Protein Folding Under Confinement

Author: Carlos Torres Jiménez*
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.
Advisor: Giancarlo Franzese

Abstract: The study of proteins near an interface is relevant to understanding their *in vivo* behavior inside biological bodies. Previous studies have shown that the folding process of the protein can be affected by nearby interfaces. Here we perform Monte Carlo simulations of a coarse-grained model of proteins, in explicit water, near a flat hydrophobic interface. Due to the periodic boundary conditions of the system, the interface effectively confines the system in one direction as in a slit pore. We consider four different proteins, two with 60 and two with 100 amino acids (aa), and study their folding in bulk and confined water by calculating the system's free energy. We observe that the 60 aa-long proteins spontaneously fold in bulk and confined water. On the other hand, the 100 aa-long proteins are stable in bulk but unable to refold within our simulation time. For the two shortest proteins, we test the effect of confinement. We observe the absorption of the folded proteins on the hydrophobic interface with no drastic changes in their folding process. We interpret this weak effect as a consequence of the large separation between the slit pore surfaces.

I. INTRODUCTION

Proteins are biological polymers that are fundamental for life. They have specific and wildly diverse tasks. Proteins are made of amino acids. Different sequences of amino acids (primary structures) fold into a few three-dimensional motives (secondary structures) that organize in a large variety of complex (tertiary) structures, some of which are ordered (folded structures), while others are (partially) disordered, giving rise to millions of proteins [1]. Temperature T, pressure P, or changes in protein concentration can induce protein denaturation (unfolding) [2]. Denatured proteins at high concentrations can aggregate [3], possibly leading to numerous diseases such as Alzheimer's, Parkinson's, or Type II diabetes [4].

These processes have been extensively studied in bulk water and, more recently, also under confinement, mimicking the cellular environment and making accessible experimentally protein conformations that are rarely detected in bulk [5-7]. For example, *in silico* simulations explain how a hydrophobic interface can destabilize the native states [5,6]. *In vitro* studies show how the hydrophobicity of the adsorbing interface can change the rate of unfolding and aggregation [7].

Previous results show that 36 aa-long proteins, with a few hydrophobic residues exposed to water in the native state, do not adsorb on the interface [3]. Here, we consider longer protein sequences, four from 60 to 100 aa, and characterize their free energy of proteins in bulk water and near a flat hydrophobic interface. Due to the periodic boundary conditions of the system, the interface effectively confines the system in one direction as in a slit pore. These proteins have been designed in such a way to have large hydrophobic patches exposed to water in their native state. Our goal is to test if these proteins adsorb on the interface and how the confinement affects the stability of their native state.

II. MODEL

We adopt the coarse-grained protein model in explicit water introduced by Bianco and Franzese [2,3,8]. The system volume V is partitioned into N cells, each with an initial volume $v \equiv V/N \geq v_0$, where v_0 is the excluded volume of a water molecule. As described in the following, the formation of hydrogen bonds (HBs) leads to small changes to the volume associated with each cell, defining a density field that coarse-grains the molecular coordinates. Each cell is occupied by a water molecule or a protein residue. Following other authors [9], we neglect the difference in volume between a residue and the surrounding water molecules, and we consider the two-dimensions (2D) model. We fix T and P near ambient conditions and numerically calculate the enthalpy $\langle H \rangle + P\langle V \rangle$ where the $\langle H \rangle$ and $\langle V \rangle$ are the thermodynamic averages of the total energy and volume, respectively, of the system. The total Hamiltonian and volume of the system have two components, the bulk water and the hydrated protein contribution.

A. Franzese-Stanley coarse-grained model for water

The Hamiltonian for the bulk water is

$$H^b_{w,w} = \sum_{ij} U(r_{ij}) - JN^b_{HB} - J_\sigma N^b_{coop}. \qquad (1)$$

The first term in Eq. (1) in a Lenard-Jones potential accounting for the Van der Waals interaction and Pauli's exclusion principle and, for all i and j water molecules at distance r_{ij}, is

$$U(r) \equiv \begin{cases} \infty & if\ r \leq r_o \\ 4\varepsilon \left[\left(\frac{r_o}{r}\right)^{12} - \left(\frac{r_o}{r}\right)^{6}\right] & if\ r_o < r < 6r_o \\ 0 & if\ 6r_o \leq r \end{cases}$$

* Electronic address: ctorresj97@gmail.com

where $r_o \equiv \sqrt{v_0}$ is the closest distance between nearest neighbor (NN) water molecules.

The second term represents the energy of the directional component of the HB. For the formation of the HB between two NN molecules i and j, the angle $\widehat{OOH}$ between them must be less than 30º. Accordingly, only 1/6 of the range [0,360º] of possible values of such an angle are associated with a HB.

Then, we define a bonding index $\sigma_{ij} = 1, \dots, 6$ of the molecule i to its NN molecule j and express the number of bulk water-water HBs as

$$N_{HB}^{b} = \sum_{<i,j>} \delta_{\sigma_{ij}\sigma_{ji}}$$

where $\delta_{ab} = 1$ if $a = b$ or $\delta_{ab} = 0$ if $a \neq b$.

The third term of Eq. (1) is the energy of the cooperative component of the HB due to quantum many-body interactions. It is represented as an effective interaction between each of the six possible pairs of the four bonding indices σ_{ij} of a molecule i. By assuming $J_\sigma \ll J$ we guarantee the asymmetry between the two HB's terms in such a way that the cooperative ordering of the four bonding indices of a molecule is possible only when the four HBs with NN molecules are formed. We express the number of cooperative HBs as

$$N_{coop}^{b} = \sum_{i} \sum_{(kl)_i} \delta_{\sigma_{ik}\sigma_{il}}$$

where the inner sum is over the six pairs of the four indices of the same molecule i.

Finally, to account that the cooperative behavior locally leads to a low-density water structure (that is tetrahedral in 3D), we express the increases of the bulk volume as $V^b \equiv Nv + N_{HB}^{b} v_{HB}^{b}$, being v_{HB}^{b} a fraction of v_0.

B. The Bianco-Franzese hydrated protein model

The proteins are represented as self-avoiding heteropolymers, whose amino acids occupy one cell. Following Ref. [8], we adopt the Miyazawa-Jerningan residue-residue interaction matrix [10] rescaled by a factor 2 to compensate the reduced surface-volume ratio in 2D. The protein-water interaction energy is ε_ζ if the residue is hydrophilic or ε_Φ if hydrophobic. We assume that only water molecules NN to residues are affected in their properties by the protein. The Hamiltonian for these hydration water molecules is

$$H_{w,w}^{h} \equiv -J^{\zeta} N_{HB}^{\zeta} - J^{\Phi} N_{HB}^{\Phi} - J^{\chi} N_{HB}^{\chi} + \\ -J_\sigma^{\zeta} N_{coop}^{\zeta} - J_\sigma^{\Phi} N_{coop}^{\Phi} - J_\sigma^{\chi} N_{coop}^{\chi} \qquad (2)$$

where the ζ terms hold when the two water molecules are NN to hydrophilic residues, the Φ terms when both are NN to hydrophobic amino acids, and the χ when one NN residue is hydrophilic, and one is hydrophobic. N_{HB}^{α} and N_{coop}^{α} are the number of HBs and cooperative HBs, respectively, in the hydration shell.

Experiments [11] show that water near hydrophobic interfaces form stronger HBs and has larger local compressibility. To account for these effects, the model associates a P-dependent volume increase $v_{HB}^{\Phi}/v_{HB,0}^{\Phi} \equiv 1 - k_1 P$ for each HB in the hydration shell, where $v_{HB,0}^{\Phi}$ is the corresponding value at $P = 0$ and k_1 the compressibility factor. Hence, the total volume of the hydrated protein plus the bulk water is

$$V \equiv V^b + V^{\Phi} \equiv V^b + N_{HB}^{\Phi} v_{HB}^{\Phi}.$$

We adopt here the simplified version of the Bianco-Franzese model where hydrophilic residues do not affect the local compressibility of the hydration water, hence $v_{HB}^{\zeta} \equiv v_{HB}^{b}$. For the mixed χ case we use averaged parameters as described below.

C. Model's parameters

Following Ref. [8], we set $\varepsilon \equiv 5.8$kJ/mol for the Lenard-Jones potential and define dimensionless units, $T^* \equiv k_B T/\varepsilon$, $P^* \equiv v_0(P - P_{atm})/\varepsilon$, and $F^* \equiv F/\varepsilon$, where F is the Monte Carlo estimate of the Gibbs free energy calculated as described below. With our choice of the parameters (Table I), the thermodynamic state at $T^* = 0.3$ and $P^* = 0$ corresponds to near-ambient conditions.

TABLE I: Parameters for bulk water and the hydrated protein models.

$\frac{J}{8\varepsilon}$	$\frac{J_\sigma}{8\varepsilon}$	$\frac{v_{HB}^{(b)}}{v_o}$	$\frac{J^{\Phi}}{8\varepsilon}$	$\frac{J_\sigma^{\Phi}}{8\varepsilon}$	$\frac{v_{HB,0}^{\Phi}}{v_o}$	$\frac{k_1\varepsilon}{v_o}$	$\frac{\varepsilon^{\zeta}}{8\varepsilon}$	$\frac{\varepsilon^{\Phi}}{8\varepsilon}$
0.3	0.05	0.5	0.55	0.05	0.5	1	0.2	0

The other parameters are $J^{\zeta} \equiv J$, $J_\sigma^{\zeta} \equiv J_\sigma$, $v_{HB}^{\zeta} \equiv v_{HB}^{(b)}$, $J^{\chi} \equiv (J^{\zeta} + J^{\Phi})/2, J_\sigma^{\chi} \equiv (J_\sigma^{\zeta} + J_\sigma^{\Phi})/2, v_{HB}^{\chi} \equiv (v_{HB}^{\zeta} + v_{HB}^{\Phi})/2$.

III. METHOD

A. Proteins

We consider two 'snake' proteins with 60 aa and two with 100 aa (Fig.1) selected from a list of 100 previously designed sequences [12] for having (*6A* and *10A*) the most hydrophilic and (*6B* and *10B*) the most hydrophobic surface in the native state. In particular, the hydrophilic surface is 71.4% for the *6A*, 50% for the *6B*, 66.67% for the *10A*, and 47.22% for the *10B*. Their amino acid residues sequences are presented in the appendix.

B. Monte Carlo Algorithm

We perform Monte Carlo (MC) simulations within a volume partitioned into a square lattice of size *L* with periodic boundary conditions (PBC). The hydrophobic interface has a length *L* and is fixed in space. Thanks to the PBC, the protein is confined between two hydrophobic walls. We use $L/r_o = 70$ for *6A* and *6B*, and $L/r_o = 110$ for *10A* and *10B* proteins.

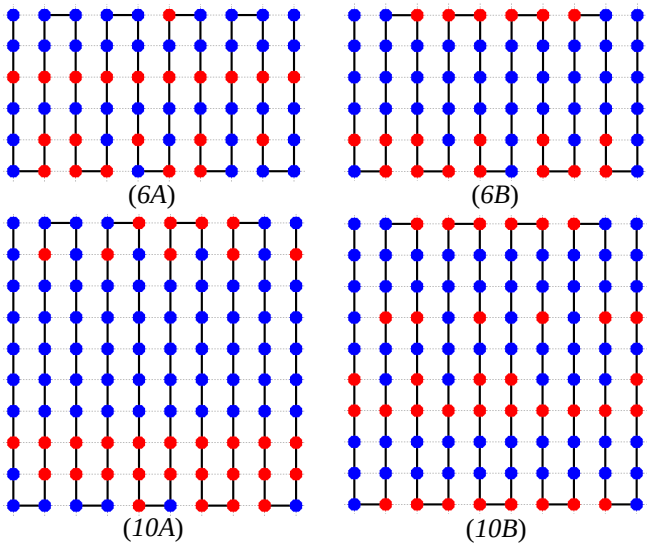

FIG. 1: The four snake-like protein sequences in their folded native structure. 6A and 6B have 60 aa, 10A and 10B 100 aa. The circles represent amino acids, blue for the hydrophilic residues and red for the hydrophobic.

In each MC step, we perform the following steps:

1. We choose a random move from the possible list and attempt to apply it to our protein. The list includes shift, rotation, crankshaft, and pivoting movements.
2. Choosing a random number m from 0 to L^2, we select the cell number m from the lattice. If it contains a water molecule, we attempt to change one of its four bonding indices σ_{ij}, chosen at random, thus, breaking or forming a HB. If it contains a protein residue, we attempt a protein corner flip.
3. We accept or reject the previous steps with a probability proportional to the exponential of the enthalpy change, due to the step, in units of T^*.

We equilibrate each initial configuration for 10^7 MC steps, produce data for the next 10^7 MC steps, make block averages of the data every 10^3 MC steps, and save configurations every 10^5 MC steps. We repeat the simulation protocol for 100 different initial seeds. Hence, we accumulate 10^4 independent configurations for each case described below. Histograms for the observables, described below, include data from 10^7 MC steps.

C. Observables

We study the protein folding and its free energy F through two observables, i) the number N_c of the protein's native contacts, i.e., the number of contacts between the amino acids matching the protein's native state, and ii) the number M_c of protein's amino acids in contact with the interface. For each observable A, we calculate the probability density distribution $P(A)$ from the normalized histograms of the simulation data and the free energy $F(A) \equiv -k_B T ln\big(P(A)\big)$ as a function of A.

We normalize N_c and M_c by their maxima: 45 native contacts for *6A* and *6B* and 81 for *10A* and *10B*; 60 and 100 interface contacts for proteins with 60 and 100 aa, respectively. Therefore, a native state corresponds to $N_c = 1$, and an unfolded protein fully adsorbed onto the interface to $M_c = 1$.

IV. RESULTS

A. Simulations from the protein's native state

First, we simulate the systems using an initial configuration with the protein in its native conformation. These simulations allow us to test the stability of the native state for the designed protein (Fig. 2). We find a global minimum in F at the N_c=1, confirming that the native state is stable for the four proteins.

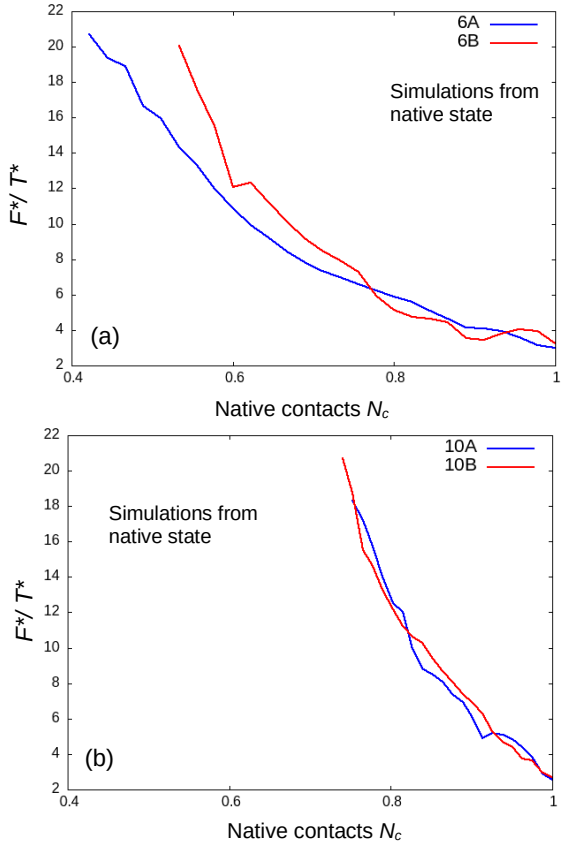

FIG. 2: Free energy F^*/T^* as a function of the normalized native contacts N_c for the proteins (a) *6A* (blue) and *6B* (red), and (b) *10A* (blue) and *10B* (red), starting from a native state. The minima are at $N_c = 1$, confirming that the four proteins are stable in their native states.

A. Simulation from an unfolded state

We simulate the four proteins from a random unfolded state to test if they refold into the native conformation within the 2x10^7 MC steps we perform for each case (Fig. 3). The proteins *6A* and *6B* refold into their native state as the minimum of the free energy F is at $N_c = 1$. Both proteins have a shallow minimum before the global minimum, with the two regions separated by a small free-energy barrier. In particular, *6B* has a meta-stable state near $N_c = 0.9$ with free energy close to the native state, suggesting that the protein could fluctuate between the two states.

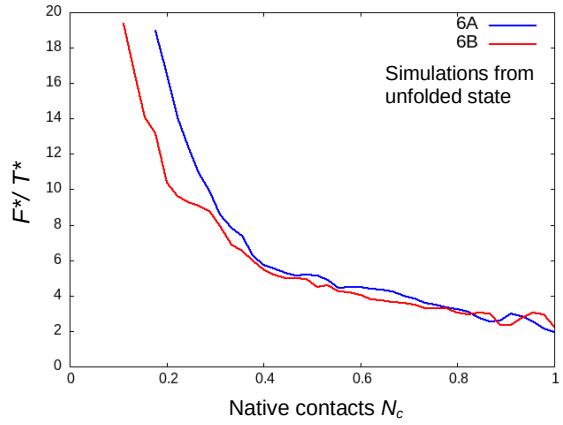

FIG. 3: As in Fig. 2a, but starting the simulations from an unfolded state. The minima at $N_c = 1$ confirm that the two proteins *6A* and *6B* refold in their native states, ensuring they are well designed.

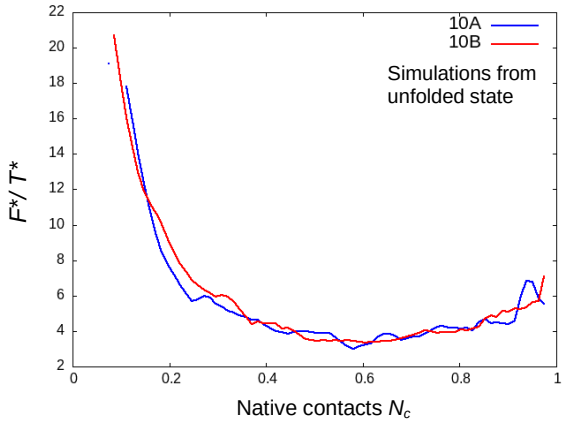

FIG. 4: As in Fig. 3. The free energy minima for the proteins *10A* and *10B* are at $N_c \cong 0.6 < 1$. *Hence* they do not refold within our simulations, being trapped within a (40%) misfolded state.

On the other hand, we find that the two longest proteins, *10A* and *10B*, cannot refold within $2\text{x}10^7$ MC steps, and are trapped within a misfolded state, being separated from the native state at $\boldsymbol{N_c = 1}$ by a sizeable free-energy barrier (Fig. 4). Although we cannot exclude that they would refold for more extended simulations, we do not use them for the confined analysis. Therefore, we focus on the two 60 aa-long proteins in the following.

B. Simulation under confinement

Next, we confine the proteins 6A and 6B as described above. We first check if the confinement affects their ability to refold, observing a weak destabilization of the native state and a minor effect on the second most stable conformation (Fig.5). In confinement, *6A*'s refolding weakly slows down due to a change of slope in the F profile, while *6B* quickly starts the refolding process but compensates around $N_c = 0.3$.

Then, we estimate the free energy as a function of the normalized interface contacts M_c (Fig. 6). We observe that both proteins have a global minimum at $M_c = 0$ (no absorption) separated by a large free-energy barrier from a region with several local minima. The two deepest among these minima are at $M_c = 0.10$ and 0.15. Given the normalization factor (60), these values of M_c correspond to the proteins adsorbed on the short and the long side, respectively, of their folded state (Fig. 7). Therefore, the proteins can explore the conformational space and fold before adsorbing, despite the (weak) confinement.

Although the two free-energy profiles are quite similar, we observe that the minimum for the short-side absorption of the *6B* protein has a larger separation barrier from states with smaller M_c, suggesting possible differences in the adsorption kinetics of the two proteins.

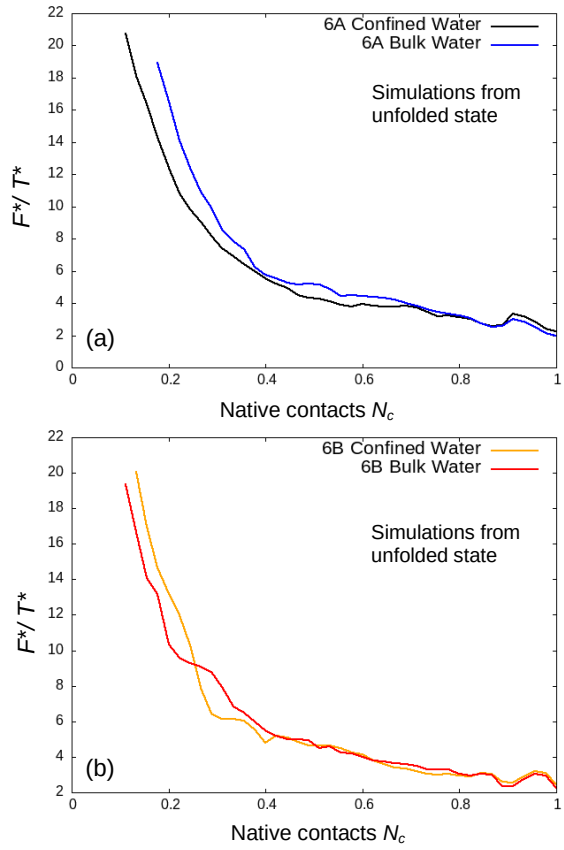

FIG. 5: Free energy F^*/T^* as a function of the normalized native contacts N_c for the proteins (a) *6A* confined (black) compared with the bulk case (blue) in Fig.3, and (b) *6B* confined (orange) compared with the bulk case (red). All the simulations start from an unfolded state.

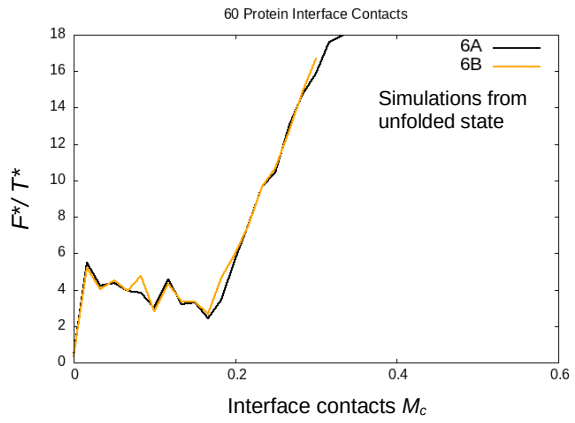

FIG. 6: Free energy F^*/T^* as a function of the normalized interface contacts M_c for the protein *6A* (black) and *6B* (orange) under confinement and initially in unfolded states. The local minima at $M_c = 0.10$ and 0.15 imply adsorption at the interface of the folded proteins' short and the long side, respectively.

V. CONCLUSIONS

We simulate four proteins in an aqueous medium near ambient conditions using coarse-grained models for water and proteins. The proteins have different sequences, ranging from 60 to 100 amino acids, and have been designed to have a 'snake' native state. We test the stability of their native state and their ability to refold within our MC simulation time steps. We find that the two 60 aa-long proteins are both stable and fast folders. We then study how the slit-pore hydrophobic confinement affects their refolding and interface adsorption.

We find that the confinement changes for both proteins the refolding kinetics and weakly destabilizes the native state. The protein *6B* folds slower than the *6A*, both in bulk and confinement, and has a meta-stable state with free-energy *F* close to the native state and separated by a small *F*-barrier, with possible quick fluctuations between the two states.

Although with possibly different adsorption kinetics due to the difference in their primary structure and surface hydrophobicity, both proteins *6A* and *6B* adsorb into the interface after folding. This result suggests that the confinement is weak enough to allow the proteins to explore the conformations space freely. Furthermore, compared with the case of shorter proteins [3], our results confirm that a more significant surface hydrophobicity facilitates the absorption onto the hydrophobic interface. However, the most stable configuration for both proteins *6A* and *6B* is desorbed.

It would be interesting to expand this study by estimating i) how stronger confinement (shorter *L*) would affect the folding and adsorption and ii) how proteins with larger hydrophobic patches would interact with the interface.

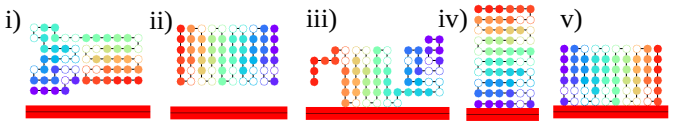

FIG. 7: Examples of *6B* protein conformations visited in our simulations when confined. i) Unfolded and not adsorbed, ii) folded and not adsorbed, iii) unfolded and adsorbed, iv) folded and adsorbed on the short side, and v) folded and adsorbed on the long side. Only one confining wall is represented (red line), with the second far out of the image. Residues are colored from blue to red in sequence order. We use full (empty) circles for hydrophilic (hydrophobic) amino acids.

VI. APPENDIX

The proteins sequences in FASTA encoding are:

- *6A*: *H R L H G S I F E I D Y N R L K M L F W E C E N T K C K M G V W E C E A H K C K M V Q W D C E Y Y K C R M P S W Q C D P*
- *6B*: *N Q P P F Y C F R Q R N I E R E M C A W K D K L V E R E M A G W K D K V L E R E M F C W K D K I Y E S D M C G W T S H H*
- *10A*: *G W S N H H R C G T Y F C D E D N D M G H W R N K K K C F P G F C E E E N E M Y A W K P K K K C L F T I C E D E Q E M A L W K P K R K C L A I V C E D E Q D M I H W R P K R R C V L S V C T Q D Q Q M Y*
- *10B*: *P Q N G H M C P T T L R Q C W E F N R Q I E N M K M C R D I L K D C W E W N K L V E H M K M C R E I A K D C F E W Y K V A E P M K G C K E A L H D C T E W Q K V S E P M K G C R Y I S H D C F D F Y R S*

Acknowledgments

I want to thank Dr. Valentino Bianco for the collaboration in this project and his advice. Also, thanks to my parents, Loli and Jose, my brother, Javi, and my partner, Júlia, for all their emotional support and concerns.

[1] Anishchenko, I., *et al.* (2021). De novo protein design by deep network hallucination. *Nature 600*, 547–552

[2] Bianco, V., & Franzese, G. (2015). Contribution of water to pressure and cold denaturation of proteins. *Physical review letters*, *115*(10), 108101.

[3] March, D., Bianco, V., & Franzese, G. (2021). Protein unfolding and aggregation near a hydrophobic interface. *Polymers*, *13*(1), 156.

[4] Knowles, T. P., Vendruscolo, M., & Dobson, C. M. (2014). The amyloid state and its association with protein misfolding diseases. *Nature reviews Molecular cell biology*, *15*(6), 384-396.

[5] Lucent, D., Vishal, V., & Pande, V. S. (2007). Protein folding under confinement: a role for solvent. *Proceedings of the National Academy of Sciences*, *104*(25), 10430-10434.

[6] Mittal, J., & Best, R. B. (2008). Thermodynamics and kinetics of protein folding under confinement. *Proceedings of the National Academy of Sciences*, *105*(51), 20233-20238.

[7] Green, R. J., Hopkinson, I., & Jones, R. A. L. (1999). Unfolding and intermolecular association in globular proteins adsorbed at interfaces. *Langmuir*, *15*(15), 5102-5110.

[8] Bianco, V., Franzese, G., & Coluzza, I. (2020). In Silico Evidence That Protein Unfolding is a Precursor of Protein Aggregation. *ChemPhysChem*, *21*(5), 377-384.

[9] Kmiecik, S., *et al.* (2016). Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* 116, 7898–7936.

[10] Miyazawa, S., & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, *18*(3), 534-552.

[11] Davis, J. G., Gierszal, K. P., Wang, P., & Ben-Amotz, D. (2012). Water structural transformation at molecular hydrophobic interfaces. *Nature*, *491*(7425), 582-585.

[12] Bianco, V., & Franzese, G. private communication.