The Water Factor in the Protein-Folding Problem

L.F.O. Rocha[1], M.E. Tarragó Pinto[2], and A. Caliri[3]

[1]*Universidade Estadual Paulista, IBILCE, Departamento de Física*
Rua Cristovão Colombo 2265, Jardim Nazareth, 15054-000, São José do Rio Preto, SP, Brazil
[2]*Pontifícia Universidade Católica do Rio Grande do Sul, Departamento de Física Teórica e Aplicada*
Av. Ipiranga, 6681, Partenon, 90.619-900, Porto Alegre, RS, Brazil.
[3]*Universidade de São Paulo, FCFRP, Departamento de Física e Química*
Av. do Café S/N, Monte Alegre, 14040.903, Ribeirão Preto, SP Brazil

Received on 20 October, 2003.

Globular proteins are produced as a linear chain of aminoacids in water solution in the cell and, in the same aqueous environment, fold into their respective unique and functional native structures. In spite of this, many theoretical studies have tried to explain the folding process in vacuum, but in this paper we adopt an alternative point of view: the folding problem of heteropolymers is analyzed from the solvent perspective. The thermodynamics of the folding process is discussed for a non homogeneous system composed by the chain and solvent together; hydrophobic effects, modulated by the polar/nonpolar attributes of the residue sequence and by its corresponding steric specificities, are proposed as basic ingredients for the mechanisms of the folding process. These ideas are incorporated in both lattice and off-lattice models and treated by Monte Carlo simulations. Configurational and thermodynamical results are compared with properties of real proteins. The results suggest that the folding problem of small globular protein can be considered as a process in which the mechanism to reach the native structure and the requirements for the globule stability are uncoupled.

1 Introduction

We still know little about the liquid that is of fundamental importance in biological processes. Particularly in the protein-folding problem, the unusual physical proprieties of water are determinant. For example, almost all proteins denature if their medium is changed from water to another solvent, as ethanol, or even in aqueous solutions containing a sufficient amount of sodium dodecyl sulfate or urea. In addition, proteins denature by simply changing intensive parameters of their solutions, such as pH, temperature, and pressure. Many physical properties of water, such as its high surface tension, specific heat, and heat of vaporization, are surprising if compared with its direct analogous H_2S and CH_4; it is impressive that water presents one of the largest known heat capacities [1] given the small size of its molecules. The effects of its special structural possibilities in either the bulk of pure water or when in contact with others substances [2] are also unique.

Such physical properties of water, concerning room temperature, are among the main reasons for referring to the protein-folding process as one of the most perplexing mysteries in science. Indeed, the task of describing how a particular long strand of amino acids twists and folds into its very specific three-dimensional structure has proved to be a considerable intellectual challenge, but the “...perplexing mystery...” *status* is somewhat exaggerated, although comprehensive: several distinct scientific and technological branches are interested in this problem, which, once understood, could lead to a better understanding of diseases and uncover possible cures. The technological implication of this problem for the pharmaceutical industry, for example, is so impressive that one of the leader computer companies started, in 1999, a research project worth one hundred million American dollars, in order to build a supercomputer – the blue gene– with more than one million processors, primarily dedicated to deal with this question.

In this paper, the protein-folding problem is considered under the solvent perspective. The thermodynamics of the folding process, in its qualitative and quantitative aspects, is discussed for a non homogeneous system composed by the chain and solvent altogether; hydrophobic effect and steric constraints are proposed as basic ingredients for the mechanisms of the folding process. In the next section, the hydrophobic effect is analyzed in connection with the folding problem, and a qualitative view of the folding thermodynamics is used to emphasize the importance of water for the folding process. Section 3 considers Monte Carlo simulations of two chain models: in the first, the chain evolves in the continuum space, and in the second model the chain’s unites are restricted to occupy the sites of a regular cubic lattice. In the last section, the simulation results for two models are discussed, emphasizing that the folding problem may be considered as a process in which the mechanism to reach the native structure and the requirement for the globule stability are uncoupled.

2 Hydrophobicity and the globular protein-folding problem

The peculiar desafinity of oil (a nonpolar substance) for water is historically known since Pliny the Elder (first century A.D.), but the first reference to the word "hydrophobicity" is not clear; it appeared at least as early as 1915, although it had been defined differently from that which is currently in use [3]. Through time, people from different technological and scientific branches have employed the term "hydrophobic effect" with distinct meanings, sometimes confusing. Some refer to it simply as desafinity of oil for water; ordering of water around nonpolar solute; or, still, free energy involved in transferring a nonpolar solute from a nonpolar environment into water. Anyway, the rigorous understanding of the so called "hydrophobic effect", at molecular level, is still in progress.

2.1 The hydrophobic effect

The term "hydrophobic effect " was coined by Charles Tanford [4] and it refers to the surprising thermodynamics of mixing nonpolar substances with water (oil/water mixing phenomenon). Recent advances suggest that the apparent "aversion" of nonpolar substances to water comes, indeed, from the versatility of water molecules in avoiding the reduction of hydrophobic bonds among water molecules [5]: on the contrary to common sense, there is an energetic preference for nonpolar substances to aggregate with water than with themselves, but it is the water-water interaction, through (relatively) strong hydrogen bonds, that ultimately segregates the nonpolar species from the water bulk, as macroscopically observed through the minimization of the nonpolar-water interface. For large enough extensions of nonpolar-water interfaces, bond breaking is inevitable and the local distribution of hydrogen bonds becomes asymmetric; in order to minimize the system's potential energy, the hydrogen bonds are redistributed into the water bulk direction. This process is substantially the same as that observed at the air-water interface, easily identified by water's high surface tension. However, for small enough nonpolar molecules (extensions $\lesssim$ 1 nanometer), hydrogen bonds simply involve the molecule, resulting in a practically unaffected average number of hydrogen bonds . This mechanism explains, for instance, the higher solubility of small nonpolar solutes[6].

The versatility of water molecules in avoiding the loss of hydrogen bonds comes from its peculiar interactional properties, which are orientation dependent; it also involves –under the conditions of many physical and biological processes– a relevant amount of energy, 10 - 20 KJ, is necessary to break 1 mol of hydrogen bonds. About 1.9 Å separates the hydrogen (donor) of one molecule from the oxygen (acceptor) of the other, and, considering all four electron pairs involved, the water structure may be considered tetrahedral. As a consequence, one water molecule may be bounded up to four other molecules, potentially forming an extensive hydrogen bonding network. The term hydrophobic, commonly used to label substances that apparently repel water, is therefore not appropriate. There is no repulsion between water and such substances; rather, the observed apparent "aversion" is owing to the strong hydrogen bonding between water molecules, also responsible for water's high specific heat, and heat of vaporization. Nonetheless, it is too late for any corrections: the term hydrophobic is going to be continuously used, but not with its etymological meaning.

Water's tetrahedral-type structure and the extension of the hydrophobic bond explain its anomalous packing density, namely $\phi_w = 0.36$. This figure is very small if compared with the protein's interior packing density $\phi_p = 0.75$ (which is greater than closest packing of spheres $\phi_{cp} = 0.74$) or the cyclohexane $\phi_{ch} = 0.44$. At 4°C and atmospheric pressure, each molecule of water is surrounded only by 4 or 5 other molecules; for closest packing of hard spheres, this number is 12 and for many simple liquid it is 8 or10; therefore, one can conclude that water has a very open structure [7].

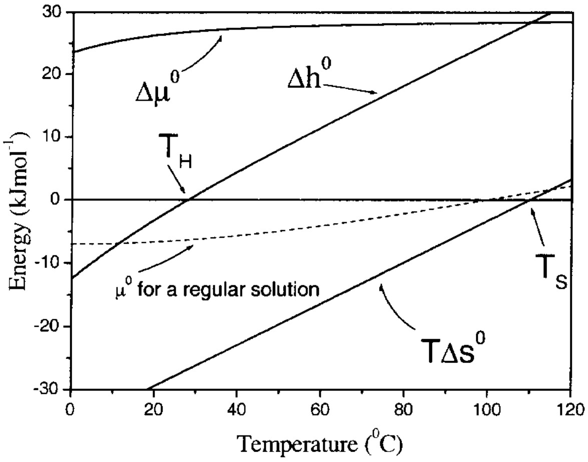

Figure 1. Molar free energy transfer, $\Delta\mu^0 = \Delta h^0 - T\Delta s^0$, of a nonpolar substance into water as a function of temperature; schematic representation. The free energy change is large and positive in the intire range for liquid water; at $T_h \simeq 25^0$C, and $T_s \simeq 120^0$C the enthalpic and entropic changes are zero, respectively. For comparasion, the molar free energy transfer for a regular solution is also shown schematically (dashed line).

Much of what is understood about the hydrophobic effect comes from studies on the oil/water partitioning of small nonpolar solutes. At equilibrium, the molar free energy of phase $i = 1$ and phase $i = 2$ are equals: $\mu_1^0 + RT\ln(c_1) = \mu_2^0 + RT\ln(c_2)$, where μ_i^0 is the affinity of the solute for phase i, and c_i is the concentration of solute in phase i. The difference $\Delta\mu^0 = \mu_2^0 - \mu_1^0$ of molar free energy, at equilibrium, is balanced by the solute partitioning between the two phases, given by the difference of concentration of solute, that is $\Delta\mu^0 = RT\ln(c_1/c_2)$. For nonpolar solutes in its on phase, $\Delta\mu^0$ is interpreted as the transfer free energy to change the solute from the nonpolar phase into water. For small nonpolar molecules (pentane, for example) , $\Delta\mu^0$ is positive (therefore, oil/water mixing is not a spontaneous process), and varies slightly with temperature, from about 23 to 28 KJmol^{-1}, for temperatures ranging from zero to 100^{0}C. Its enthalpic Δh^0 and entropic $T\Delta s^0$ components ($\Delta\mu^0 = \Delta h^0 - T\Delta s^0$) change almost linearly along the entire interval of temperature of liquid water, which gives a nearly constant molar heat capacity in this temperature range. Fig.1 shows a schematic representation of the main characteristics of the hydrophobic effect [2] (mixing of nonpolar solute in water).

2.2 The folding process

Under physiological conditions (namelly, adequate pH and temperature) many proteins fold by themselves, that is, they fold without intermediation of any biological machinery. This experimental observation suggests that the folding problem may be described only by chemical and physical parameters, despite its complexity. However, even considering the anomalous behavior of water and its solutions, most of the models used in studying the folding problem has tried to include water only implicitly, by integrating, *a priori,* the solvent degrees of freedom. This simplified treatment is derived through statistical mechanics, by first writing the configurational partition function of the entire system, which is constituted by a protein molecule in the presence of the solvent:

$$Z = \sum_{\{R_i\}} \sum_{\{r_j\}} \exp[-E(R_{1,} ... R_N;\, r_1,\, ...\, r_m)/k_BT], \quad (1)$$

where E is the total potential energy function, k_B is the Boltzmann constant, T is the absolute temperature, and the summation involves the coordinates of the chain $\{R_i\}$ and the coordinates of the solvent $\{r_j\}$ (treated here as discrete variables for simplicity). The partition function may be rewritten as

$$Z = \sum_{\{R_i\}} \exp[-\Phi(\{R_i\})/k_BT], \quad (2)$$

with

$$\Phi(\{R_i\}) = -k_B \ln \sum_{\{r_j\}} \exp[-E(R_1...R_N; r_1...r_m)/k_BT]. \quad (3)$$

The function $\Phi(\{R_i\})$ can be seen as the free energy for the hypothetical case in that the chain is maintained “frozen” in a particular configuration through the set $\{R_i\}$ of fixed coordinates $\{dR_i/dt = 0\}$. Therefore, at this point, one has two possible ways of focusing the problem, from the statistical mechanics point of view:

(i)- One may see the free energy Φ as *potential of mean force* [8], obtained by the summation over the solvent degree of freedom, that is $\Phi(R_1...R_N)$ is averaged over all possible solvent coordinates $\{r_j\}$. This is what is assumed in most simplified models of protein folding: the *solvent factor* is averaged out and appears implicitly as a potential of mean force, and so the problem is reduced to intra-chain interactions through the potential of mean force Φ. However, althought this approach has been extensively employed, even for dealing with the dynamical aspect of the problem, a definitive potential of mean force was never found: it is not feasible to condense all solvent-chain information (in an operational manner) in a single function of the chain coordinates $\{R_i\}$ and at the same time, keep the problem treatable. Therefore, simplified potentials of mean force are commonly employed in considering solvent implicitly. Anyway, the price to be paid for using this approach is that of having to consider the solvent at macroscopic equilibrium with the chain during all the process.

(ii)-On the other hand, if one considers that the solvent should play a more explicit role in the dynamic process of folding, an alternative approach is then to emphasize the solvent-chain interactions. The details of how the correct chain’s segments get together and keep in contact, clearly has to do with the stereochemical pattern encoded in the chain sequence. But, before getting into the details of this alternative view of the folding problem, let us analyze its interplaying parts through some thermodynamic quantities, in order to associate the hydrophobic effect –as observed in the nonpolar/water system– with the chain-solvent equilibrium state.

2.3 Thermodynamics of the folding process

Indeed, theoretical studies and experimental evidences point to the hydrophobic effect as the main folding driving force generator[9] [10]. One of the most convincing arguments comes from the thermodynamic analysis of the *folding* $\rightleftharpoons$ *unfolding* reaction. Measurements of thermodynamic amounts may provide insights about the relevant forces involved in the protein stability and folding process. In this analysis, it is usual to consider the enthalpic and entropic changes through the folding process, taking separately the chain and solvent contributions by writing the entire free energy changes $\Delta G = G_{folded} - G_{unfolded}$ between the final folded structure and the initial unfolded conformation, as

$$\Delta G = \Delta H_{chain} + \Delta H_{solvent} - T\Delta S_{chain} - T\Delta S_{solvent}; \quad (4)$$

the enthalpic changes ΔH also means $\Delta H = H_{folded} - H_{unfolded}$, as well as $\Delta S = S_{folded} - S_{unfolded}$. The interaction energy between chain and solvent and among those solvent molecules affected by the chain presence are all contained in ΔH_{chain}. It is necessary to bear in mind, however, that this procedure of breaking down a full thermodynamic quantity into a sum of components is not a rigorous method; indeed, special conditions, such as free energy additivity, are *a priori* assumed. But our interest here is only to provide a qualitative analysis of the thermodynamics of the folding process, and so we proceed without major considerations about.

First, we consider the polar (^{p}) groups (side-chains). In this case, the entalpic change between the folded structure and the unfolded conformation is positive, $\Delta H^{p}_{chain} > 0$, that is, the chain enthalpic change contributes to the unfold conformation. The reason is that these groups, as the backbone, form much more hydrogen bonds and salt interactions when exposed to water than among themselves. On the other hand $\Delta H^{p}_{solvent} < 0$, favoring the folded structure. The amount $\Delta H^{p}_{solvent}$ is negative because the water molecules interaction with themselves is stronger than with the chain. Therefore, the terms in the sum $\Delta H^{p}_{chain} + \Delta H^{p}_{solvent}$ compete and, although close to zero, the sum usually favors the folded structure slightly, that is: $\Delta H^{p}_{chain} + \Delta H^{p}_{solvent} \lesssim 0$, indicating that the water molecules interaction with themselves prevails.

The chain entropic change is negative, that is, $-T\Delta S^{p}_{chain} > 0$, because the chain in the unfolded conformation has much more assessable configurations than in the folded structure and so it contributes to the unfolded conformation. However, $-T\Delta S^{p}_{solvent} < 0$ because the water has more configurational choices with the chain in the folded structure, and so this term favors the folded structure. The balance is found to be slightly negative, that is: $\Delta S^{p}_{chain} + \Delta S^{p}_{solvent} \lesssim 0$. Then, summing up the contribution from the chain and solvent, namely $\Delta G^{p} = \Delta H^{p}_{chain} + \Delta H^{p}_{solvent} - T(\Delta S^{p}_{chain} + \Delta S^{p}_{solvent})$, one usually finds that polar groups favor the unfolded conformation, that is: $\Delta G^{p} \gtrsim 0$.

Now we turn to the nonpolar ($^{\widetilde{p}}$) groups. As in the polar case, $\Delta H^{\widetilde{p}}_{chain} \gtrsim 0$ contributing to the unfold conformation. Although the chain's nonpolar groups interact (slightly) stronger with water than with themselves, when the chain is open, the backbone may form more hydrogen bonds with water than with themselves. But, it is not so favorable to the unfolded conformation as in the polar case, because the van der Waals (attractive) interactions are weaker between nonpolar groups and water than among themselves. On the other hand, water molecules interact much more strongly among themselves than with nonpolar groups; therefore $\Delta H^{\widetilde{p}}_{solvent} < 0$, favoring the folded structure. Here, again, the terms in the sum $\Delta H^{\widetilde{p}}_{chain} + \Delta H^{\widetilde{p}}_{solvent}$ compete, slightly favoring the folded structure, that is: $\Delta H^{\widetilde{p}}_{chain} + \Delta H^{\widetilde{p}}_{solvent} \lesssim 0$.

Following, it is found that the sum $\Delta S^{\widetilde{p}}_{chain} + \Delta S^{\widetilde{p}}_{solvent} > 0$, that is, it strongly induces the chain into the folded conformation. The term $-T\Delta S^{\widetilde{p}}_{chain} > 0$ contributes to the unfolded conformation, like in the polar case, but the solvent contribution strongly favors the folded structure, that is $-T\Delta S^{\widetilde{p}}_{solvent} << 0$, and is the leading driving force for protein folding. The structural arrangement of water molecules at the water-nonpolar interface are significantly affected: due (again) to the stronger water-water interaction (much stronger than the water-nonpolar groups interaction), the lost hydrogen bonds are partially compensated by redirecting them towards the water bulk. Therefore, when by thermal fluctuation two nonpolar groups get closer, the resulting reduction of exposed nonpolar surface to water increases the number of hydrogen bonds (a water molecule always "looks" for other water molecules). The net effect is the collapse of the chain, reducing the overall water-nonpolar interface and increasing the number of possible hydrogen bonds for water molecules; that is, the folded structure increases the solvent entropy. Therefore, adding all the contribution from nonpolar groups, namely $\Delta G^{\widetilde{p}} = \Delta H^{\widetilde{p}}_{chain} + \Delta H^{\widetilde{p}}_{solvent} - T(\Delta S^{\widetilde{p}}_{chain} + \Delta S^{\widetilde{p}}_{solvent})$, one finds that it favors the folded structure, that is $\Delta G^{\widetilde{p}} < 0$.

Finally, adding up all terms, at room temperature, it is found that, for proteins, $\Delta G = \Delta G^{p} + \Delta G^{\widetilde{p}}$ is negative, that is, the folding is a spontaneous process. However, the change ΔG is very small, indicating that proteins are marginally stable. For typical proteins, ΔG is found to be between -10 and -50 KJ/mol, which corresponds to few hydrogen bonds. In short, one finds that enthalpic and entropic contributions due to the chain favor the unfold conformation ($\Delta H^{p}_{chain} > 0$; $\Delta H^{\widetilde{p}}_{chain} \gtrsim 0$; $-T\Delta S^{p}_{chain} > 0$; $-T\Delta S^{\widetilde{p}}_{chain} > 0$); while all terms due to the solvent contribution favor the folded structure ($\Delta H^{p}_{solvent} < 0$; $\Delta H^{\widetilde{p}}_{solvent} < 0$; $-T\Delta S^{p}_{solvent} < 0$; $-T\Delta S^{\widetilde{p}}_{solvent} << 0$). Then, it seems clear that the solvent plays a dominant role in the spontaneous folding process, and a hypotheses may be now set up: The solvent "folds the chain" in a such way that the number of lost hydrogen bonds is reduced to a minimum level; the final structure specificity and stability are decurrent from the particular stereochemical instruction encoded along the chain. Therefore, in the following topic, we will approach the folding problem focusing the solvent part, specially thinking of using the Monte Carlo method.

2.4 Folding of grobule protein driven by the hydrophobic effect

In order to emphasize the hydrophobic effect as the folding driving force generator, let us consider a microcanonical representation of a system constituted by a single chain immersed into N_0 solvent molecules (water), a fixed volume V_0, total constant energy E_0, and Γ_0 accessible states, all of them evenly probable. The system as a whole (actually a non-homogeneous system) reveals two distinct parts as illustrated in Fig.2: (i)- a subsystem constituted by the chain and its neighborhood, and (ii)- the complementary system. The subsystem, in an arbitrary state α, occupies a volume V_α, which has all N_α water molecules that 'perceive' the chain, and energy E_α; the complementary system is constituted by the complementary volume V'_α, the reminder number N'_α of bulk water molecules, energy E'_α and having a number Γ'_α of accessible states. Clearly, $V_\alpha + V'_\alpha = V_0$; also $N_\alpha + N'_\alpha = N_0$; and $E_\alpha + E'_\alpha = E_0$. The energy of the subsystem is given by the sum $E_\alpha = E^{cc}_\alpha + E^{cs}_\alpha$, where E^{cc}_α is the intra-chain energy and E^{cs}_α includes the chain-solvent interaction energy and the energy between those solvent molecules that can be affected by the chain presence.

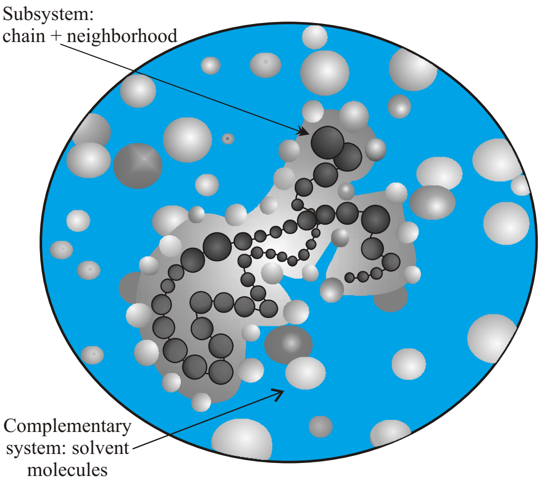

Figure 2. Microanonical system: representation of the chain-water system. The system as a whole has N_0 solvent (water) molecules, fixed volume V_0, total (constant) energy E_0 and Γ_0 accessible states, is divided in a subsystem constituted by the chain and solvent molecules at its neighborhood, and the complementary system (just solvent molecules).

Therefore, one may assume that along the configurational evolution, a given state α of the system as a whole has probability of occurrence

$$P_\alpha = \frac{\Gamma'_\alpha}{\Gamma_0}, \tag{5}$$

because, for the same state α, the complementary system can assume Γ'_α distinct states out of Γ_0. Therefore, the probability of occurrence of a local specific physical condition, such as a microscopic fluctuation specified by the state α of the subsystem, depends on the number of accessible states left to the complementary system. Dynamically, this means that along a large enough time interval τ in which a number $\aleph(\tau)$ of states are visited , state α occurs $\aleph(\tau)\times \Gamma'_\alpha/\Gamma_0$ times. As the system as a whole is considered in macroscopic equilibrium, P_α is then the 'chance' of occurrence of a specific local microscopic fluctuation specified by the subsystem at state α. Eq.5 can be re-written by taking the natural logarithm of its both sides and after defining $\ln\Gamma = S/k_B$, one gets

$$P_\alpha = \exp(-\Delta S'_\alpha/k_B), \tag{6}$$

where $\Delta S'_\alpha = S_0 - S'_\alpha$; note that $\Gamma_0 = \sum_{\{\alpha\}} \Gamma'_\alpha$. The amount $\Delta S'_\alpha$ is not the entropy of the subsystem; the correct thermodynamic corresponding amount is obtained by the ensemble (or temporal) average $S = \overline{\Delta S'_\alpha} = S_0 - S'$. On the other hand, the amount S'_α can be seen as the entropy of the whole system when the subsystem is considered 'frozen' in state α exactly as considered in Eq.3. Therefore, through this interpretation, Eq.5 may be used along with the master equation to get a prescription of the transition probability between two (consecutive) states of the whole system.

$$T_{\alpha,\beta} = \frac{P_\alpha}{P_\beta} = \min(1, \exp[(S'_\alpha - S'_\beta)]/k_B), \tag{7}$$

where the function $\min(a, b)$ means the minimum amount between a and b, and $\delta S' = S'_\alpha - S'_\beta$ is the entropy difference between two distinct states of the system as a whole, namely: the state corresponding to the subsystem 'frozen' in state α and the state corresponding to the subsystem 'frozen' in state β. The idea now is to estimate the amount $\delta S'$ through experimental data, with respect to the change on the free energy $\Delta G = \Delta H - T\Delta S$ involved in transferring a solute from a nonpolar environment (interior of a protein) into water (solvent). Such transfer free energy, determined for all 20 natural aminoacids, are used to construct what is called "hydrophobic scales". Fig.3 shows two of such scales [4]. One was obtained from experimental measurements and the other from theoretical calculations. They were chosen here, among tens of others, because they agree qualitatively about which are the "hydrophobic" and the "hydrophilic" aminoacids – it is common to find qualitative disagreement between different scales for some aminoacids. Therefore, writing the entropic change of the system as $\Delta S' = -(\Delta G' - \Delta H')/T$,one may use experimental results to weigh distinct configurations . As a first approximation that emphasizes the solvent factor, one may drop the entalpic term and write $\Delta S' \simeq -\Delta G'$; particularly, this approximation becomes exact for models using only hard-core potential for the intra-chain interactions.

However, as discussed above, on considering the thermodynamics of the folding process, proteins are marginally stable and so some other "non energetic" ingredients must play a role in its overall stability. And, indeed, it is surprising how "hydrophobic potentials ", constructed from hydrophobic scales, in association with steric effects may mimic the protein process, as it will be considered in the next section.

3 Hydrophobic induction and steric constraints as two main ingredients in the folding process

The mechanisms and thermodynamics of the hydrophobic effect, as discussed above for oil/water mixing, are not easily and directly applied to protein systems, although the hydrogen bonding is at the root of both problems. Differently from single small nonpolar molecules, proteins are linear long heteropolimers with polar and nonpolar groups mixed in the same chain. In spite of some characteristic behavior of oil/water mixing being qualitatively equivalent to protein systems, its corresponding quantitative parameters are markedly distinct. Indeed, the molar amount of energy contained in the changes ΔH and $T\Delta S$, separately, in the exposure of nonpolar groups during the protein unfolding reaction is much higher than its corresponding amount in oil/water mixing, as illustrated in Fig.4, for lysozyme [9];

Residues	Trp	Ile	Phe	Leu	Met	Val	Cys	Tyr	Pro	Ala	Thr	His	Gly	Ser	Gln	Glu	Asn	Asp	Lys	Arg	
h_i	-3,07	-2,46	-2,44	-2,32	-1,68	-1,66	-1,34	-1,31	-0,98	-0,42	-0,35	-0,18	0	0,05	0,30	0,87	0,82	1,05	1,35	1,37	*
	-2,6	-1,9	-2,3	-1,9	-2,4	-1,5	-0,38	-1,6	-1,2	-0,67	-0,52	-0,64	0	-0,01	0,22	0,76	0,60	1,2	0,57	2,1	†
3-letter alphabet	H	H	H	H	H	H	H	H	H	H	H	H	N	N	P	P	P	P	P	P	

Legend- *: experimental scale; †: theoretical scale; h_i: hydrofobicity level of each aminoacid

Figure 3. Hydrophobic scales: there are many distinct scales determined experimentally; these two were choosen because there is a qualitative agreement between them and theoretical calculations.

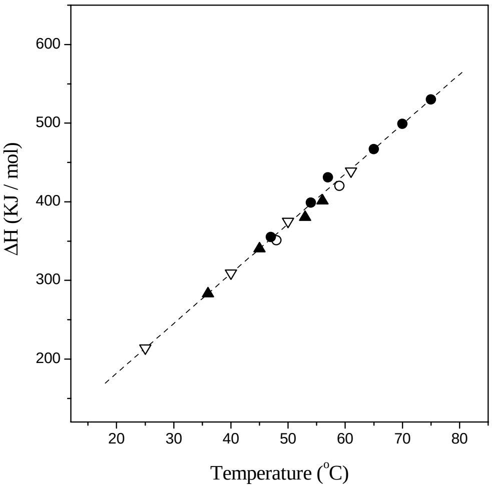

Figure 4. Classical result of the enthalpy change ΔH of the conformational transition (denaturation) of lysozyme as a function of temperature . There is no difference in heat capacity (slope of the enthalpy change as a function of temperature) upon denaturation induced by temperature (open circles) at constant pH; acid (filled circles) at constant temperature; denaturation by temperature (open triangle) at constant concentration of denaturant guanidine chloride; and denaturation by guanidine chloride (filled triangle) at constant temperature. For many proteins studied [9], the same linear dependence ΔH with the temperature was found, but with different slopes for each protein.

compare it with Fig.1. The resulting balance for protein unfolding free energy $\Delta G = \Delta H - T\Delta S$, although positive, is surprisingly small. In addition, while in oil/water mixing of small solutes, the temperatures T_h in which the enthalpic contibution is zero, and the temperatures T_s in which the entropic contibution is zero, are separated by about 80^0C, they are about nearly the same for protein [3]. In oil/water mixing, as well as in protein systems, the entropic term is the predominant one, that is $|T\Delta S| > |\Delta H|$, from lower temperatures up to T_ε, in that, by definition, $-T_\varepsilon \Delta S = \Delta H$; note that $T_h < T_\varepsilon < T_s$. The value for T_ε varies from protein to protein: it is about 10^0C for T4 lysozyme [11], 26^0C for myoglobin [12], and about 25^0C for several other proteins[3]. Such thermodynamic results certainly reinforce the idea that the protein chain folding is driven by the solvent's "entropic forces"; the enthalpic component only starts to favor the folded structure for temperatures $T > T_h$ when the unfolding enthalpy change ΔH becomes positive.

The stability of the native structure of single-domain proteins is measured by the work required for its disruption, and, for the case of proteins having only two states, namely folded and unfolded, the work required for transition from the native structure to the denatured conformation is given by $\Delta\widetilde{G} = \Delta H - T\Delta S$, the unfolding free energy. But, as already emphasized, the amount $\Delta\widetilde{G}$ is positive but surprisingly small, producing questions such as: Why do globular proteins have to be only marginally stable?; or: Are there any other hidden stabilizing mechanisms that are not revealed by final state thermodynamic amounts? Indeed, it can be observed that, in living cells, there is a continuous process of production and degradation of most of their constitutes. Many functional globular proteins are systematically produced and eliminated, in a wide range of time characteristic (from fraction of second to days, or years). Therefore, the fact that proteins are only marginally stable structures seems to be important for the maintenance of the biologic machinery of life. But, on the other hand, under the same physiological conditions, how can distinct proteins present distinct half-live within the same range of small thermodynamic stability? A complex combination of factors, such as sensitiveness for local chemical fluctuation, number of covalent cross-links, protein size and shape, etc., would usually be evoked. However, among all these possible factors, a good starting point for investigation seems to be the *steric specificities* of the chain because they are present in all proteins through their elementary constituents: the twenty natural aminoacids presenting a rich repertory of sizes and shapes. Therefore, one gets a simple but consistent proposition in order to study the folding process: in the linear sequence of aminoacids of a particular protein, a stereochemical instruction to the *solvent* should be encoded, dictating *how* and in *what* structure the protein has to be folded. In the next topic, Monte Carlo simulation results of two models are presented in order to analyze the consequence of considering the hydrophobic effect (entropic forces) and the steric specificities (or constraints) as the two main ingredients in governing the folding process.

3.1 Off-lattice model: the hydrophobic induction

Entropic forces originated by the sequence of specific residues of a protein are efficient for packing and inducing the chain to visit the native configuration, but they fail to provide it with enough stability [13]. In order to primarily examine the exclusive efficiency and limitations of this "modulated hydrophobic effect", an off-lattice model is employed, in that the stereochemical diversity of the 20 natural amino acids is reduced to a $three$-letter alphabet, representing polar (P), hydrophobic (H), and neuter (N) monomers, as depicted in Fig.3, at the same time that all geometrical constraints were deliberately eliminated. Technically, the chain-solvent system is represented as a pearl necklace in the solution (the solvent is treated explicitly), in which each monomer is represented by a hard-sphere of diameter D connected to its neighbors by ideal flexible strings with defined length $D+\varepsilon$, where $\varepsilon \simeq 0.2D$. The 12.565 solvent molecules are also represented by hard-spheres of the same diameter D. The magnitude of the specific hydrophobic levels $\{h_i\}$ is equivalent to the one used in the lattice model, with each monomer of the chain having one of three possible values: $h_P = +1$, $h_H = -2$, or $h_N = 0$. The solvent-solvent interaction, $e_{s,s'}$, as well as the monomer-monomer interaction, $e_{m,m'}$ is a hard core-type potential. The solvent-monomer interaction involves, additionally, the hydrophobic energy, $e_{s,m} = e_0 - n_s h_m$, where e_0 is an arbitrary constant, h_m is the hydrophobic level of monomer m (h_P, h_H, or h_N), and n_s is the number of solvent molecules surrounding it. Note that the energy $e_{s,m}$ increases with n_s if the monomer m is hydrophobic ($h_m = h_H$), decreases if it is hydrophilic ($h_m = h_P$), and is indifferent otherwise ($h_m = h_N$). Each monomer in the HPN sequence corresponds, one by one, to the polar/nonpolar attribute of the 35 amino acids of a real protein, and its corresponding 3-D structure was specially chosen as a target configuration . The polar/nonpolar attribute of each residue was chosen based on the scale proposed in Fig.2, and the configurational evolution is governed by Eq.7 above. The amount $\delta S_{\alpha,\beta} = \Delta S_\alpha - \Delta S_\beta$ is the system's entropic change with respect to the transition between the distinct (frozen) chain configurations α and β, and may be expressed by its corresponding changes on energy $e_{\alpha,\beta}$, number of water molecules bounded in the chain $n_{\alpha,\beta}$, and molecule volume $v_{\alpha,\beta}$, in the form $\delta S_{\alpha,\beta} = (-\delta e_{\alpha,\beta} + \mu\, \delta n_{\alpha,\beta} - P\delta v_{\alpha,\beta})/T$; see ref. 15 for more details. In the present model, we set $\delta n_{\alpha,\beta} = \delta v_{\alpha,\beta} = 0$ and, because only hard-core energies are considered in the inter-monomer interactions, $\delta S_{\alpha,\beta}$ depends only on the change of the hydrophobic energy $\delta e_{\alpha,\beta}$. Essentially, this change depends on the reorganization of hydrogen bonds in the layer of water molecules between the solvent bulk and the monomers surface[7], establishing a linear dependence on the accessible surface area of the residues. In this work, the amount $\delta e_{\alpha,\beta}$ is estimated by associating a specific hydrophobic level $\gamma = 2\gamma_0$ for hydrophobic monomers, $\gamma = -\gamma_0$ for polar ones, and $\gamma = 0$ for neuters, for each solvent contact; the amount $\gamma_0 < 0$ is measured in units of $k_B T$; the value $\gamma_0 = -1$ was used in this work. Therefore, each new generated configuration, say configuration β, is obtained from a previous configuration α by trying to change the spatial coordinates of a specific chain's monomer or a solvent molecule, which are chosen randomly along the MC simulation. There are three possible situations to be considered with respect to the chance $T(\alpha \rightarrow \beta)$ in accepting the new configuration β:

(i)- moving a hydrophobic monomer:

$$T(\alpha \rightarrow \beta) = \min\{\exp(\mathbf{2\gamma_0 \Delta n_S / k_B T}), \mathbf{1}\}; \tag{8}$$

(ii)- moving a polar monomer:

$$T(\alpha \rightarrow \beta) = \min\{\exp(-\mathbf{\gamma_0 \Delta n_S / k_B T}), \mathbf{1}\}; \tag{9}$$

(iii)-moving a solvent molecule:

$$T(\alpha \rightarrow \beta) = \min\{\exp[\mathbf{\gamma_0}(\mathbf{2\Delta n_H} - \mathbf{\Delta n_P})/\mathbf{k_B T}], \mathbf{1}\}, \tag{10}$$

where Δn_S is the change of the number of solvent molecules in contact with the monomer being moved, and Δn_H and Δn_P are, respectively, the change on the number of hydrophobic and polar monomers around of the solvent molecule being moved. The Monte Carlo sampling technique, then, becomes straightforward: each new generated configuration is first checked with respect to the hard-core constraints and, if no superposition is verified, the configuration is accepted according to $T(\alpha{\rightarrow}\beta)$, which is calculated by Eqs. 8-10 above.

Firstly, in following analysis of the simulation results, the global chain behavior of the packing process is presented through the standard deviation SD_G of the average radius of gyration R_G against $k_B T$, as shown by Fig.5, which was obtained by using the last 10^5 MC steps –representing one fifth of the total time window t_w, namely $t_w = 5 \times 10^5$ MC steps that corresponds to a total of about 6×10^9 generated configurations. Three distinct regions are identified: For $k_B T < 1.5$ (region A) the amount SD_G depends strongly on the initial conditions. For $1.5 \leq k_B T \leq 3.0$ (region B) the globule is well defined; the smaller value for SD_G occurs at $k_B T = 1.5$, and then increaes slowly up to $k_B T = 3.0$. Finally for $k_B T > 3.0$ (region C), SD_G changes rapidly with the temperature until saturating at $k_B T \gtrsim 5.0$.

¿From $k_B T = 1.5$, up to $k_B T = 3.0$, the size of the globule can be thermodynamically defined, independently of the initial condition: thermal fluctuations are already significantly large to disrupt the non-optimized hydrophobic contacts and so, independently of the initial conditions, the chain always collapses into a compact globule-like conformation. At $k_B T = 2.0$, and $k_B T = 3.0$, SD_G is only about 3% and 5% lager than SD_G at $k_B T = 1.5$, respectively.

More detailed configurational behaviors are provided by contact maps of four real proteins and for their corresponding models, as shown in Fig.6. Owing to severe topological simplifications introduced by the model used here, two precautions were taken before configurational comparisons: (*i*)- the inter-monomers distances $d_{i,j}$(center of mass), for protein and model, were properly translated and re-scaled to fit the same interval from zero to one, that is, $0 \leq d_{i,j} \leq 1$, and (*ii*)- black regions in the maps correspond to all distances $d_{i,j}$ satisfying $0 \leq d_{i,j} < 0.3$, that is, distances up to 30% of the largest distance (for each case: model and protein), and as white regions for distances $0.3 \leq d_{i,j} \leq 1$.

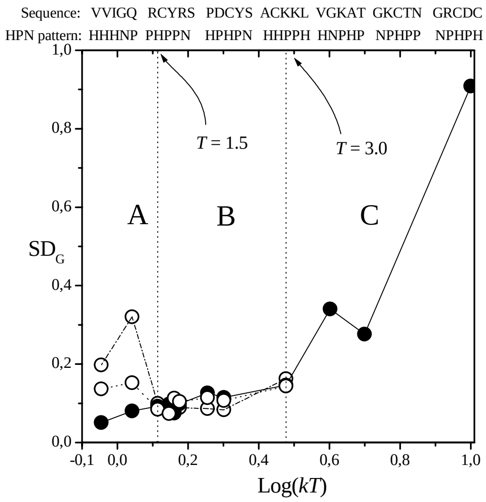

Figure 5. The behavior of the Standard Deviation SD_G of the radius of gyration R_G against k_BT. The protein 1-tsk sequence (one letter symbol) and its HP pattern are shown at the top. The physical system as a whole, is represented by a single linear chain of 35 units, surrounded by $12,565$ solvent molecules confined in a cubic box. The beads of the chain, as well as the solvent molecules, are hard-spheres of the same diameter, but the monomer-solvent interaction depends additionally on the hydrophobic attribute of each interacting pair.

The contact map for each model corresponds to a particular configuration chosen among ten of them, taken from the last quarter of the simulation time.

The simulated and real protein maps can not be conclusively compared given the severe geometrical changes imposed into the model. However it is possible to see specific propensities (even though somewhat distorted) in each model's map, resembling the corresponding real protein's map. First, it is helpful to recognize the distinct configurational peculiarities of the simulated maps, and that they are exclusively due to its distinct HPN patterns. So, henceforth one is ready to look for the similitudes between the protein's map and its corresponding model's map. Note that several contacting residues in the real protein (black regions of Fig.6) have corresponding contacting pairs in the model system; many of them presenting a relatively high frequency of contact along the simulation. Although the globule is very compact, it preserves great malleability, indicating that the chain is not pinned in any particular configuration . The resemblances between proteins and model maps are recurrent, appearing and disappearing from time to time along the simulation; the chain-model configurations shown in Fig.6 were selected for visual purpose only to illustrate our arguments. They have, indeed, a short lifetime, but even with similarity alternations between model's results and the native structure, many contacting pairs of monomers last along the whole simulation time. When the chain is compacted in a globule conformation, an continuous succession of swelling, shrinkage, and twisting of the globule takes place; the chain rambles through the compact configurational sub-space, eventually visiting configurations that present more resemblances with its respective native configuration. More detailed results and comments about this model will appear soon elsewhere.

Therefore, the main virtues of the chain-solvent interactions, governed by the hydrophobicity of the residues, are the efficiency to compact the chain maintaining the globule malleable, and, once packed, the capability to induce the chain through conformations near the native state, but without providing configurational definition to the globule. In the next topic, dealing with a lattice model, the same type of configurational induction by the hydrophobic effect is reproduced, but with an additional remedy: the introduction of steric interaction specificities.

3.2 Lattice model: steric constraints and its thermodynamics implications

The intrachain contact energy $\varepsilon_{i,j}$ between a pair of residues i and j, used in most lattice simplified models, obey the so called segregation principle, $2\,\varepsilon_{i,j} - \varepsilon_{ii} - \varepsilon_{jj} \geq 0$. A particular class of potentials, denominated hydrophobic potentials, has the property of satisfying marginally the segregation principle, through the equal sign, that is, $2\,\varepsilon_{i,j} - \varepsilon_{ii} - \varepsilon_{jj} \equiv 0$. Such potentials are based on a hydrophobic scale for the aminoacids $\{h_i\}$ and, in general, the effective intermonomer potentials are written as $\varepsilon_{i,j} = h_i + h_j$, that is, a linear combination of the hydrophobicity level of each interacting monomer [14]. The interaction potential obtained by this way has an important property: the energy change between two chain configurations is exactly equivalent to that obtained from considering the exclusive chain-solvent interaction [15], in that all lattice sites are occupied by either monomers or solvent molecules. The model studied here is composed by a single protein-like chain constituted by $N = 27$ monomers, which are effective residues taken from a repertory of stereochemically different elements; the residues occupy consecutive and distinct sites of a three-dimensional infinity cubic lattice; the interactions are assumed to occur between nearest-neighbor pairs of residues through a set of contact energy $\{\varepsilon_{i,j} = h_i + h_j\}$ and steric constraints $\{c_{i,j}\}$. Together, the set of hydrophobic levels $\{h_i\}$ and steric interactional specificities $\{c_{i,j}\}$ of the residues, constitute a 10-letter alphabet, as shown in Fig.7. The strength of the interactions $\{h_i\}$ are expressed in units of k_BT, (arbitrary energy units).

Compact Self-Avoiding (CSA) configurations, characterized by their corresponding relative contact order χ, are used as native or target structures [13]. The sequence of residues assigned to each structure is determined through a specific "syntax" that emerges from the constraints $\{c_{i,j}\}$ and from the application of the "hydrophobic inside" rule; see Ref.15 for details.

3.2.1 Heat capacity and configurational activity

In this paper, we consider the analysis of heat capacity and configurational activity for a particular structure, featured by its relative contact order $\chi = 0.2381$, in order to thoroughly discuss the effect of steric constraints in selecting folding pathways and on the overall globule stability; the monomer's sequence for this case is [CBCIA ECBCE ADHRH DAECB CEAIC BC]; see alphabet's details in Fig. 7.

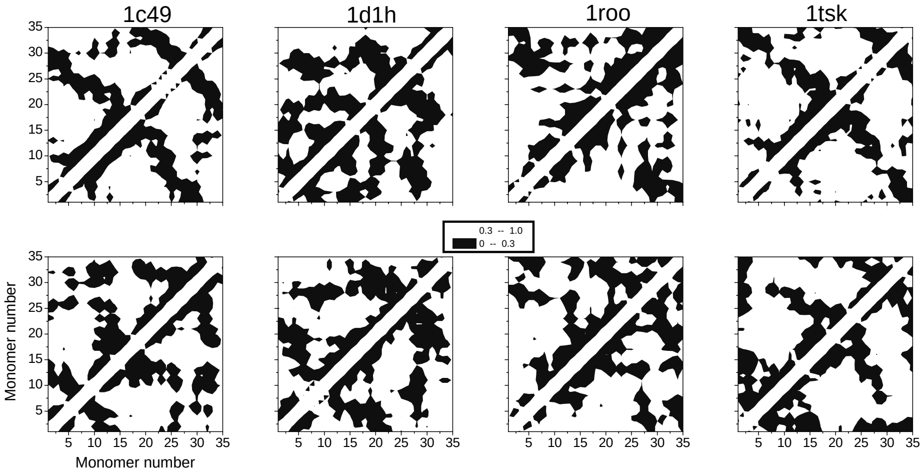

Figure 6. Contact maps for four globular proteins, namely 1c49, 1dh1, 1r00, and 1tsk. The spatial scales were properly translated and rescaled in such that distances $d_{i,j}$ fit the range $0 \leq d_{ij} \leq 1$ (real protein and models). Black and white regions mean distances smaller or equal to 0.3 and larger than 0.3, respectively. Real proteins and models are to be compared by columns. See text for details.

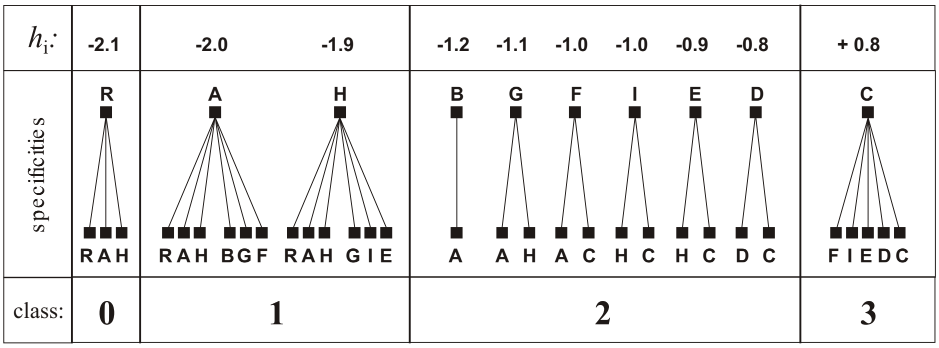

Figure 7. Steric specificities and hydrophobic level for a 10-letter alphabet, namely: {R, A, H, B, G, F, I, E, D, C}. The lines connecting pairs of letters indicate the residues allowed to be first-neighbors in the cubic lattice.The hydrophobic level for each "residue" is indicated at the top of the Figure. When the chain is in the native configuration , those monomers making zero, one, two, and three contacts with the solvent are chosen, respectively, from the classes **0; 1; 2;** or **3**.

Initially, using only inter-monomers contact potential just as $\varepsilon_{i,j} = h_i + h_j$, the heat capacity curve is broad and its peak occurs at about $k_B T = 0.9$, as shown by open circles in Fig.8. This behavior reveals that, as the temperature modifies, the system exchanges relatively small amounts of energy with its surroundings. But, if appropriate steric constraints are introduced, $\varepsilon^*_{i,j} = \varepsilon_{i,j} + c_{i,j},$ the heat capacity curve changes drastically (solid circles): two distinct temperatures stand out, namely that corresponding to $k_B T_{\max} \simeq 1.5$ and $T_\kappa < T_{\max}$; the temperature $T_{\max}$ corresponds to the peak of the heat capacity, and at $k_B T_\kappa \simeq 1,$ where perturbations are observed (single solid circle out of the curve in Fig. 8). Such fluctuations at T_κ have a singular meaning because the simulations always started with the chain in native structure (except for some checking runs), which corresponds to unfolding-like computational experiments. For details, see ref. 13.

The change on the chain's configurational space, imposed by the steric constraints, is the cause for the remarkable transformation on the shape of the heat capacity curve. To follow details of such alterations, some aspects of the configurational activity as function of temperature are shown in Fig. 9. Thus, Ψ is defined as the average number of contacting first-neighbor monomers, normalized by the total number of contacts for any CSA configuration , which is 28, so $0 \leq \Psi \leq 1$. The behavior Ψ with temperature, when the total number of contacts (native and non-native) is considered, is represented by Ψ_u for the system without steric constrains (open squares in Fig. 9), and by Ψ_c for the system with steric constraints (solid squares), hereafter denominated as *unconstrained* and *constrained* systems, respectively. Ψ_u decreases smoothly as the temperature increases, in the interval $0.5 < k_B T < 3.0$, whereas Ψ_c presents an accomplished sigmoidal shape. Steric constraints effect indicate that part of the conformational space, corresponding to globular-like conformations, was significantly affected: the number of configurations that link the distended chain

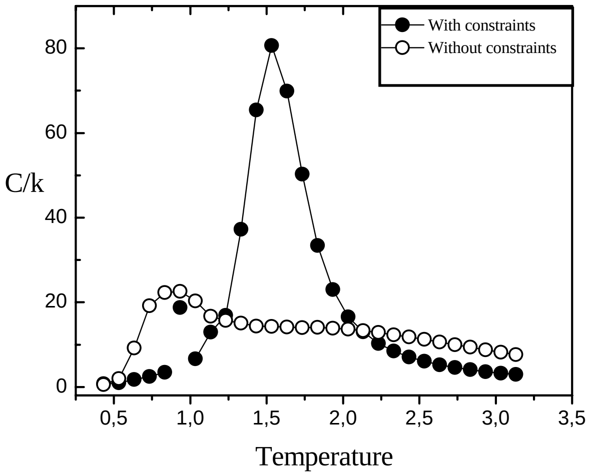

Figure 8. Heat capacity as a function of k_BT (arbitrary energy units) for the system with constraints (steric specificities) and without constraints, solid and open circles, respectively: the remarkable change on the shape of the curve indicates that the chain's configurational activity is substantially distinct for each system. Note that at $k_BT_k = 0.93$, for the system with constraints (solid circles), the amount C/k_B depends on the initial conditions (see text); for most of all other values of T the discrepancy between the results for independent simulations is smaller than 3%.

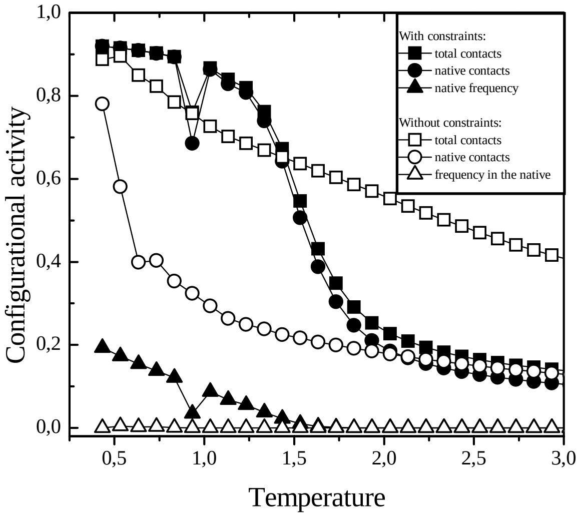

Figure 9. The chain configurational activity as a function of k_BT. Solid and open marks refer to the system with and without constraints (steric specificities), respectively. The solid and open squares (■ and □) represent the normalized average number of total contacts for the unconstrained (Ψ_u) and constrained (Ψ_c) system, respectively; the solid and open circles (• and ◦) represent the normalized average number of native contacts for the constrained ($\Psi_c^{(n)}$) and unconstrained ($\Psi_u^{(n)}$) system, respectively; and the solid and open triangles (▲ and △) are the relative frequency in the native state for the constrained (Φ_c) and unconstrained (Φ_u) system, respectively.

configurations from the most compact ones is severely reduced, which explains the sharp peak observed in the heat capacity curve. Such peak indicates that the chain's internal energy and entropy exhibit a jump, rapidly changing its corresponding amounts for temperatures about $T = T_{\max}$.

Now, let us consider the behavior of the average relative number of *native contacts* for constrained $\Psi_c^{(n)}$ and unconstrained $\Psi_u^{(n)}$systems. For the unconstrained one, a relatively low value for the average native contacts is observed for most temperatures, as shown in Fig. 9 (open circles). But for low enough temperatures, when the globule is very compact, namely for $\Psi_u > 0.8$, the average number of native contacts is significantly enlarged, with $\Psi_u^{(n)}$ quickly approaching Ψ_u but still $\Psi_u^{(n)} < \Psi_u$. A close look at the configurational evolution along the simulation showed that, even though very compact, that is $\Psi_u > 0.8$, the globule shows significant malleability: the amount $\Psi_u^{(n)}$ oscillates intermittently between 15 and 80%, whereas the instantaneous Ψ_u changes continuously from 60 to 100%.

Now, for the constrained system, the number of native contacts $\Psi_c^{(n)}$ (solid circles in Fig. 9) closely follows Ψ_c (solid squares). For $T < T_\kappa$, almost all contacts are native, that is, the condition $\Psi_c^{(n)} = \Psi_c$ is practically satisfied; but even for temperatures as high as $k_BT > 2$ most of the contacts are native contacts, as displayed by Fig.9. This result should be understood as an effect of steric specificities: for $k_BT > 1.5$ the radius of gyration for the constrained system is significantly larger than that for the unconstrained one [16], as depicted in Fig. 10. So, in average, many chain contacts are local contacts, but such contacts are restricted by the steric constraints that favor the native ones because of the design of sequence. This result also indicates that steric constraints work as a folding guide, inducing the chain to native contacts, even at higher temperatures above $T_{\max}$.

As a remarkable result, we point out that at the peak of heat capacity, $k_BT_{\max} \simeq 1.5$, the average number of native contacts approaches 50%, that is $\Psi_c^{(n)} \simeq 1/2$; Fig.9. Therefore, $T_{\max}$ can be seen as the temperature that separates two distinct behaviors of the configurational activity: below $T_{\max}$, the configurational activity –limited by steric constraints and relatively small thermal fluctuations– defines a compact globular shape for the chain ($\Psi_c > 1/2$), and quickly becomes denser for smaller temperatures, as for increasing temperatures above $T_{\max}$ the chain's globular shape is destroyed because, at this moment, the distended configurations are statistically more significant.

Finally, we analyze the relative frequency Φ at which the chain is found in the native state. It is the ratio $\Phi = \phi^{(n)}/\phi$ between the number $\phi^{(n)}$ of times the chain was found in the native structure, and the total number ϕ of configurations. For the unconstrained system, the native configuration can eventually be visited, but it is unprovided with enough stability, that is, $\Phi_u < 10^{-5}$ for all temperatures $T > T_k$; open triangles in Fig.9. However, the fact that Φ_u is not exactly zero has an important meaning; it suggests that the hydrophobic-type potentials, such as $\varepsilon_{i,j} = h_i + h_j$, are efficient in compacting the chain and reaching the native state, although they fail to sustain it properly in that state. Howe-

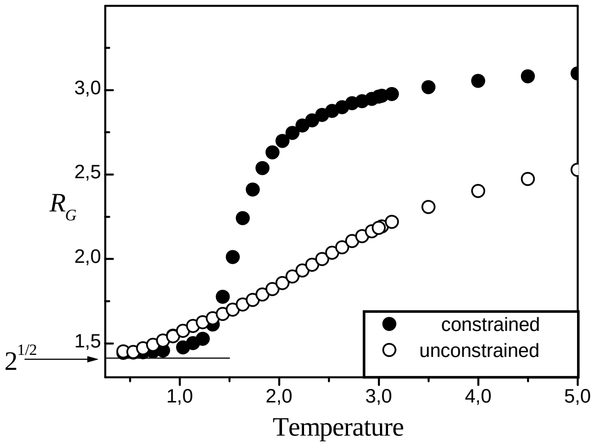

Figure 10. The radius of gyration R_G as a function of k_BT. For the constrained system, at low temperatures corresponding to $k_BT \lesssim 1.3$, R_G is reduced with respect to the unconstrained system owing to the synergism between energetic (local) minima and topological restrictions. But, above $k_BT = 1.3$ the Boltzmann factor becomes systematically less influential and so the steric constraints considerably affect the original configurational space, swelling the globule. As the steric specificities do not allow many of the local contacts, this effect persists even for $T \rightarrow \infty$.

ver, if appropriate steric interactional specificities are introduced, they work as a type of topological labyrinth for the native configuration . This configurational barrier increases its efficiency as T decreases from $T_{\max}$, and so the relative frequency Φ_c in the native state (solid triangles) assumes significant values, reaching 10% about $k_BT = 1$; numerically Φ_c is at least five orders of magnitude larger than Φ_u. Note that just at T_κ, the value Φ_c is smaller than the curve tendency should suggest, which agrees with the comments above, regarding heat capacity.

The same effect on the heat capacity was observed for all CSA target structures, characterized by distinct contact order χ and other topological attributes of the native structure, simulated through unfolding (and also some folding) computational experiments

Results qualitatively equivalent to those described here were observed for many CSA target structures studied[17]. These target structures used in the unfolding (and also some folding) computational experiments were selected in order to cover the entire range of topological attributes, as contact order χ. It was observed that the temperature $T_{\max}$ of the peak of heat capacity, as well as its values at $T_{\max}$ change with χ. As a rule, $T_{\max}$ slightly increases with χ, but other topological characteristics also may be influential, such as the number of structural patterns resembling secondary structures. Yet, with respect to the constrained system discussed here, the time to reach the native state for the first time is smaller than $T_{\max}$, quickly becoming larger as the temperature deviates above or bellow it.

4 Coments and conclusion

In the present work, the hydrophobic effect and classical thermodynamic results of protein folding / unfolding are used as evidence to propose *entropic forces* and *steric constraints* as two basic ingredients for the folding process; these premisses emphasize water as the *protagonist* of the folding process. The entropic forces, or hydrophobic effect, are originated from the versatility of water molecules in re-arranging themselves, as well as their surroundings, in order to minimize the loss of hydrogen bonds. The combination of chemical specificities (polar, nonpolar and neuter residues) and steric interactional specificities of the residues (size and shape), encoded along the chain, are employed here in two simplified (lattice and off-lattice) models; the results from Monte Carlo simulations are compared against properties of real proteins, such as its native structures and characteristic thermodynamic fundamental behavior. First, an off-lattice model is employed to estimate the effectiveness of entropic forces in producing a malleable globule and driving the chain through configurations that intermittently approach the native conformation. Then, a lattice model is used to show that contact energy based in pure hydrophobic potentials may be efficient, indeed, in packing the chain and in finding the native structure, and also to confirm that this kind of energetic interaction fails to provide configurational stability to the globule. A heuristic set of steric specificities is then added to the hydrophobic potential and it is shown that such steric interactional specificities help to select folding pathways and improve the overall stability condition of the globule, in the native structure. Through comparisons between two sets of Monte Carlo simulation results, it is shown that suitable steric specificities dramatically change the system's configurational activity. This effect has the following consequences: (*i*)– it transforms the original broad curve of the heat capacity, obtained using a pure hydrophobic-type potential as pair contact energy, into a peaked and symmetric curve; and (*ii*)– it significantly increases the frequency in which the chain stays in the native state in five or more orders of magnitude.

The results presented here suggest that the folding problem of small globular protein can be thought as a process in which the mechanism to reach the native structure and the requirements for the globule stability are uncoupled. In this view, the stereochemical code, expressed through the hydrophobic pattern and the steric interactional specificity of each residue, provide the governing mechanism through which the chain reaches the native state, which must then be considered as a special and unique state. Once in the native conformation, the steric specificities of the residues also work as hindrances, topologically trapping the chain in its native conformation, as it was shown in this work. However, the native state is indeed very special; other energetic ingredientes, not added explicitly in the present model, start to act exclusive and cooperatively in the direction that maximizes the stability conditions for the globule: At the native conformation, most of the intra-chain hydrogen bonds are protected from the medium and, as the competition with the solvent is minimized, they effectively contribute to the globule stability; additionally, the overall steric complementariness of the residues increases the internal contact area (increasing the dispersion forces), at the same time that it reduces the external contact with the solvent, also producing a net contribution for the stability of the globule.

Acknowledgments

This work was supported, in part, by grant from FAPESP, Proc. No. 00/12107-4.

References

[1] W. Kauzmann, Nature **325**, 763 (1987).

[2] N.T. Southall and K.A. Dill, Phys. Chem.B **10**(6), 139 (2000).

[3] N.T. Southall, K.A. Dill, and D.J. Haynet; J. Phys. Chem. B **106**, 521 (2002).

[4] C. Tanford, *The hydrophobic effect* John Wiley and Sons, New York, 1980.

[5] D. Chandler, Nature **417**, 491 (2002).

[6] K. Lum, D. Chandler and J. D. Weeks; J. Phys. Chem B **103**, 4570 (1999).

[7] M. Daune, *Molecular Biophysics - Structures in motion* Oxford University Press, New York, 1999.

[8] M. Karplus and E. Shakhnovich in: *Protein Folding*, ed. by T.E. Creighton, W.H.Freeman and Company, New York, (1992), cahpter 4.

[9] [1]P. L Privalov in: *Protein Folding*, ed. by T.E. Creighton, W.H. Freeman and Company, New York, (1992), chapter 3.

[10] J.C. Nelson, J.G. Saven, J.S. Moore, and P.G. Wolynes, Science **277,** 1793 **(**1997).

[11] J.A. Schellman, Biophys. J. **73**(6), 2960 (1997).

[12] L.P. Privalovi, Adv. Prot. Chem. **33**, 167 (1979).

[13] M.E.P. Tarragó, L.F.O. Rocha, R. A. da Silva and A. Caliri, Phys. Rev. E, **67**, 031901 (2003).

[14] A.F.P de Araujo, PNAS **96**, 12482 (1999).

[15] R. A. da Silva, M.A.A. da Silva and A. Caliri, J. Chem. Phys. **114**, 4235 (2001).

[16] A. Caliri and M. A. A. da Silva, J. Chem. Phys. **106**, 7856 (1997).

[17] M.E.P. Tarragó, *Potencial estéreo-hidrofóbico e propriedades topológicas no enovelamento de proteínas*.Tese (Doutorado) - FFCLRP - Universidade de São Paulo (2003).