Protein folding in explicit bulk water

Author: Sotiris Samatas
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Giancarlo Franzese

Abstract: Water is thought to play a crucial role in the process of cold and pressure denaturation of proteins as it has been experimentally confirmed that an important driving force behind the folding process is the minimisation of the protein's hydrophobic surface. Water-protein interactions are therefore a vital ingredient for the understanding of such phenomena. In this report, extending previous results calculated in 2 dimensions, we use Monte Carlo simulations of a 3 dimensions coarse-grain protein model in explicit bulk water to show that the changes of water properties at the interface between the solvent and hydrophobic self-avoiding homopolymer effectively lead to a stability region in the temperature-pressure plane in which the protein is folded. We find that the model is able to reproduce and rationalize the protein folding and the protein pressure denaturation at high and low pressures.

I. INTRODUCTION

Proteins perform most of the body's functions at the cellular level [1]. They have a long string-like structure made up of amino acid residues and are able to carry out their biological function when they are in the so-called *native state*, i.e. when they are properly folded and fully operational [2]. This is also known as the *tertiary structure* of the protein, corresponding to a three-dimensional structure, in contrast to the protein's *primary structure* which simply refers to the sequence of monomeric subunits (amino acid residues) forming the protein.

The process in which a protein unfolds, losing its functionality, is called denaturation. It is a well known fact that in general, a protein can be found in its native state for a given range of temperatures and pressures, beyond which the protein naturally unfolds [3]. At first glance, given the Gibbs free energy $G = H - TS$ of a system, high temperature denaturation can be easily understood as due to an increase in entropy (and therefore minimisation of the free energy), whereas cold and pressure denaturation turn out to be a little more tricky to explain.

This is where the water solvent comes into play. In biological environments proteins are surrounded by water, whose molecules have the characteristic property of being polar. A water molecule is composed of two hydrogen atoms covalently bonded to a single oxygen atom. Once the water molecule is formed, the eight electrons initially corresponding to the oxygen atom tend to stay away from the two electrons forming the covalent bonds due to electrostatic repulsion, thus leading to the formation of the electronegative part of the water molecule; whilst for each of the two hydrogen atoms an electropositive part is formed on the opposite side of the covalent bond (see Fig. 1). This gives rise to the "V" shape of water molecules, enabling the formation of *hydrogen bonds* and introducing interactions amongst themselves.

The presence of an inert hydrophobic body submerged within a given volume of water affects the hydration water (water at the water-protein interface) properties [4], inducing an effective interaction between the protein solute and water solvent.

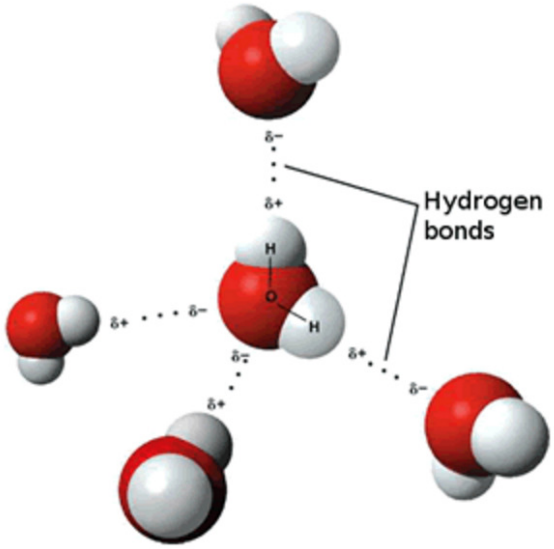

FIG. 1: Hydrogen bonding in water forming a tetrahedral structure.

II. A COARSE-GRAIN MODEL OF PROTEIN FOLDING IN EXPLICIT BULK WATER

The first step in building our many-body water model consists in partitioning the total volume V into a constant number of cells, N, with volume $v \equiv V/N \geq v_0$, v_0 being the van der Waals water excluded volume. We represent our protein as a polymer made of monomers (protein residues) that fits the cell volume v. At high temperature T the protein will occupy a random sequence of nearest neighbour cells. Next we solvate the protein adding water molecules to all the remaining cells in V. In the next section we give details about the water-water interactions and the water-protein interactions.

A. Water-water interaction away from the protein

To simplify the model we replace the coordinates of each water molecule with a discretized density field for each cell i, $n_i = 0, 1$ depending if the cell is gas-like ($v_0/v < 0.5$) or liquid-like ($v_0/v \geq 0.5$). Furthermore, we consider the system homogeneous, hence $n_i = n_j$ for any cells i and j not occupied by a protein residue. We consider our system in the ensemble at constant number of water molecules N_w, constant pressure P and constant temperature T. Therefore v is free to change assuming continuous values (compressible cells) allowing the continuous change of the water-water distance.

Next, we define the Hamiltonian for those water molecules away from the protein first hydration shell (bulk water):

$$H = \sum_{ij} U(r_{ij}) - JN_{HB}^{(b)} - J_\sigma N_{coop} \tag{1}$$

The first term of the Hamiltonian refers to the van der Waals interaction between every pair $i - j$ of molecules separated by a (continuous) distance r_{ij} between the pair's oxygen atoms (O-O). Hence we have:

$$U(r_{ij}) \equiv \begin{cases} \infty & for \quad r_{ij} < r_0 \\ 4\epsilon\left[\left(\frac{r_0}{r_{ij}}\right)^{12} - \left(\frac{r_0}{r_{ij}}\right)^{6}\right] & for \quad r_c > r_{ij} \geq r_0 \\ 0 & for \quad r_{ij} \geq r_c \end{cases} \tag{2}$$

Where $r_0 \equiv v_0^{1/3} = 2.9$ Å is the water van der Waals diameter $\epsilon \equiv 5.8$ kJ/mol and $r_c \equiv 6r_0$ is the cutoff.

The second term accounts for the energy associated to the directional component of the hydrogen bonds (HBs) present in the bulk water, J being the energy of each HB and $N_{HB}^{(b)}$ the total number of bulk HBs:

$$N_{HB}^{(b)} \equiv \sum_{<ij>}^{NN} n_i n_j \delta_{\sigma_{ij},\sigma_{ji}} \tag{3}$$

The sum extending over nearest neighbours (NN). While $n_i n_j$ makes sure that for a hydrogen bond to exist both cells must be liquid-like, $\delta_{\sigma_{ij},\sigma_{ji}}$ makes up for the fact that two molecules are H-bonded in only one sixth of all the possible angular configurations made by the H in the O-H-O plane. The angle (ϕ) is defined as the one formed between the hydrogen atom and the straight line unifying the two oxygen atoms forming the bond. Therefore hydrogen bonding is possible for angles of up to $\pm 30°$ (hence: $360°/60°=6$). Based on this information, we assign six possible different values to the bonding index of each cell ($\sigma_{ij} = 0, 1, 2, 3, 4, 5$) that must be equal for nearest neighbour cells in order for them to be able to form a hydrogen bond. We also assume that each water molecule can have up to 4 hydrogen bonds, as is the case corresponding to the water tetrahedral structure (see Fig. 1).

The third term represents the hydrogen bond cooperativity due to the quantum many-body interaction [5] with:

$$N_{coop} \equiv \sum_i n_i \sum_{(l,k)_i} \delta_{\sigma_{ik},\sigma_{jl}} \tag{4}$$

where $(l, k)_i$ goes over all six different combinations of σ_{ik} and σ_{il}, both bonding variables of the same molecules i. Because we choose the energy J_σ of the many-body interaction as $J_\sigma \ll J$, the term in Eq.(4) represents an effective interaction among all the HBs formed by the molecules i that takes place at an approximate temperature J_σ/k_B, once the HBs are formed at a higher temperature J/k_B approximately.

For low enough pressures P water takes the form of an open hydrogen-bonded tetrahedral structure with low density [6]. By increasing the pressure or temperature, hydrogen bonds are broken, consequently leading to an increase of density. We incorporate this behaviour into our model by adding to the free energy an enthalpic term PV where

$$V \equiv Nv_0 + N_{HB}^{(b)} v_{HB}^{(b)} \tag{5}$$

and $v_{HB}^{(b)}$ is the volume associated to a hydrogen bond. Pressure and temperature therefore contribute to enthalpy variations through the formation/destruction of hydrogen bonds with an average enthalpy variation of $Pv_{HB}^{(b)}$ per hydrogen bond.

B. Water-water interactions at the protein interface

We consider the protein to be a hydrophobic homopolymer (made up of the same monomer) with no interactions among its residues other than the excluded volume. Therefore the only effect the protein has in our model is how its presence affects the surrounding water at the interface. Experiments suggest that water-water hydrogen bonds near a hydrophobic residue are more stable than in bulk, compensating the enthalpy gain during the denaturation process [7]. Hence the energy associated to a hydrogen bond at the interface, J_ϕ, has to be greater than the hydrogen bond energy in the bulk, J, ($J_\phi > J$).

The local density and compressibility of the hydration water are also affected by the interfacial interactions [8-10]. This is introduced in the model using a linear dependence on pressure of the average volume change per water-water hydrogen bond at the hydrophobic interface:

$$v_{HB}^{(\phi)}/v_{HB,0}^{(\phi)} \equiv 1 - kP \tag{6}$$

where $v^{(\phi)}_{HB,0}$ is the volume change associated to hydrogen bond formation in the hydrophobic shell at zero pressure and k is a positively defined parameter. It has been tested that the use of a $v^{(\phi)}_{HB}$ dependance of up to the third order in P has not introduced significant qualitative differences [11] since the pressure values we are interested in are small, being in the range of biological atmospheric pressures. The total volume V of the system is therefore given by:

$$V \equiv V^{(b)} + V^{(\phi)} \equiv V^{(b)} + N^{(\phi)}_{HB} v^{(\phi)}_{HB} \tag{7}$$

III. COMPUTATIONAL METHODS

As already mentioned, we simulate the system fixing the pressure (P), temperature (T), and number of cells (N), that is, working in the NPT thermodynamic ensemble. We choose the density in such a way that $n_i = 1$ for any i making sure the protein has a completely hydrated surface. Given the thermodynamic ensemble we have chosen, the probability of a microstate of energy $E \equiv H$ and volume V is proportional to:

$$p = e^{-\beta(E+PV)}. \tag{8}$$

For the sampling of the phase space we use the Wolff algorithm:

A. The Wolff algorithm

The Wolff algorithm is based on the formation of a cumulated cluster of correlated degrees of freedom. This is done according to the following steps:

1. A random molecule i is chosen and one of its four bonding indexes (or *arms*) σ_{ij} is randomly chosen as the first element of the cluster.

2. Any or the three remaining arms that are in the same state q as the initial arm are added to cluster with probability:

$$p_{J_\sigma} = 1 - exp(-\beta J_\sigma) \tag{9}$$

3. The value q' of the arm of the nearest neighbour to the initial arm is checked and added to the cluster with probability:

$$p_J = 1 - exp(-\beta |J_{eff}|) \tag{10}$$

if $J_{eff} \equiv J - PV > 0$ and $q' = q$ or if $J_{eff} < 0$ and $q' \neq q$.

4. Steps 1,2 and 3 are repeated for any new element added to the cluster until no more elements are added.

5. If $J_{eff} > 0$ a new bonding index value q^* is chosen and all elements of the cluster are flipped to this value; otherwise, if $J_{eff} < 0$ each arm is changed according to:

$$\sigma^{new} \equiv (\sigma^{old} + q^*) \ \ mod\, 6 \tag{11}$$

("6" corresponding to the six different possible values a bonding index can have).

With each Monte Carlo step (a whole loop of the Wolff algorithm) a change in the cell volume is also attempted according to the probability in equation (8). Changes of the cell volume modify the Gibbs free energy of the system in the following way:

$$\Delta G \equiv \Delta U + P\Delta V_r - T\Delta S \tag{12}$$

where ΔV_r refers to the volume change due to the variation of the cells' size and not due to a variation in the number of hydrogen bonds. HBs produce an additional change in volume because of the volume associated to them (v_{HB}). Both the Lennard Jones potential and the entropy of the system are affected by ΔV_r since both depend on the cell size; the former according to equation (2) and the latter according to:

$$\Delta S = 2NT \ln(1 + \Delta v/v) \tag{13}$$

We choose small random cell variations in each step ranging from $-0.01r_0$ to $0.01\mathrm{r}_0$ with an acceptance probability of:

$$p_v \equiv \left\{ \begin{array}{ll} 1 & for \quad \Delta G \leq 0 \\ e^{-\beta \Delta G} & for \quad \Delta G > 0 \end{array} \right\} \tag{14}$$

B. Simulations

For the simulations we choose a protein length of 30 (i.e. the protein is made up of 30 *equal* segments), in an infinite (wrapped) three dimensional space, using cubic cells, therefore each cell has 6 nearest neighbours, able to bond with up to 4 of them. We begin our simulations by generating an initial random configuration for the protein, and starting from high temperatures ($0.6\epsilon/k_BT$) we move along isobars towards lower temperatures (i.e. through an *annealing* process). This is done to ensure that the protein has enough time to reach equilibrium, since starting the simulations of a random initial configuration at low temperatures would most likely result in the protein being in an out-of-equilibrium state (this was effectively tested through simulations).

To test wether the protein is folded or not (or how much "folding" there is) we count the number of residue-residue contacts and compare it to the maximum possible value corresponding to the case of 100% folding; which is 30 in our case.

The program runs 10000 steps at the beginning of each simulation (at fixed pressure and temperature), looping

through the Wolff algorithm with each step. This is the tolerance time chosen for the system to equilibrate. Once this is done, another 500 steps are performed allowing the protein to move by undergoing 90° or 180° rotations until it "reaches" equilibrium as well, and finally the last 30000 steps are carried out with the protein supposedly in equilibrium (Fig. 2); where the total energy of the system, total volume, number of bulk and cooperative hydrogen bonds, cell radius, pure van der Waals potential energy and number of residue-residue contacts are recorded and saved using 10-step intervals. The average number of residue-residue contacts will give us the folding of the protein at a given temperature and pressure.

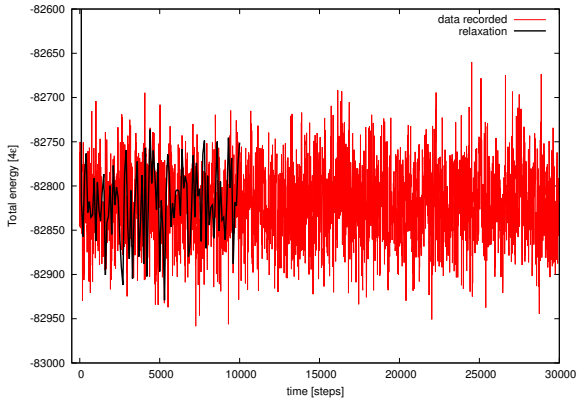

FIG. 2: Red line: Energy fluctuations once the system is at equilibrium; black line: Energy fluctuations during the "equilibration" process.

We fix $v_{HB}^{(b)} = v_{HB,0}^{(\phi)} = v_0/2$ according to experimental suggestions [12], $J_\sigma = 0.05$ (in units of 4ϵ), $J_\phi/J = 1.5$ and the pressure coefficient in equation (6) to: $k = 1v_0/4\epsilon$; meanwhile different values for the hydrogen bond energy J will be tested.

IV. RESULTS AND DISCUSSION

Hawley proposed a theory [13] predicting a close stability region in the pressure-temperature plane with an elliptic shape by hypothesising that proteins can be found in either the unfolded or folded state and that the transition form one state to the other was a reversible process.

We define how folded the protein is by counting the number of contact points between non-subsequent residues (Fig. 3). Our results depict a semi-elliptical shape in the pressure-temperature plane indicating the region where the protein is folded (Fig. 4). All three figures have a similar shape with the difference between them lying in the "shrinking" of the folded region for smaller values of the HB strength J, since the energy term related to hydrogen bonding is less likely to compensate the free energy decrease associated to high temperature and pressure denaturation.

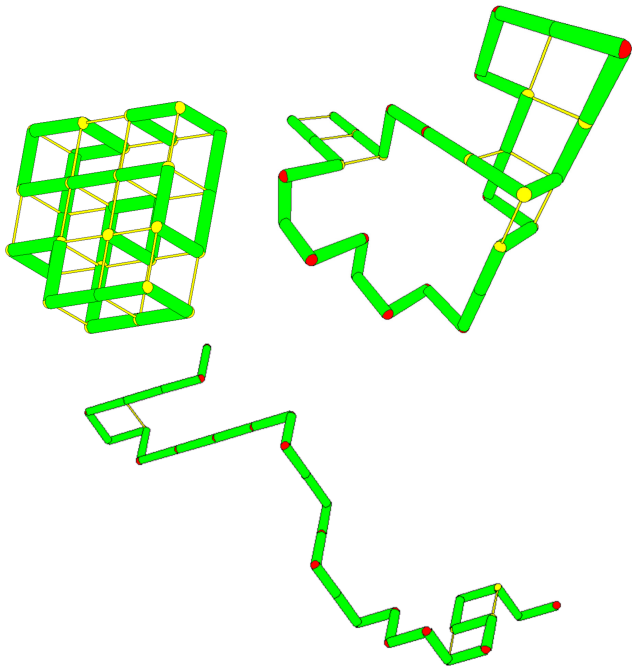

FIG. 3: Examples of proteins folded 100% (top left), 30% (top right), 10% (bottom). Green indicates the protein "backbone". **Fictitious** yellow bonds are drawn to indicate residue-residue contact. Residues in contact with a neighbour are painted yellow while those that are not are painted red.

We notice that the isobaric entropy-driven unfolding for high temperatures is accurately reproduced by the model since an unfolded protein configuration corresponds to a greater number of compatible micro-states.

The isothermal denaturation of the protein for increasing pressure observed is associated to a decrease in the *total* volume (given by equation (7)) due to the breaking of bulk hydrogen bonds (whose abundance would imply the formation of the *low-density* tetrahedral structure in water). A slight increase in the number of surface hydrogen bonds is also present, leading to an increase in volume (even though $v_{HB}^{(\phi)}$ is reduced as a consequence of the high P value according to equation (6)); nevertheless unable to compensate the volume decrease due to the bulk hydrogen bond breaking. During this process there is an increase in the internal energy of the system due to the smaller cell size and bulk hydrogen bond breaking according to equation (1), but it is outweighed by the contribution of the pressure-volume decrease leading to a lower Gibbs free energy.

Finally, the isothermal denaturation for low pressures has to do with the enthalpy minimisation through the formation of more hydrogen bonds at the interface given by:

$$\Delta H = (P v_{HB}^{(\phi)} - J_\phi)\Delta N_{HB}^{(\phi)} \tag{15}$$

which is effectively what we observe in our simulations.

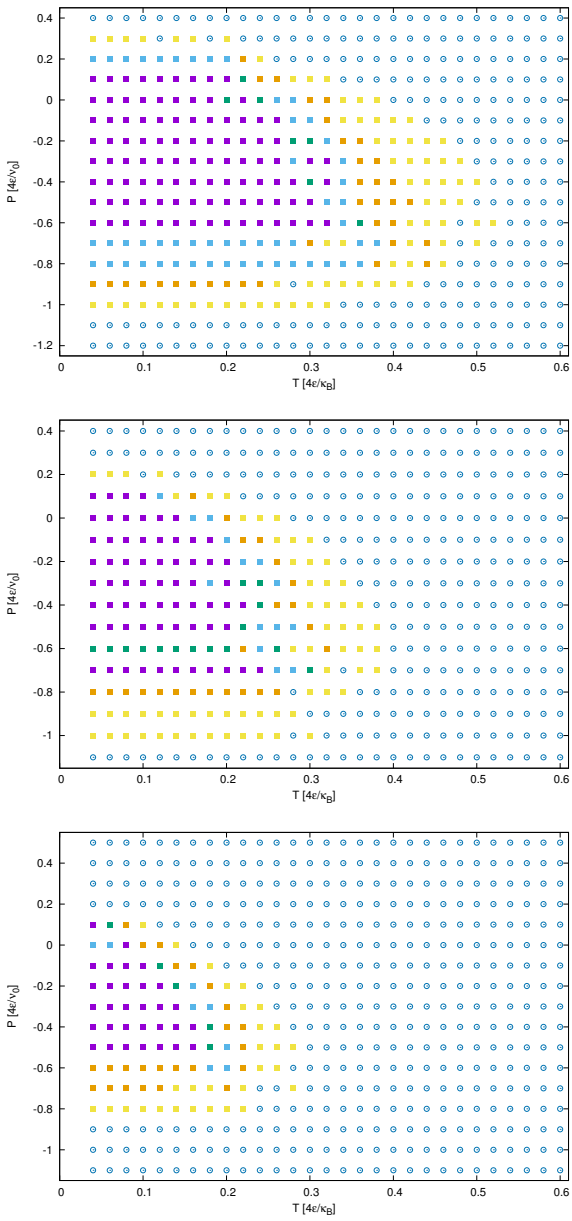

FIG. 4: Pressure-temperature phase diagram for simulations with: $J = 0.75$ (top), $J = 0.5$ (middle), $J = 0.3$ (bottom). Colour indicates the folding percentage: purple for $\geq 70\%$, green for $\geq 60\%$, blue for $\geq 50\%$, orange for $\geq 40\%$, yellow for $\geq 30\%$ and blue circles for $< 30\%$.

V. CONCLUSIONS

Concluding, in this report we have shown using a coarse-grained model of protein folding in explicit bulk (3 dimensions) water, how the pressure and high temperature denaturation mechanisms can arise through the competition of different terms in the Gibbs free energy of bulk and hydration water.

The fact that our model is able to produce accurate results for pressure and high temperature protein denaturation by solely focusing on the water properties, neglecting any protein-protein interactions, highlights the fundamental role that water plays in protein folding.

However, further investigation is needed in the case of cold denaturation, which would eventually lead us to the obtention of a completely elliptical close stability region in the pressure-temperature plain as proposed by Hawley. The answer to this possibly lies in the variation of one of the parameters left constant in our simulations (most probably the relative surface bond strength J_ϕ/J). This is indeed the case for the 2 dimensions version of the present model that has been studied in detail in Ref. [11].

Furthermore, taking into consideration residue-residue interactions would be an interesting detail to add to the model since in the absence of these interactions the protein has several native states, all equivalent due to the fact that folding is entirely based on the number of residue -residue contacts.

Acknowledgments

I would like to thank my advisor, Giancarlo Franzese, for guiding me through the whole process and giving me the opportunity to carry out this project in the field of research. Special thanks to Arne Zantop and Valentino Bianco as well for all the help provided.

[1] Alberts B, Johnson A, Lewis J, et al, *Molecular Biology of the Cell. 4th edition.*, (New York: Garland Science; 2002).
[2] J. R. Claycomb, Jonathan Tran, *Introductory Biophysics: Perspectives On The Living State*, (Jones Bartlett Publishers, Apr 1, 2010).
[3] S. Lapanje, *Physicochemical Aspects of Protein Denaturation.*, (John Wiley and Sons Limited, New York and Chichester. 1978).
[4] Y. Levy and J. N. Onuchic, Annu. Rev. Biophys. Biomol. Struct. 35, 389 (2006)
[5] L. Hernndez de la Pea and P. G. Kusalik, J. Am. Chem. Soc. 127, 5246 (2005).
[6] A. K. Soper and M. A. Ricci, Phys. Rev. Lett. 84, 2881 (2000).
[7] J. G. Davis, K. P. Gierszal, P. Wang, and D. Ben-Amotz, Nature (London) 491, 582 (2012).
[8] D. A. Doshi, E. B. Watkins, J. N. Israelachvili, and J. Majewski, Proc. Natl. Acad. Sci. U.S.A. 102, 9458 (2005).
[9] R. Godawat, S. N. Jamadagni, and S. Garde, Proc. Natl. Acad. Sci. U.S.A. 106, 15119 (2009).
[10] V. M. Dadarlat and C. B. Post, Biophys. J. 91, 4544 (2006)
[11] Valentino Bianco and Giancarlo Franzese, Phys. Rev. Lett. 115, 108101 (2015)
[12] R. C. Dougherty, J. Chem. Phys., 109(17):7372?7378, 1998.
[13] S. A. Hawley, Biochem. 10, 2436 (1971)