descriptors are specific cases of local amino acid-based indices. First, it should be remarked that **ACFPI**, **ACUSI** and **LSI**, can be computed for each residue in the protein and contain intrinsic and topological structural information from all other amino acids in contact with it and within the structure. Besides, the residue-type indices can also be calculated. These novel local amino acid descriptors provide much useful information allowing it to be applied to a wide range of problems. The residue-type descriptors are calculated by adding the *k*-th residue-level indices for all amino acid of the same type in the protein. This residue-type index lends itself to use in a group additive-type scheme in which an index appears for each amino acid type in the molecule. That is to say, the global indices can be expressed as the sum of the local indices of the `z´ fragments of different residue types. For instance:

$$LSI_j = \sum_{i=1}^{k} LSI_{ij} \quad (21)$$

$$GSI = \sum_{j=1}^{z} LSI_j \quad (22)$$

In the formalism of the residue-type descriptors, each amino acid in the structure is classified according to its nature as non-polar, charged polar, non- charged polar, aromatic, aliphatic, their own ID and so on. The residue-type descriptors combine three important aspects of structural information: 1) Topologic accessibility to the residues of the same type, 2) presence/absence of the residue type and 3) count of the amino acid in the residue-type sets. Moreover, these local amino acid descriptors can be calculated by a group in the structure. The group-level indices are the sum of the individual residue-level indices for a particular group of amino acids, e.g. residues of an active site. These two formalisms lately mentioned can provide important information for QSAR/QSPR studies.

Folding Degree (**FD**): This index depends of how much bended would be the structure over a single amino acid and is exponentially related with the conformational entropy variation:

$$FD = e^{\frac{1}{P}\sum_{1}^{P}\frac{|i-i|\geq l}{N^2}} = e^{\frac{1}{P}\sum_{1}^{P}\frac{2\Delta S_{conf}}{N^2}} = \left[\prod_{1}^{P} N\sqrt{\prod_{1}^{N} LFD_i}\right]^{1/P} e^{\frac{|i-i|\geq l}{N}} \qquad (23)$$

3. Concluding Remarks

It has been defined a set of eight thermodynamic and topological protein indices and several related descriptors (terms) for energetic and entropic contributing factors. We described two illustrative analogies that were taken as the basis to understand a protein-water system as a disperse system. Then, we interpreted the hydrophobic collapse as a hybrid of the known phenomena coalescence and formation of micelles. In one hand and in consonance with its coarse grained character, our model has some limitations. We do not consider surrounding factors like pH and ionic strength, do not include interactions with other species in the media, and depends on empirical parameters like hydrophobicity, superficial area, and the number of coordination's water molecules of every amino acid. In the other hand and by constrast, with our model we developed a comprehensive formalism introducing several postulates to formalize every term of the free energy potential (Eq. 1) of the system. With these indices, it is possible to characterize the unfolded state of a protein (**USI**) and any folded state if its coordinates are known (**FPI**, **GSI** and **FD**). In addition, we extended these global indices to their local versions (**ACUSI**, **ACPFI**, **LSI** and **LFD**). These local indices lend itself to a widespread field of applications since they can be calculated at a single residue level or at fragment level combining several of the firsts. They can be grouped by type (i.e. polar, non polar, aromatic), by spatial zone or any other criteria that the researcher can adopt. The evaluation of the descriptors requires a predefinition of several parameters needed for the calculations such as the cutoffs for the spatial and topological distances involved in the definition of the constrained pair of residues, the minimum solvent accessible surface area for atoms to determine the superficial residues, the inter-Cα distance in a HSG, etc. The users could vary them according to their experience and goals.

4. Future Outlook

The indices presented in this report will allow calculations of protein stability (ΔG_{fold}), Φ-values, association free energy in protein-protein complexes, predicting folding rate constants, protein QSAR/QSPR as well as protein alignment studies. These indices could also be used as scoring function in protein-protein docking or 3D protein structure prediction algorithms and any others applications which need a numerical code for proteins and/or residues from 2D or 3D format.

5. Coming Applications

Currently we are immersed in two main jobs in order to probe our formalism. First, the prediction of the folding rate constant using a comprehensive physic model in accordance with the theory exposed in the present report. Second, the development of an *ab initio* protein structure prediction program employing a term weighted version of the FPI as scoring function. The program for generating all the indices and descriptors described in this paper will be freely available after publishing the first application.

References

[1] P.Baldi, Bioinformatics 15 (1999) 937.
[2] Dill, Protein Sci. 4 (1995) 561.
[3] Y.Cui, Proteins: Structure, Functions Genetics 31 (1998) 247.
[4] P.S.Kim, Annu. Rev Biochem. 50 (1990)
[5] Y. D. P.A.Kollman, M.R.Lee, J. Mol. Graph. Model. 19 (2001) 146.

[6] J. R. Desjarlais and T.M. Handel, J. Mol. Biol. 290 (1999) 305.
[7] K. Leonhard, Protein Sci. (2004) 358.
[8] Skolnick, Curr. Opin. Struct. Biol. 16 (2006) 166.
[9] M.J.Sippl, Curr. Opin. Struct. Biol. 5 (1995) 229.
[10] I. B. R.L.Jernigan, Curr. Opin. Struct. Biol. 6 (1996) 195.
[11] J.Moult, Curr. Opin. Struct. Biol. 7 (1997) 194.
[12] G. K. H.Gohlke, Curr. Opin. Struct. Biol. 11 (2001) 231.
[13] R. R. W. P.Russ, Curr. Opin. Struct. Biol. 12 (2002) 447.
[14] J. E. S. N.V.Buchete, D.Thirumalai, Curr. Opin. Struct. Biol. 14 (2004) 225.
[15] R. R. A.M.Poole, Curr. Opin. Struct. Biol. 16 (2006) 508.
[16] M. K. T.Lazaridis, Curr. Opin. Struct. Biol. 10 (2000) 139.
[17] H. Z. Y.Zhou, C.Zhang S.Liu, Cell Biochem. Biophys. 46 (2006) 165
[18] J. C. Luis P.B. Scott, Jose R. Ruggiero, Applied Mathematics and Computation 195 (2008) 515.
[19] K. P. Murphy, E.Freire Advan. Protein Chem. 43 (1992) 313.
[20] S. B. A. Pace C. N., McNutt M. & Gajiwala K. , J.FASEB 10 (1996) 75.
[21] W. S. J. Rooman M. J. Protein Eng. 8 (1995) 849.
[22] R. M. Gilis D., J. Mol. Biol. 157 (1996) 1112.
[23] R. M. Gilis D., J. Mol. Biol. 272 (1997) 276.
[24] S. N. Topham C. M., Blundell T. L., Protein Eng. 10 (1997) 7.
[25] A. P. Bordo D., J. Mol. Biol. 217 (1991) 721.
[26] D. A. R. Maxwell K. L., Biochemistry 37 (1998) 721.
[27] D. A. R. Larson S. M., Protein Sci. 9 (2000) 2170.
[28] W. S. J. Prevost M., Tidor B., Karplus M. , Proc. Natl Acad. Sci. 88 (1991) 10880.
[29] K. P. A. Pitera J. W., Proteins: Structure, Functions Genetics 41 (2000) 385.
[30] Radja, Phys. Rev. 72 (2005)
[31] R.W.Peterson, Protein Sci. 13 (2004) 735.
[32] W. W. J. Wang, S. Huo, M. Lee, A. Kollman, J. Phys. Chem. 105 (2001) 505.
[33] C. S. B. Rost, Proteins: Structure, Functions Genetics 19 (1999) 55.
[34] F. M. R. J.W. Ponder, J. Mol. Biol. 193 (1987) 775
[35] J. M. T.J. Pendersen, J. Mol. Biol. 269 (1997) 240.
[36] I. Belda, J. Comput. Aid. Mol. Des. 19 (2005) 585.
[37] J. E. N. Raphael Guerois, Luis Serrano, J. Mol. Biol. 320 (2002) 369.
[38] R. S. P. Felix Deanda, Journal of Molecular Graphics and Modelling 20 (2002) 415.
[39] F. M. Richards, Ann. Rev. Biophys. Bioeng 6 (1977) 151.
[40] Lehninger, in Biochemistry, 2005, p. 52.
[41] B. D. T. Finney J.L., Daniel R.M., Timmins P.A., Roberts M.A., Biophys. Chem. 105 (2003) 391.
[42] D. K.A., Biochemistry 29 (1990) 7133.
[43] B. R.L., Science 295 (2002) 1657.
[44] S. B. A. Pace C.N., McNutt M., Gajiwala K., FASEB J. 10 (1996) 75.
[45] Lehninger, in Biochemistry, 2005, p. 149.
[46] J. T. Carl Branden, in Introduction to protein structure (1999).
[47] B. W. Matthews, Structural and genetic analysis of protein stability (1993).
[48] K. W., Advan. Protein Chem. 14 (1959) 1.
[49] M. M. Peter, E. Leopold, Jose Nelson Onuchic, Biophysics 89 (1992) 8721.
[50] S. Dao-pin S., Wosniak J. A., Sauer U., Mathews B. W. , J. Mol. Biol. 221 (1991)
[51] R. F. M. Lee B., J. Mol. Biol. 55 (1971) 379.
[52] H. T. F. Hasel W., Still W.C., Tetrahedron Comput. Methodol 1 (1988) 103.
[53] E. B. David Dynerman, Julie C. Mitchell, J. Comp. Bio. 16 (2009)
[54] W. J. D. I. Elizabeth R. Collantes, J. Med. Chem 38 (1995) 2705.
[55] B. B. J. Zhou R., Proc. Natl Acad. Sci. 99 (2002) 12777.
[56] G. A. E. Nymeyer H., Proc. Natl Acad. Sci. 100 (2003) 13934.
[57] E. J. S. Young Min Rhee, Guha Jayachandran, Erik Lindahl, Vijay S. Pande. , Proc. Natl Acad. Sci. 101 (2004) 6456
[58] K. R. W. THOMAS P. HOPP, Proc. Natl Acad. Sci. 78 (1981) 3824.
[59] S. Hellberg S., M., Skagerberg B., Wold, S., J. Med. Chem 30 (1987) 1126.
[60] D. R. F. Kyte J., J. Mol. Biol. 157 (1982) 105.
[61] G. Y. Wang R., Lai L. , Perspect Drug Discov. Des. 19 (2000) 47.
[62] R. B. Ertl P., Selzer P., J. Med. Chem 43 (2000) 3714.