A EUROPEAN JOURNAL

CHEMPHYSCHEM

OF CHEMICAL PHYSICS AND PHYSICAL CHEMISTRY

Accepted Article

Title: In silico evidence that protein unfolding is as a precursor of the protein aggregation

Authors: Valentino Bianco, Giancarlo Franzese, and Ivan Coluzza

To be cited as: *ChemPhysChem* 10.1002/cphc.201900904

Link to VoR: http://dx.doi.org/10.1002/cphc.201900904

A Journal of

ChemPubSoc Europe

WILEY-VCH

www.chemphyschem.org

In silico evidence that protein unfolding is as a precursor of the protein aggregation

Valentino Bianco
*Faculty of Chemistry, Chemical Physics Deprtment, Universidad Complutense de Madrid, Plaza de las Ciencias, Ciudad Universitaria, Madrid 28040, Spain**

Giancarlo Franzese
Secció de Física Estadística i Interdisciplinària–Departament de Física de la Matèria Condensada, Facultat de Física & Institute of Nanoscience and Nanotechnology (IN2UB), Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain†

Ivan Coluzza
CIC biomaGUNE, Paseo Miramon 182, 20014 San Sebastian, Spain. IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain.‡

We present a computational study on the folding and aggregation of proteins in an aqueous environment, as a function of its concentration. We show how the increase of the concentration of individual protein species can induce a partial unfolding of the native conformation without the occurrence of aggregates. A further increment of the protein concentration results in the complete loss of the folded structures and induces the formation of protein aggregates. We discuss the effect of the protein interface on the water fluctuations in the protein hydration shell and their relevance in the protein-protein interaction.

I. INTRODUCTION

Proteins cover a range of fundamental functions in the human body: i) the enzymes and hormones are proteins; ii) proteins can carry other biomolecules within the cellular environment; iii) proteins are a source of energy; iv) proteins are necessary to build and repair tissues [1]. A protein is synthesised in the ribosome and, despite the fact that the cellular environment is very crowded, it is capable of reaching its native conformation (mostly dictated by the protein sequence). This process is usually spontaneous–at least for small protein–or is driven by complex interactions with other biomolecules, like the chaperones. Proteins can aggregate after they folded in the native state — through the formation of chemical bonds or self-assembling — or via unfolded intermediate conformations and their propensity to aggregate is related to a series of factors, like the flexibility of the protein structure [2] or the sub-cellular volume where the protein resides [3]. In particular, non-native protein aggregates are commonly formed through a multistep process and are composed by native-like–partially folded intermediate structures [4–7]. Inappropriate protein aggregation represents a crucial issue in biology and medicine, is associated with a growing number of diseases such as Alzheimer's and Parkinson's disease [8–11]. Proteins have evolved to have a low enough propensity to aggregate within a range of protein expression required for their biological activity, but with no margin to respond to external factors increasing/decreasing their expression/solubility [3, 12, 13]. Indeed, protein aggregation is mostly unavoidable when proteins are expressed at concentrations higher than the natural ones.

The mechanisms leading to the failure of the folding process and the formation of potentially dangerous protein aggregates are a matter of large scientific debate [14], where computational tools have largely contributed to elucidate some crucial aspects. Nevertheless, to date an extensive computational study of protein aggregation with all-atom simulations including the solvent explicitly remains not practicable, making the coarse-grain approach an ultimate tool to rationalise those complex systems [15, 16]. In particular, lattice models have been largely exploited to address fundamental questions on protein folding and aggregation [17–27]. According to these studies, the presence of more than one chain leads to aggregate—although each protein contains a considerable fraction of native structure—with consequent loss of the funnel-like free-energy landscape [17, 19, 24].

All these studies, usually performed with a fixed sequence [15, 17] of with Go-like models [19, 20], miss the explicit contribution of water, which instead is supposed to play an essential role in the protein-protein recognition and aggregation [28–33]. Moreover, works implicitly accounting for water show that proteins with hydrophobic amino acids on the surface are prone to aggregate [25], although in nature many proteins present a considerable fraction of hydrophobic and non-polar amino acids on their native surface.

Here we present a computational study on the folding, stability and aggregation of proteins optimised according to the environment. We consider a series of native protein structures, and for each, we determine one or more sequences designed to make the protein fold into the aqueous environment [34, 35]. Each sequence exhibits a different ratio between the number of hydrophilic amino

* vabianco@ucm.es
† gfranzese@ub.edu
‡ icoluzza@cicbiomagune.es

acids exposed to the solvent and the number of hydrophobic amino acids buried into the core of the protein in its native conformation. For each protein, we study its capability to fold as a function of its concentration. We show that the propensity to aggregate is not strictly related to the hydrophobicity of the protein surface. Moreover, for all the designed sequences, at the thermodynamic equilibrium, the concentration at which the (partial) unfolding occurs is lower than the concentration where the aggregation is observed. This phenomenon would suggest a possible two-steps smooth transition between the folded, unfolded and aggregated states of proteins. Finally, focusing on binary systems (i..e solutions with only two proteins), we discuss the extent of the water statistical fluctuations – related to the hydrogen bond dynamics – between two folded or unfolded proteins in relation with the effective protein-protein interaction.

II. THE METHOD

To perform this study we adopt a coarse-grained lattice representation of proteins which is computationally affordable and has been widely adopted in literature [34–41]. A protein is represented as a self-avoiding heteropolymer, composed of 20 amino acids. The residues interact through a nearest-neighbour potential given by the Miyazawa Jernigan interaction matrix [42–44] [45].

The protein is embedded in water, explicitly modelled via the Franzese-Stanley water model which expressly accounts for many-body interactions and has been proven to reproduce, at least qualitatively, the thermodynamic and dynamic behaviour of water [46–50], including its interplay with proteins [34, 35, 40, 51–53]. The coarse-grain representation of the water molecules, adopted to describe water at a constant number of molecules N, constant temperature T and constant pressure P, replaces the coordinates and orientations of the water molecules by a continuous density field and discrete bonding variables, respectively. The discrete variables describe the local hydrogen-bond (HB) formation and its cooperativity, leading to a local open–tetrahedral structure of the water molecules.

Since the protein is composed by hydrophilic ζ and hydrophobic Φ amino acids, we assume that the first interact with water decreasing the local energy, while the second affect the water–water HB in the Φ hydration shell [54]. In particular, we assume that i) the water–water HB at the Φ interface are stronger than HB formed in bulk consistent with the observation that water-water HBs in the Φ hydration shell are more stable and more correlated with respect to the bulk HBs [55–60]; ii) the local density fluctuations at the Φ interface are reduced upon pressurisation, as observed in [61–64].

A detailed description of the model is reported in the next section, and Ref. [34, 40, 53]. Here we resort to using the 2D version of the model because it is faster to simulate and simpler to visualise The analysis of the 3D version is object of current systematic investigation, with preliminary results that confirm our findings qualitatively in 2D: (i) we checked that the water model in 3D is in agreement with our current understanding of the bulk water phase diagram [65]; (ii) we verified that the protein folding analysis performed in 2D can be extended with similar results in 3D [66]; (iii) we also extended the protein design results to 3D with preliminary results consistent with those in 2D [67]. Since the design and folding in 2D is easier than in 3D because the conformational space is smaller [68–71], the unfolding event we discuss in this work should be even easier to find in 3D. We then expect that our results would be not only confirmed in a 3D model but would be even stronger.

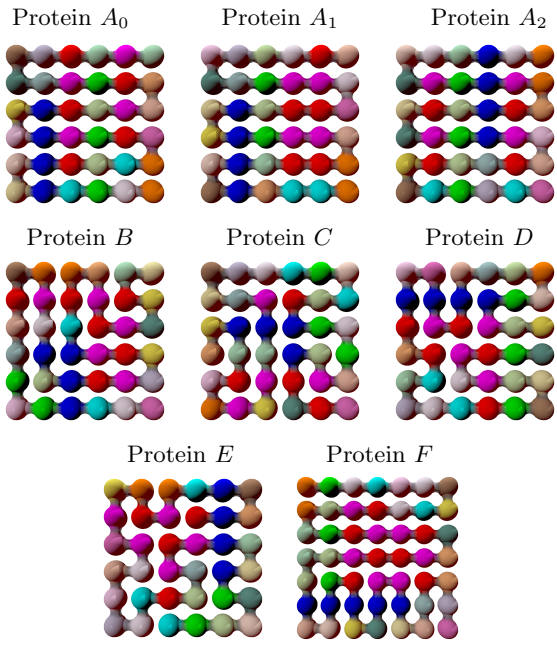

FIG. 1. Here we show all the proteins considered in our simulations. Each amino acid is represented with a different color. By shifting one sequence with respect to the other we establish the maximum overlap between them. We find that A_0 and A_1 have 10 amino acids in the same position ; A_0 and A_2 have 6 corresponding amino acids; A_0 and B have 5 overlapping amino acids; A_0 and C share 5 amino acids; A_0 and D have 6 amino acids in common; A_0 and E have 8 amino acids in common; A_0 and F share 6 amino acids; B and C share 6 amino acids.

We consider 8 different proteins, which we label as A_0, A_1, A_2, B, C, D, E and F, which native states are shown in Fig. 1. Each capital letter in the protein label identifies a different native structure, while different subscript numbers refer to different sequences associated with the structure. Therefore, proteins A_0, A_1 and A_2 share the same native structure but have different sequences. All the native structures have been selected considering maximally compact conformations, composed of 36 or 49 amino acids. Then, for each native structure, the protein sequence has been established through a design

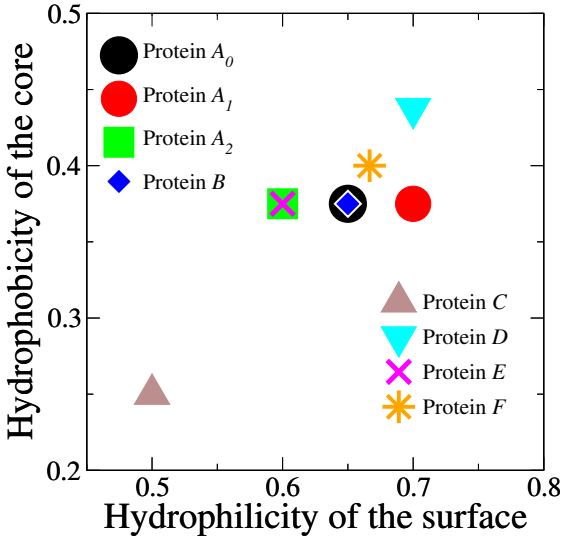

FIG. 2. Composition of the designed proteins. The hydrophilicity (hydrophobicity) of the protein surface (core) is given by the ratio between the number of hydrophilic (hydrophobic) amino acids on the surface (core) and the total number of amino acids exposed to the solvent (buried into the core) when the protein attains its native conformation.

scheme, based on the standard approach introduced by Shakhnovich and Gutin [72, 73] and successfully adopted to design realistic off-lattice proteins [68, 69, 74, 75], but accounting explicitly for the water properties in the protein hydration shell [34]. We perform Monte Carlo (MC) simulations, of the model described in the next section, fixing the pressure and temperature and changing the volume of the system continuously. To simplify the MC algorithm:

(i) We partition the space into a regular square lattice such that each lattice cell has, by definition, the volume coinciding with the proper volume of each molecule (water or residue);

(ii) We assume that all proper volumes are equal (but continuous) if there are no hydrogen bonds, in such a way that by changing the total volume of the system we change (of the same proportion) also the volume of each molecule when there are no hydrogen bonds;

(iii) We assume that all the (local) volume heterogeneities are due to the presence of the hydrogen bonds: the proper volume of each molecule is a linear function of the number of hydrogen bonds made by the same molecules.

Therefore, the space is continuous but partitioned in a regular way that allows us (a) to account for local volume heterogeneity due to the hydrogen bonds, and (b) to simplify the MC algorithm by defining the neighbouring list.

In our isobaric–isothermal simulations at ambient conditions, we keep fixed the protein conformation in its native state and mutate the amino acids, to explore the phase space of sequences. For each sequence, the surrounding water is equilibrated, and the average enthalpy H of the hydrated protein (residue–residue energy plus the average enthalpy of the water molecules in the hydration shell) is computed. The sequence to whom corresponds the minimum value of H is selected as the best folder.

Our MC algorithm includes (I) water moves and (II) protein moves.

(I) The water moves consist of (1) forming or breaking the hydrogen bonds and (2) rescaling the total volume of the simulation box [50].

(II) The protein moves depend on if we are performing (A) design or (B) folding:

(A) Design consists in (1) point mutations of the proteins, (2) residue identity swapping, and (3), after every mutation, several water moves to equilibrate the system [34].

(B) The folding moves are (1) pivot moves, (2) corner flips and (3) crankshaft moves same as used in [43] but in 2D [40, 52].

The design scheme leads to sequences which are not perfectly hydrophilic on the surface and hydrophobic into the core, consistent with what is observed in real proteins [76, 77]. The hydropathy of the designed protein surface and core is shown in Fig. 2, while the full amino acid composition of each sequence is shown in Fig. 6. It is worth to be noted that all the sequences generated differ from each other (the maximum overlap between the sequences is of 10 amino acids[78]), sampling a range of values of the hydrophilicity (hydrophobicity) of the protein surface (core), irrespective of the native structure. Each designed sequence is folded alone at ambient conditions to prove its capability to reach the native state. Once the proteins have been designed, we simulate the folding of multi-protein systems in a range of concentrations $c \in [1\%, 55\%]$, considering homogeneous solutions, i.e. when all the sequences are equal. Along with the simulations, we compute the free energy landscape as a function of the total number of native contacts N_c and inter-protein contacts I_c to study the folding–aggregation competition. The free energies are computed from the natural logarithm of the probability of observing a given value of the order parameter $F(A) = -k_B T \ln P(A)$. In particular we studied the free energy $F(N_c, I_c) = -k_B T \ln P(N_c, I_c)$, where N_c are the numbers of contacts in common with the native structure normalised by total number of native contacts, and I_c are the inter-protein contacts normalised by the maximum number of inter-contacts.

III. THE MODEL

The coarse-grain representation of the water molecules replaces the coordinates and orientations of the water molecules by a continuous density field and discrete bonding variables, respectively. The density field is defined on top of a partition of the volume V into a fixed number N of cells, each with volume $v \equiv V/N \geq v_0$, being $v_0 \equiv r_0^3$ the water excluded volume and $r_0 \equiv 2.9$ Å the water van der Waals diameter. The size of a cell $r \geq r_0$ is a stochastic variable and coincides, by construction, with the average distance between first-neighbour water molecules. The general formulation of the model envisages to each cell i an index $n_i = 1$ or $n_i = 0$ according to the size r (which varies a lot from the gas phase to the super-cooled one), to distinguish when the molecule can form hydrogen bonds (HBs) or not, respectively. Here, since we perform the study at ambient conditions, we assume that all the molecules can form HB, placing $n_i = 1$ to all cells. Therefore such an index is removed from the following expression for the sake of simplicity (for general formulation see for example Ref. [34, 40, 46–53, 79–81]).

The Hamiltonian of the bulk water is

$$\mathscr{H}_{\mathrm{w,w}} \equiv \sum_{ij} U(r_{ij}) - J N_{\mathrm{HB}}^{(\mathrm{b})} - J_\sigma N_{\mathrm{coop}}^{(\mathrm{b})}. \qquad (1)$$

The first term, summed over all the water molecules i and j at oxygen-oxygen distance r_{ij}, is given by $4\epsilon[(r_0/r)^{12} - (r_0/r)^6]$ for $r_0 < r < 6r_0$, $U = \infty$ for $r \leq r_0$, and $U = 0$ for $r \geq 6r_0$ (cutoff distance). We fix $\epsilon = 2.9$ kJ/mol.

The second term of the Hamiltonian represents the directional component of the water-water hydrogen bonds (HB). By assuming that a molecule can form up to four HBs, we discretize the number of possible molecular conformations introducing four bonding indices σ_{ij} for each water molecule i. the variable σ_{ij} describes the bonding conformation of the molecule i with respect to the neighbour molecule j. Each variable σ_{ij} has q possible states, and if $\sigma_{ij} = \sigma_{ji}$ an HB between the molecules i and j is formed, with the characteristic energy $J/4\epsilon = 0.3$. The number of HB is then defined as $N_{\mathrm{HB}}^{(\mathrm{b})} \equiv \sum_{\langle ij \rangle} \delta_{\sigma_{ij},\sigma_{ji}}$, with $\delta_{ab} = 1$ if $a = b$, 0 otherwise. An HB is broken if the oxygen-oxygen-hydrogen angle exceeds the 30°; therefore, only 1/6 of the entire range of values [0,360°] of this angle is associated with a bonded state. Fixing $q = 6$ we correctly account for the entropic loss due to the HB formation.

The third interaction term in Eq. (1) corresponds to the cooperative interaction of the HBs due to the oxygen-oxygen-oxygen correlation. This effect originates from quantum many-body interactions of the HB [82] and in bulk leads the molecules toward an ordered tetrahedral configuration [83]. This term is modelled as an effective interaction–with coupling constant J_σ–between each of the six different pairs of the four indexes σ_{ij} of a molecule i. Hence, we have $N_{\mathrm{coop}}^{(\mathrm{b})} \equiv \sum_{ikl} \delta_{\sigma_{ik},\sigma_{il}}$ which defines the cooperativity of the water molecules. By assuming $J_\sigma \ll J$, we guarantee the asymmetry between the two terms [46].

For any HB formed in bulk, the local volume increases of the quantity $v_{HB}^{(b)}/v_0$. The associated enthalpic variation is $-J + P v_{HB}^{(b)}$, being P the pressure. It accounts for the P disrupting effect on the HB network. Here $v_{HB}^{(b)}/v_0$ represents the average volume increase between high-density ices VI and VIII and low-density ice Ih [46]. Hence, the volume of bulk molecules is given by $V^{(b)} = Nv + N_{\mathrm{HB}}^{(b)} v_{\mathrm{HB}}^{(b)}$.

The water-water hydrogen bonding in the protein hydration shell depends on the hydrophobic (PHO) or hydrophilic (PHI) nature of the hydrated amino acids, and is described by the Hamiltonian

$$\mathscr{H}_{\mathrm{w,w}}^{(h)} \equiv -\left[J^{\mathtt{PHO}} N_{\mathrm{HB}}^{\mathtt{PHO}} + J^{\mathtt{PHI}} N_{\mathrm{HB}}^{\mathtt{PHI}} + J^{\mathtt{MIX}} N_{\mathrm{HB}}^{\mathtt{MIX}}\right] + \qquad (2)$$
$$-\left[J_\sigma^{\mathtt{PHO}} N_{\mathrm{coop}}^{\mathtt{PHO}} + J_\sigma^{\mathtt{PHI}} N_{\mathrm{coop}}^{\mathtt{PHI}} + J_\sigma^{\mathtt{MIX}} N_{\mathrm{coop}}^{\mathtt{MIX}}\right],$$

where $N_{\mathrm{HB}}^{\mathtt{PHO}}$, $N_{\mathrm{HB}}^{\mathtt{PHI}}$ and $N_{\mathrm{HB}}^{\mathtt{MIX}}$ indicate respectively the number of HB formed between two molecules hydrating two hydrophobic amino acids, two hydrophilic amino acids, one hydrophobic amino acid and one hydrophilic amino acid. Analogously $N_{\mathrm{coop}}^{\mathtt{PHO}}$, $N_{\mathrm{coop}}^{\mathtt{PHI}}$ and $N_{\mathrm{coop}}^{\mathtt{MIX}}$ represent the cooperative bonds at the hydrophobic, hydrophilic and mixed interface.

The hydrophobic interface strengthens the water-water hydrogen bonding in the first hydration shell [57, 59, 84, 85] and increases the local water density upon pressurization [57, 86–88]. The first effect is included by assuming $J^{\mathtt{PHO}} > J$ and $J_\sigma^{\mathtt{PHO}} > J_\sigma$. This condition guarantees that the solvation free energy of a hydrophobic amino acid decreases at low temperature T [89]. The second one is accounted assuming that the volume associate to the HB at the PHO interface decreases upon increasing P, $v_{\mathrm{HB}}^{\mathtt{PHO}}/v_{\mathrm{HB},0}^{\mathtt{PHO}} \equiv 1 - k_1 P$ [40]. In this way, the density fluctuations at the PHO interface are reduced at high P. The volume contribution $V^{\mathtt{PHO}}$ to total volume V due the HBs in the hydrophobic shell is $V^{\mathtt{PHO}} \equiv N_{\mathrm{HB}}^{\mathtt{PHO}} v_{\mathrm{HB}}^{\mathtt{PHO}}$. We assume that the water-water hydrogen bonding and the water density at the hydrophilic interface are not affected by the protein. Therefore, $J^{\mathtt{PHI}} = J$, $J_\sigma^{\mathtt{PHI}} = J_\sigma$ and $v_{\mathrm{HB}}^{\mathtt{PHI}} = v_{\mathrm{HB}}^{(b)}$. Finally, we fix $J^{\mathtt{MIX}} \equiv (J^{\mathtt{PHO}} + J^{\mathtt{PHI}})/2$ and $J_\sigma^{\mathtt{MIX}} \equiv (J_\sigma^{\mathtt{PHO}} + J_\sigma^{\mathtt{PHI}})/2$.

Lastly, we assume that the protein-water interaction energy is $-\varepsilon^{\mathtt{PHO}}$ or $-\varepsilon^{\mathtt{PHI}}$, depending if the residue is hydrophobic or hydrophilic, respectively. As reported in Ref. [34], we express the model parameters in units of 8ϵ, and fix the value to $J = 0.3$ and $J_\sigma = 0.05$ (bulk water), $J^{\mathtt{PHI}} = J$ and $J_\sigma^{\mathtt{PHI}} = J_\sigma$ (water at hydrophilic interfaces), $J^{\mathtt{PHO}} = 1.2$ and $J_\sigma^{\mathtt{PHO}} = 0.2$ (water at hydrophobic interfaces), $\varepsilon^{\mathtt{PHO}} = 0$ or $\varepsilon^{\mathtt{PHI}} = 0.48$. Finally, we fix $k_1 = 4$, $v_{\mathrm{HB}}^{(b)}/v_0 = 0.5$ and $v_{\mathrm{HB},0}^{\mathtt{PHO}}/v_0 = 2$. These choices balance the water-water, the water-residue and the residue-residue interactions, making the proteins stable for thermodynamic conditions comprised in the (stable and metastable) liquid phase, including ambient conditions. Moreover, by enhancing the interface interactions, we account for the lower surface volume ratio of

the model (formulated in two dimensions) with respect to a three-dimensional system.

All the results presented in this work have been tested under the change of parameters. In particular, we have decreased the effect of the protein interface on the water-water interaction observing a decrease in the concentration thresholds at which the proteins unfold and aggregate, but the phenomenology observed is substantially the same.

IV. FOLDING VS AGGREGATION IN HOMOGENEOUS PROTEIN SOLUTIONS

In Fig. 3 we show the free energy landscape $F(N_c, I_c)$ of proteins A_0, B and C as function of N_c and I_c simulated in a concentration range $c \in [1\%, 55\%]$. In all the cases we observe that for low concentrations, $c \lesssim 5\%$, the minimum of the free energy correspond to $N_c = 1$ and $I_c = 0$, i.e. all the proteins reach their native folded state and, on average, are not in contact to each other.

By looking at the separate free energy profile as function of N_c (Fig. 4a,b,c) and I_c (Fig. 4d,e,f) (obtained integrating the free energy profiles shown if Fig. 3 along the axes I_c and N_c respectively), respectively indicated with $F(N_c)$ and $F(I_c)$, we can identify three different states for each protein: i) the native state FOL; ii) the unfolded and not aggregated state UNF; iii) the unfolded and aggregated state AGG. The FOL state occurs when all the proteins recover their native conformation and the minima $F_{\min}(N_c)$ and $F_{\min}(I_c)$, respectively of the free energy profiles $F(N_c)$ and $F(I_c)$, occur at $N_c = 1$ and $I_c = 0$. The unfolded and not-aggregated state UNF takes place when the protein looses part of its native contacts leading to $F_{\min}(N_c)$ for $0.8 \lesssim N_c < 1$ while the aggregated state is still less favourable being $F_{\min}(I_c)$ for $I_c = 0$. Similar behaviour is observed for proteins A_1, A_2, D, E, and F shown in the supplementary Figs. S7-S11. The peculiar characteristic of the UNF state is that there are no inter-protein contacts ($I_c = 0$ remains by far the lowest free energy minima Fig.4b).

In Fig. 5 we prove that, for isolated protein pairs, the unfolding starts before the residues can interact directly. Since the proteins are not interacting directly, and in the model, there are no long-range interactions the logical conclusion is that the water is mediating the interaction that stabilises the misfolded states compared to the folded one. When we switched off the water terms in the model the UNF state disappears, and the systems go directly into the AGG state at even lower concentrations c (see Fig. 13 in the Supplementary Information). Hence, it is clear that the water is creating a barrier against aggregation.

Water properties are considered key to prevent protein aggregation in the community. However, the origin of the aggregation barrier has, to the best of our knowledge, never been observed before.

The UNF state holds for quite large values of c, where protein gradually unfold by increasing c. Eventually, at very high concentrations ($c \geq 20\%$ for protein A_0), we observe the appearance of a clear minimum in the free energy ($I_c > 0$ in Fig.3) signifying that we reached an aggregated state AGG. The occurrence of aggregates AGG comes with a loss of the native conformations ($F_{\min}(N_c)|_{N_c<0.8}$) consistent with previous observations [19].

It is important to stress that the concentration thresholds of the $FOL \rightarrow UNF$ and $UNF \rightarrow AGG$ transitions, which we indicate with symbols $c^{(i)}_{FOL\rightarrow UNF}$ and $c^{(i)}_{UNF\rightarrow AGG}$ with $i = A_0, B, C$, depend on the specific sequence (Fig. 4).

By comparing the $UNF \rightarrow AGG$ transition points for proteins A_0 and B (Fig. 4d,e), which have the same fraction of hydrophilic amino acids on the surface and hydrophobic amino acids into the core (Fig. 2), we observe that the protein A_0 is less prone to aggregate with respect to B, since $c^{(A_0)}_{UNF\rightarrow AGG} > c^{(B)}_{UNF\rightarrow AGG}$. On the other hand, by comparing the same transition points between proteins B and C (Fig. 4b,c), we find that both transitions occur at similar values of c within the numerical error. This effect occurs although their surface and core composition are quite different, being the protein C more hydrophobic on the surface and less hydrophobic into the core with respect to the protein B. As long as the proteins are "designed" in a water environment [34], the propensity of proteins to aggregate is not strictly related to the hydrophobic content of their surface.

Similar $FOL \rightarrow UNF$ and $UNF \rightarrow AGG$ transitions are observed also in the proteins A_1, A_2, D, E and F (shown in the supplementary information), and in heterogeneous mixtures [35]. It is interesting to observe that, although proteins A_0, A_1 and A_2 share the same native structure (the sequence of each protein has been obtained with an independent design procedure), the concentration threshold for the $FOL \rightarrow UNF$ and $UNF \rightarrow AGG$ transitions are different in each case.

V. WATER-MEDIATED PROTEIN-PROTEIN INTERACTION

In this section, we focus on the protein-protein interaction mediated by water molecules, for binary systems. In particular, we consider the cases A_0–A_0 proteins and C_0–C_0 proteins (homogeneous systems). In Fig. 5 we report the average number of native contact $\langle N_c \rangle$ [90] as function of the minimum protein distance [91]. We observe that the value of $\langle N_c \rangle$ is constant for a wide range of protein distances, with higher or lower values (respectively for the systems A_0–A_0 and C–C) reflecting the width of the free energy minima and hence the intrinsic stability of the native conformation. The interesting feature in Fig. S5 occurs when $\langle N_c \rangle$ starts decreasing linearly when the protein gets close to each other. These results demonstrate that the proteins start to unfold before interacting directly. Moreover, the transition distances cor-

relate with the protein stability as A_0 (red square points), being overall more stable than the protein C (blue circle points), show an interaction radius smaller ($\sim 3r_0$) with respect the one of protein C ($\sim 5r_0$). We hypothesise that the distance under which $\langle N_c \rangle$ decreases can be considered as the water-mediated the interaction radius of a protein. With this respect, following a recent percolation mapping [81], we have performed a preliminary analysis of the extent of "cluster size of statistically correlated water molecules" at the protein interface, depending on the protein folded/unfolded state and on the protein-protein distance. Such an extent is a measure of the correlation length in water and quantifies the perturbation exerted by the protein on the surrounding water. Our data, shown in Fig. 12 the Supplementary Information, reveal an increase of the cluster size of statistically correlated water molecules when two proteins unfold upon approaching each other. It is also important to notice that the transition distances are close to the distance between proteins at the $FOL \rightarrow UNF$ transition concentrations. Finally, the transition distances correlate with the protein stability as A_0 (red square points), being overall more stable than the protein C (blue circle points), show an interaction radius smaller ($\sim 3r_0$) with respect the one of protein C ($\sim 5r_0$).

VI. CONCLUSIONS

We have presented a computational study on the competition between folding and aggregation of proteins in homogeneous solutions. Employing an efficient coarse-grain model, we have designed a series of proteins according to the water environment at ambient condition. Then, we have tested the capability of each designed protein to fold alone, and in the presence of multiple copies (i.e. changing the protein concentration). The main conclusion of this work is that proteins tend to fold uninfluenced by the presence of other proteins in the solution provided that their concentration is below their specific unfolding concentration $c_{FOL \rightarrow UNF}$. Our simulations predict an unexpected and not previously observed role of the water in the inducing the unfolded regime UNF that is a precursor of the fully aggregated state AGG. We believe that such prediction should be testable first in more detailed protein models and supports the need for new intriguing experiments.

The results presented here have a profound implication on the physiology of the cell where protein aggregation is a fundamental parameter for expression regulation [13]. Using the results obtained in this work, we have shown that indeed that proteins can be optimised to fold and regulated independently of the other proteins [34]. In other words, provided that each protein does not pass their aggregation threshold concentration (counting only that particular species and not the total protein concentration), cross interactions do not affect the folding and the aggregation. Such a simplification of the regulatory process would not have been possible without the barrier that water creates against aggregation. Here, we showed a possible origin for such a barrier that makes water key for life on our planet.

Another important implication of our combined studies (present and in ref [34]) is that water is mediating a long-range interaction that depends on the protein sequence. The unfolding occurs only when the concentration of one protein species crosses over the aggregation threshold (see Fig. 2 in ref [34]) while it is rather unaffected by the interactions with different sequences. Hence, we conclude that water is mediating a long-range molecular recognition process.

Correlated to our study, there is an extensive literature about the role of cellular crowding on aggregation and folding. A sample of pioneering works in the field are [92–96]. The central message of these studies is that the role of the steric crowding does not significantly affect the folding. However, when globular proteins replace crowding agents, the behaviour of the system becomes difficult to explain because of the influence of protein-protein. Our results offer a qualitative description of such an impact, separating the role of water, protein and steric interactions at different concentrations.

Acknowledgements

VB acknowledges the support from the Austrian Science Fund (FWF) project M 2150-N36 and from the European Commission through the Marie Skodowska-Curie Fellowship No. 748170 ProFrost. IC gratefully acknowledges support from the Ministerio de Economià y Competitividad (MINECO) (FIS2017-89471-R). This work was performed under the Maria de Maeztu Units of Excellence Program from the Spanish State Research Agency Grant No. MDM-2017-0720. VB and IC acknowledge the support from the FWF project P 26253-N27. GF acknowledges support by ICREA Foundation (ICREA Academia prize) and Spanish grant PGC2018-099277-B-C22 (MCIU/AEI/ERDF). Simulations have been performed using the Vienna Scientific Cluster (VSC-3).

[1] A. V. Finkelstein and O. B. O. B. Ptitsyn, *Protein physics* (Elsevier, 2016).
[2] A. De Simone, C. Kitchen, A. H. Kwan, M. Sunde, C. M. Dobson, and D. Frenkel, Proceedings of the National Academy of Sciences of the United States of America **109**, 6951 (2012).
[3] G. G. Tartaglia and M. Vendruscolo, Molecular BioSystems **5**, 1873 (2009).

[4] D. Eliezer, K. Chiba, H. Tsuruta, S. Doniach, K. O. Hodgson, and H. Kihara, Biophysical Journal **65**, 912 (1993).
[5] A. L. Fink, Folding and Design **3**, R9 (1998).
[6] C. J. Roberts, Biotechnology and Bioengineering **98**, 927 (2007).
[7] P. Neudecker, P. Robustelli, A. Cavalli, P. Walsh, P. Lundström, A. Zarrine-Afsar, S. Sharpe, M. Vendruscolo, and L. E. Kay, Science **336** (2012).
[8] C. A. Ross and M. A. Poirier, Nature Reviews Molecular Cell Biology **6**, 891 (2005).
[9] F. Chiti and C. M. Dobson, Annual Review of Biochemistry **75**, 333 (2006).
[10] A. Aguzzi and T. O'Connor, Nature Reviews Drug Discovery **9**, 237 (2010).
[11] T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson, Nature Reviews Molecular Cell Biology **15**, 384 (2014).
[12] M. Schröder, R. Schäfer, and P. Friedl, Biotechnology and bioengineering **78**, 131 (2002).
[13] G. G. Tartaglia, S. Pechmann, C. M. Dobson, and M. Vendruscolo, Trends in biochemical sciences **32**, 204 (2007).
[14] C. M. Dobson, Seminars in Cell & Developmental Biology **15**, 3 (2004), protein Misfolding and Human Disease and Developmental Biology of the Retina.
[15] T. Cellmer, D. Bratko, J. M. Prausnitz, and H. W. Blanch, Trends in biotechnology **25**, 254 (2007).
[16] J. Nasica-Labouze, P. H. Nguyen, F. Sterpone, O. Berthoumieu, N.-V. Buchete, S. Coté, A. De Simone, A. J. Doig, P. Faller, A. Garcia, A. Laio, M. S. Li, S. Melchionna, N. Mousseau, Y. Mu, A. Paravastu, S. Pasquali, D. J. Rosenman, B. Strodel, B. Tarus, J. H. Viles, T. Zhang, C. Wang, and P. Derreumaux, Chemical Reviews **115**, 3518 (2015).
[17] R. a. Broglia, G. Tiana, S. Pasquali, H. E. Roman, and E. Vigezzi, Proceedings of the National Academy of Sciences of the United States of America **95**, 12930 (1998).
[18] L. Toma and S. Toma, Biomacromolecules **1**, 232 (2000).
[19] D. Bratko and H. W. Blanch, Journal of Chemical Physics **114**, 561 (2001).
[20] N. Combe and D. Frenkel, The Journal of Chemical Physics **118**, 9015 (2003).
[21] R. Dima and D. Thirumalai, Protein Science **11**, 1036 (2002).
[22] M. T. Oakley, J. M. Garibaldi, and J. D. Hirst, Journal of Computational Chemistry **26**, 1638 (2005).
[23] Y.-Y. Ji, Y.-Q. Li, J.-W. Mao, and X.-W. Tang, Physical Review E **72**, 41912 (2005).
[24] T. Cellmer, D. Bratko, J. M. Prausnitz, and H. Blanch, Proceedings of the National Academy of Sciences of the United States of America **102**, 11692 (2005).
[25] L. Zhang, D. Lu, and Z. Liu, Biophysical Chemistry **133**, 71 (2008).
[26] S. Abeln, M. Vendruscolo, C. M. Dobson, D. Frenkel, and C. Riekel, PloS one **9**, e85185 (2014).
[27] A. Morriss-Andrews and J.-E. Shea, Annual Review of Physical Chemistry **66**, 643 (2015).
[28] E. Y. Chi, S. Krishnan, T. W. Randolph, and J. F. Carpenter, Pharmaceutical Research **20**, 1325 (2003).
[29] A. De Simone, G. G. Dodson, C. S. Verma, A. Zagari, and F. Fraternali, Proceedings of the National Academy of Sciences of the United States of America **102**, 7535 (2005).
[30] M. G. Krone, L. Hua, P. Soto, R. Zhou, B. J. Berne, and J.-E. Shea, Journal of the American Chemical Society **130**, 11066 (2008).
[31] D. Thirumalai, G. Reddy, and J. E. Straub, Accounts of chemical research **45**, 83 (2012).
[32] Y. Fichou, G. Schirò, F.-X. Gallat, C. Laguri, M. Moulin, J. Combet, M. Zamponi, M. Härtlein, C. Picart, E. Mossou, H. Lortat-Jacob, J.-P. Colletier, D. J. Tobias, and M. Weik, Proceedings of the National Academy of Sciences of the United States of America **112**, 6365 (2015).
[33] S. Arya and S. Mukhopadhyay, Biophysical Journal **110**, 398a (2016).
[34] V. Bianco, G. Franzese, C. Dellago, and I. Coluzza, Physical Review X **7**, 021047 (2017).
[35] V. Bianco, M. Alonso-Navarro, S. Di Silvio, Desire ans Moya, A. L. Cortajarena, and I. Coluzza, The Journal of Physical Chemistry Letters **10**, 4800 (2019).
[36] G. Caldarelli and P. De Los Rios, Journal of Biological Physics **27**, 229 (2001).
[37] M. I. Marqués, J. M. Borreguero, H. E. Stanley, and N. V. Dokholyan, Phys. Rev. Lett. **91**, 138103 (2003).
[38] B. A. Patel, P. G. Debenedetti, F. H. Stillinger, and P. J. Rossky, Biophysical Journal **93**, 4116 (2007).
[39] S. Matysiak, P. G. Debenedetti, and P. J. Rossky, The Journal of Physical Chemistry B **116**, 8095 (2012).
[40] V. Bianco and G. Franzese, Physical Review Letters **115**, 108101 (2015).
[41] E. van Dijk, P. Varilly, T. P. J. Knowles, D. Frenkel, and S. Abeln, ArXiv e-prints **116**, 78101 (2015).
[42] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[43] I. Coluzza and D. Frenkel, Physical Review. E **70**, 51917 (2004).
[44] J. Kyte and R. F. Doolittle, J Mol Biol **157**, 105 (1982).
[45] To account for a lower surface-volume ratio in two dimensions we scaled the matrix by a factor 2. Hence, increasing the effective amino acid-amino acid interaction.
[46] K. Stokely, M. G. Mazza, H. E. Stanley, and G. Franzese, Proceedings of the National Academy of Sciences of the United States of America **107**, 1301 (2010).
[47] M. G. Mazza, K. Stokely, S. E. Pagnotta, F. Bruni, H. E. Stanley, and G. Franzese, Proceedings of the National Academy of Sciences **108**, 19873 (2011).
[48] F. de los Santos and G. Franzese, The Journal of Physical Chemistry B (2011), 10.1021/jp206197t.
[49] G. Franzese and V. Bianco, Food Biophysics **8**, 153 (2013).
[50] V. Bianco and G. Franzese, Scientific Reports **4**, 4440 (2014).
[51] G. Franzese, V. Bianco, and S. Iskrov, Food Biophysics **6**, 186 (2011).
[52] V. Bianco, S. Iskrov, and G. Franzese, J. Biol. Phys. **38**, 27 (2012).
[53] V. Bianco, N. Pagès-Gelabert, I. Coluzza, and G. Franzese, Journal of Molecular Liquids **245**, 129 (2017).
[54] The hydration shell is defined by the water molecules which are first-neighbours of the amino acids.
[55] C. L. Dias, T. Ala-Nissila, M. Karttunen, I. Vattulainen, and M. Grant, Physical Review Letters **100**, 118101 (2008).
[56] C. Petersen, K.-J. Tielrooij, and H. J. Bakker, The Journal of chemical physics **130**, 214511 (2009).

[57] S. Sarupria and S. Garde, Phys. Rev. Lett. **103**, 37803 (2009).
[58] Y. I. Tarasevich, Colloid Journal **73**, 257 (2011).
[59] J. G. Davis, K. P. Gierszal, P. Wang, and D. Ben-Amotz, Nature **491**, 582 (2012).
[60] D. Laage, T. Elsaesser, and J. T. Hynes, Chemical Reviews , acs.chemrev.6b00765 (2017).
[61] S. Sarupria and S. Garde, Phys. Rev. Lett. **103**, 37803 (2009).
[62] P. Das and S. Matysiak, The Journal of Physical Chemistry B **116**, 5342 (2012).
[63] T. Ghosh, A. E. García, and S. Garde, Journal of the American Chemical Society **123**, 10997 (2001).
[64] C. L. Dias and H. S. Chan, Journal of Physical Chemistry B **118**, 7488 (2014).
[65] L. E. Coronas, V. Bianco, A. Zantop, and G. Franzese, "Liquid-Liquid Critical Point in 3D Many-Body Water Model," (2016), arXiv:1610.00419 [cond-mat.stat-mech].
[66] S. Samatas and G. Franzese, "Protein folding in explicit bulk water," (2016).
[67] J. Àguila-Rojas and G. Franzese, "Amino acid sequence correlation in optimally designed proteins," (2019).
[68] C. Cardelli, F. Nerattini, L. Tubiana, V. Bianco, C. Dellago, F. Sciortino, and I. Coluzza, Adv. Theory Simulations , 1900031 (2019).
[69] C. Cardelli, L. Tubiana, V. Bianco, F. Nerattini, C. Dellago, and I. Coluzza, Macromolecules **51**, 8346 (2018).
[70] I. Coluzza, J. Phys. Condens. Matter **29**, 143001 (2017).
[71] E. Bianchi, B. Capone, I. Coluzza, L. Rovigatti, and P. D. J. van Oostrum, Phys. Chem. Chem. Phys. **19**, 19847 (2017), arXiv:1705.04383.
[72] E. Shakhnovich and A. Gutin, "Protein Engineering, Design and Selection" **6**, 793 (1993).
[73] E. I. Shakhnovich and a. M. Gutin, Proceedings of the National Academy of Sciences **90**, 7195 (1993).
[74] C. Cardelli, V. Bianco, L. Rovigatti, F. Nerattini, L. Tubiana, C. Dellago, and I. Coluzza, Scientific Reports **7**, 4986 (2017).
[75] F. Nerattini, L. Tubiana, C. Cardelli, V. Bianco, C. Dellago, and I. Coluzza, Journal of Chemical Theory and Computation **15**, 1383 (2019).
[76] L. Lins, A. Thomas, and R. Brasseur, Protein science : a publication of the Protein Society **12**, 1406 (2003).
[77] S. Moelbert, E. Emberly, and C. Tang, Protein science : a publication of the Protein Society **13**, 752 (2004).
[78] The maximum overlap between two sequences is computed shifting and overlapping one sequence with respect to the other, and counting the number of amino acids on both sequences which coincide along the overlapped region.
[79] E. G. Strekalova, J. Luo, H. E. Stanley, G. Franzese, and S. V. Buldyrev, Phys. Rev. Lett. **109**, 105701 (2012).
[80] M. G. Mazza, K. Stokely, S. E. Pagnotta, F. Bruni, H. E. Stanley, and G. Franzese, Proceedings of the National Academy of Sciences **108**, 19873 (2011), https://www.pnas.org/content/108/50/19873.full.pdf.
[81] V. Bianco and G. Franzese, Journal of Molecular Liquids **285**, 727 (2019).
[82] L. Hernández de la Peña and P. G. Kusalik, Journal of the American Chemical Society **127**, 5246 (2005), pMID: 15810860, https://doi.org/10.1021/ja0424676.
[83] A. K. Soper and M. A. Ricci, Phys. Rev. Lett. **84**, 2881 (2000).
[84] N. Giovambattista, P. J. Rossky, and P. G. Debenedetti, Phys. Rev. E **73**, 41604 (2006).
[85] C. L. Dias, T. Ala-Nissila, M. Karttunen, I. Vattulainen, and M. Grant, Physical Review Letters **100**, 118101 (2008).
[86] P. Das and S. Matysiak, The Journal of Physical Chemistry B **116**, 5342 (2012).
[87] T. Ghosh, A. E. Garc\'ia, and S. Garde, Journal of the American Chemical Society **123**, 10997 (2001).
[88] C. L. Dias and H. S. Chan, The journal of physical chemistry. B **118**, 7488 (2014).
[89] M. S. Moghaddam and H. S. Chan, The Journal of Chemical Physics **126**, 114507 (2007).
[90] The average is calculated over all conformations hence the maximum average value will smaller compared to the global minimum that is 1 for all proteins.
[91] The minimum protein distance is the minimum value between all the possible distances among any amino acid of the first protein and any amino acid of the second protein.
[92] B. van den Berg, R. J. Ellis, and C. M. Dobson, The EMBO Journal **18**, 6927 (1999).
[93] A. H. Gorensek-Benitez, A. E. Smith, S. S. Stadmiller, G. M. Perez Goncalves, and G. J. Pielak, Journal of Physical Chemistry B **121**, 6527 (2017).
[94] P. H. Schummel, A. Haag, W. Kremer, H. R. Kalbitzer, and R. Winter, Journal of Physical Chemistry B **120**, 6575 (2016).
[95] M. Feig, I. Yu, P. H. Wang, G. Nawrocki, and Y. Sugita, Journal of Physical Chemistry B **121**, 8009 (2017).
[96] H.-X. X. Zhou, J. Mol. Recognit. **17**, 368 (2004).
[97] P. Kasteleyn and C. M. Fortuin, Physical Society of Japan Journal Supplement, Vol.˜26.˜ Proceedings of the International Conference on Statistical Mechanics held 9-14 September, 1968 in Koyto., p.11 **26**, 11 (1969).
[98] A. Coniglio and W. Klein, Journal of Physics A: Mathematical and General **13**, 2775 (1980).

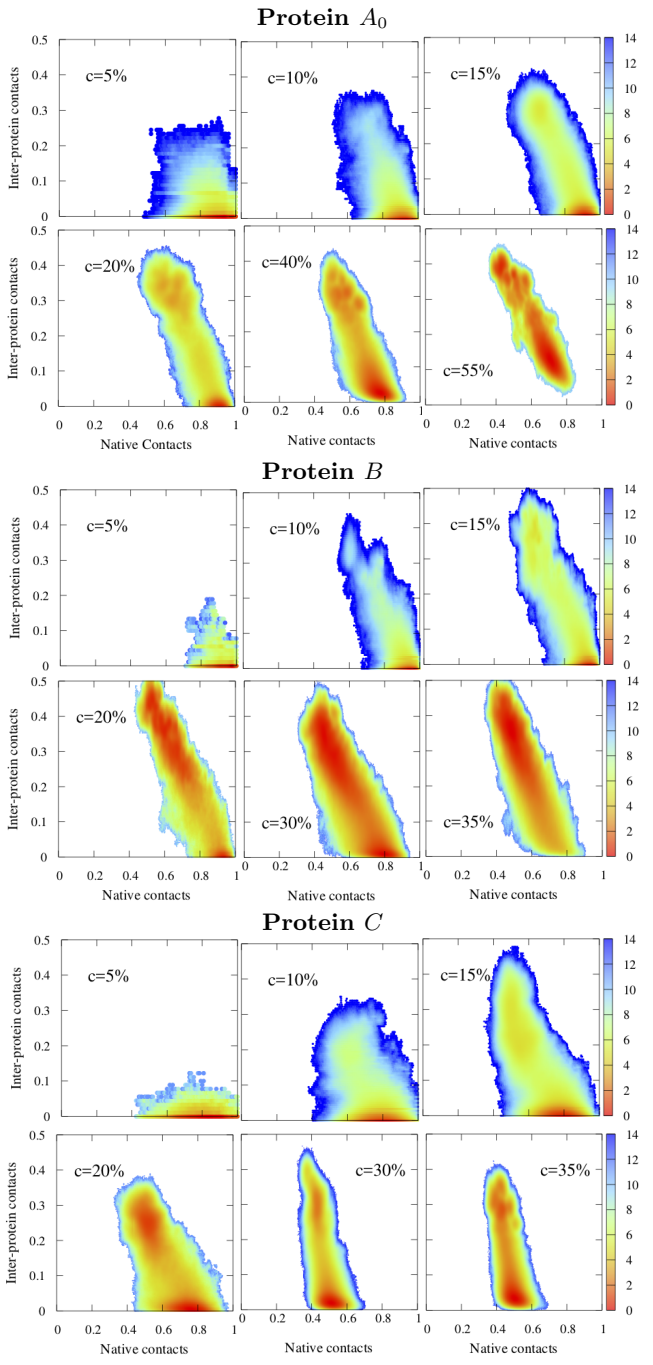

FIG. 3. Color map of the free energy profile $F(N_c, I_c)$ of the protein A_0, B and C, as function of the native contacts and inter-protein contacts, for different protein concentration c. We computed $F(N_c, I_c) = -k_B T \ln P(N_c, I_c)$, where N_c are the numbers of contacts in common with the native structure, and I_c are the inter-protein contacts. Native contacts N_c have been normalised by total number of native contacts (i.e. 1 is fully folded) and inter-protein contacts have been normalized by the total number of monomers ln, where n is the number of proteins simulated and l is the length (number of amino acids) of a single protein. In the shown cases, the size of the simulation box have been chosen such that $c = n$, i.e. a single protein occupies a volume corresponding to the 1% of the available volume.

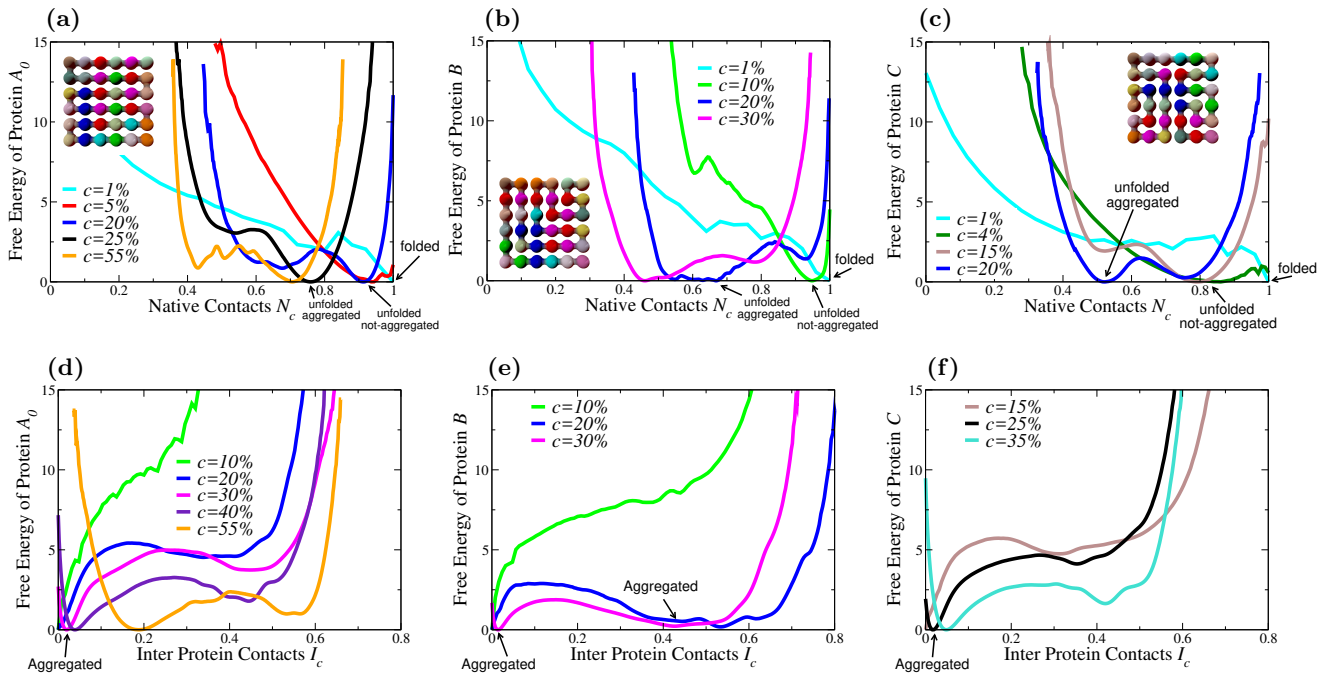

FIG. 4. Free energy profile of the protein A_0, B and C as function of N_c (upper panels) and I_c (lower panels) for different concentrations. All the free energy curves are in k_BT units and have been shifted such that the minimum coincides with 0. The N_c axes has been normalized dividing the number of native contacts for its maximum possible value (corresponding to all the proteins in their native conformation). The axes Ic has been normalized dividing the number of inter-protein contacts for the total number of amino acids. We find that, for the protein B (C), the $FOL \to UNF$ transition occurs at $c^{(B)}_{FOL\to UNF} \sim 8 \pm 1\%$ ($c^{(C)}_{FOL\to UNF} \sim 4 \pm 1\%$) and the $UNF \to AGG$ transition occurs at $c^{(B)}_{UNF\to AGG} \sim 16.5 \pm 1.5\%$ ($c^{(C)}_{UNF\to AGG} \sim 18 \pm 2\%$).

10.1002/cphc.201900904

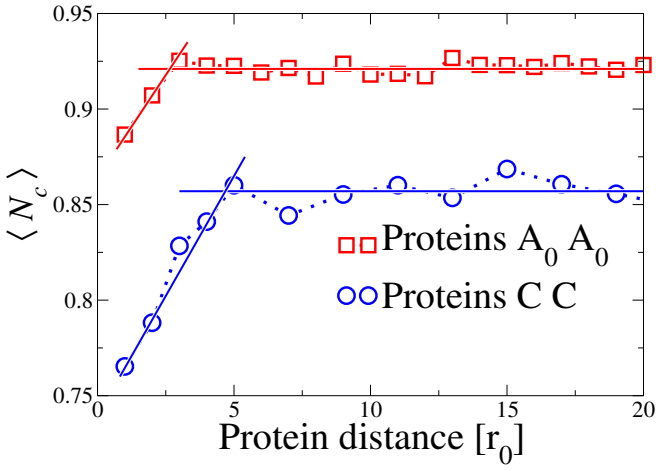

FIG. 5. Average number of native contacts $\langle N_c \rangle$ for the binary solutions with i) two proteins A_0 (red squares); ii) two proteins C (blue circles). Data are plotted as a function of the minimum distance between the two proteins d_{pp}. We considered the minimum inter-protein distance between each pair of residues. In this sense, any distance larger than one implies that the closest residues (i.e. the whole proteins) are not interacting directly but necessarily through the water. Lines are guides for the eye showing the increasing trend of $\langle N_c \rangle$ at smaller values of d_{pp}, and the constant value of $\langle N_c \rangle$ at larger d_{pp}. The intersection between the lines identifies the interaction radius of the proteins. The protein unfolds at a distance 2.5 for A and 5 for C both close to the average protein-proteins distances at the $FOL \rightarrow UNF$ transition concentrations

Supplementary Information

Protein Sequences

The sequences in FASTA encoding of the proteins are the following.

- Protein A_0: $NRMDCVACKWDNPKMECTICKWEQGKMEHLSYKFEF$;
- Protein A_1: $NDDGCSACFKKNQEEMCIVCWKKPREEMHLTYWRKQ$;
- Protein A_2: $PDYMDSIKWHKQTECMELVKWCKGNECMELARFCRN$;
- Protein B: $PRDCMTMHQKSNERCWCKEYIKECDKNGEKELIKFV$;
- Protein C: $PKLKCWEQMRMCKWDAMDRYSVHECFEIENTKFCIG$;
- Protein D: $LMKEREWVSMKDRYFDKGKCTPCEKCQHWNAMCEWI$;
- Protein E: $DMWALMCVFCEKEHWKDRYTQREPEINKENDCSGCK$;
- Protein F: $PYGKHQVCEECLICKMCAQCWFWEKKEGLKEEKMFNWEKRDISRTDRMN$.

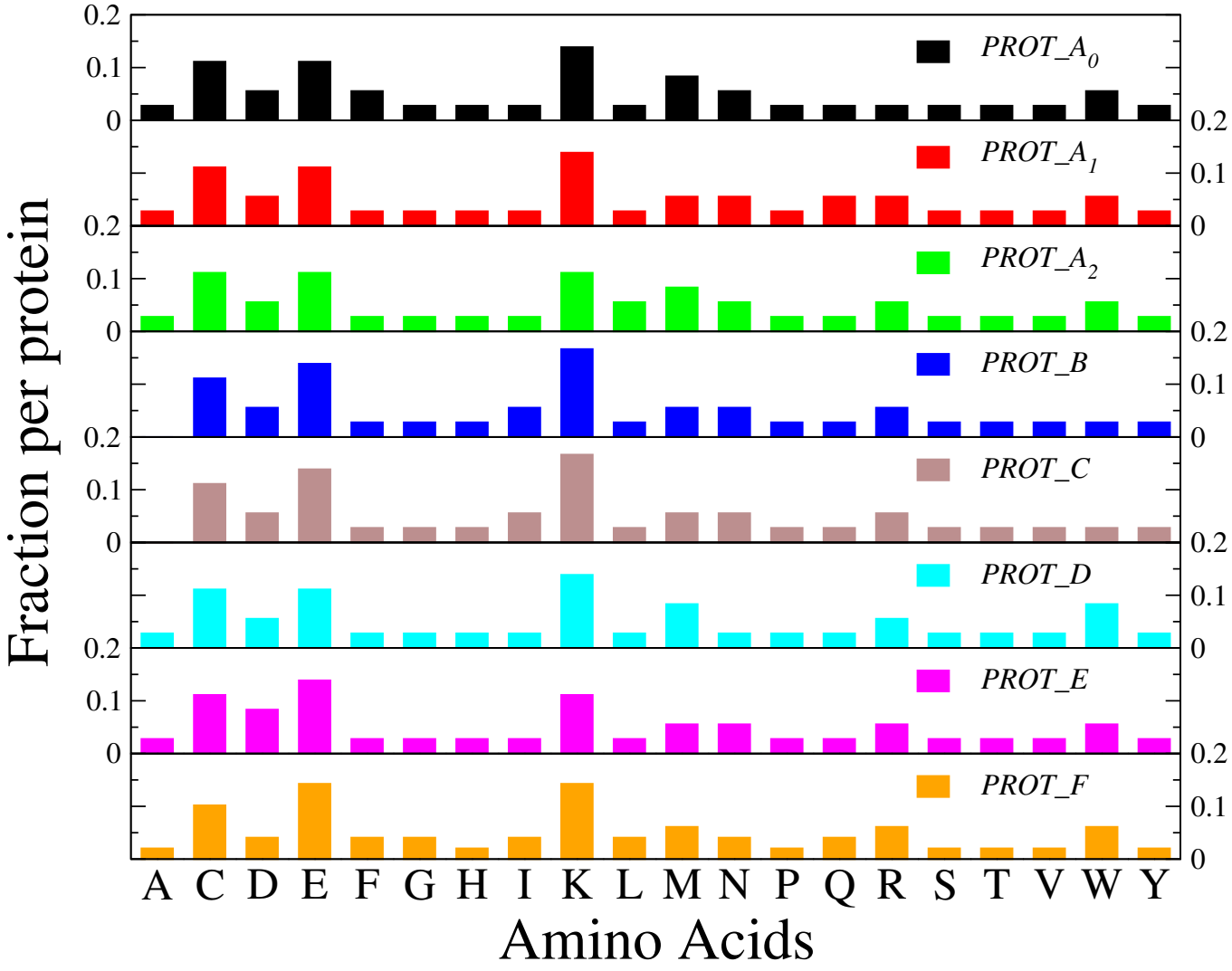

FIG. 6. Amino acid composition of the designed proteins.

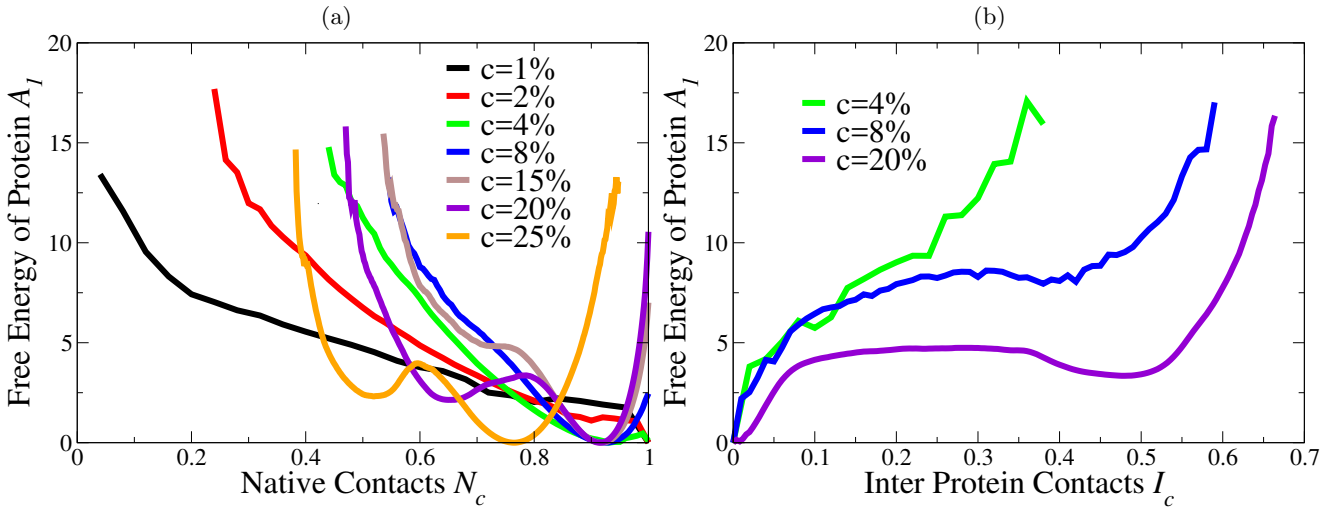

FIG. 7. a) Free energy profile of the protein A_1 as function of the native contacts. b) Free energy profile of the protein A_1 as function of the contacts between different molecules.

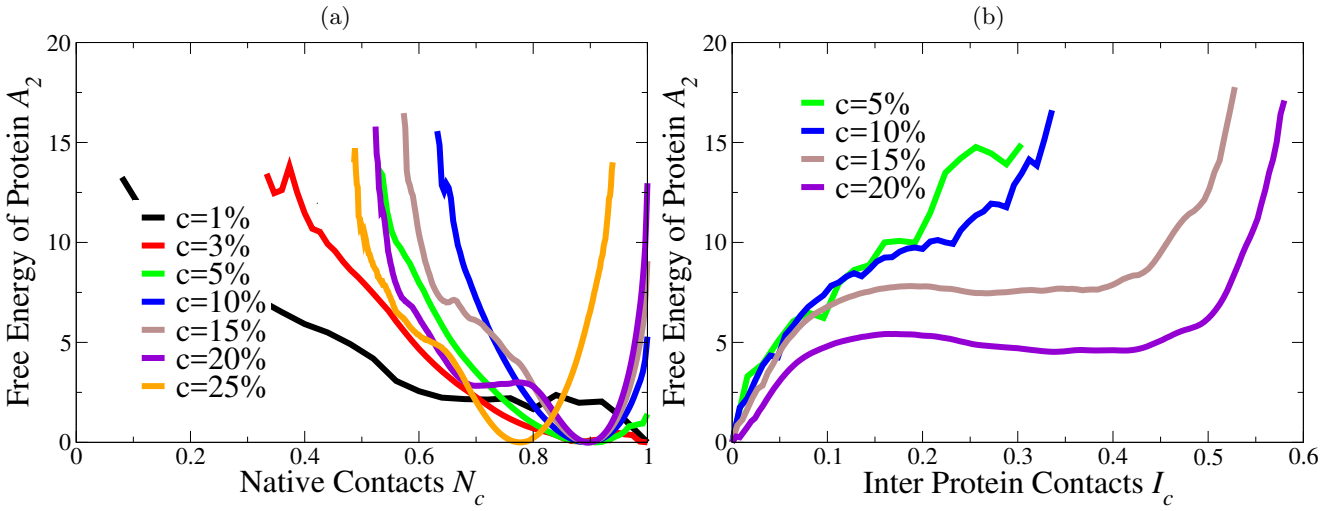

FIG. 8. a) Free energy profile of the protein A_2 as function of the native contacts. b) Free energy profile of the protein A_2 as function of the contacts between different molecules.

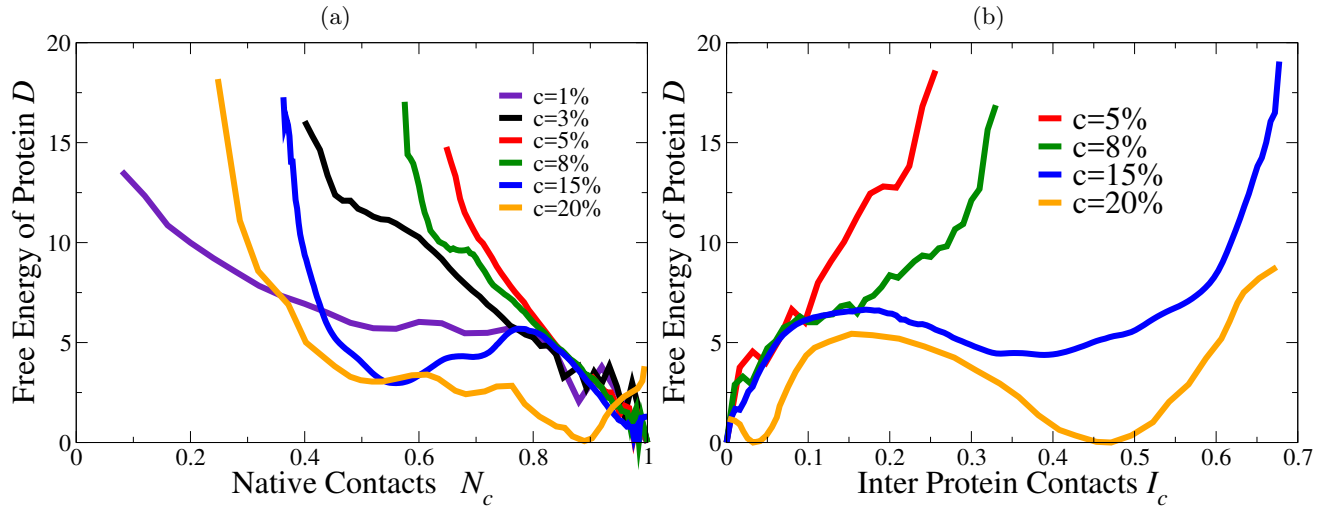

FIG. 9. a) Free energy profile of the protein D as function of the native contacts. b) Free energy profile of the protein D as function of the contacts between different molecules.

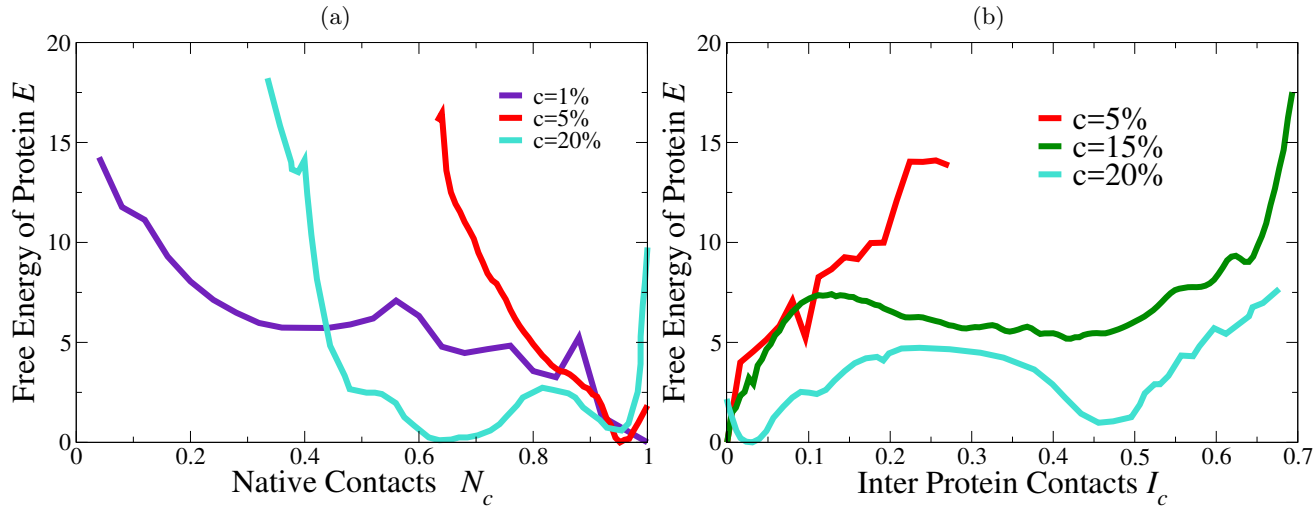

FIG. 10. a) Free energy profile of the protein E as function of the native contacts. b) Free energy profile of the protein E as function of the contacts between different molecules.

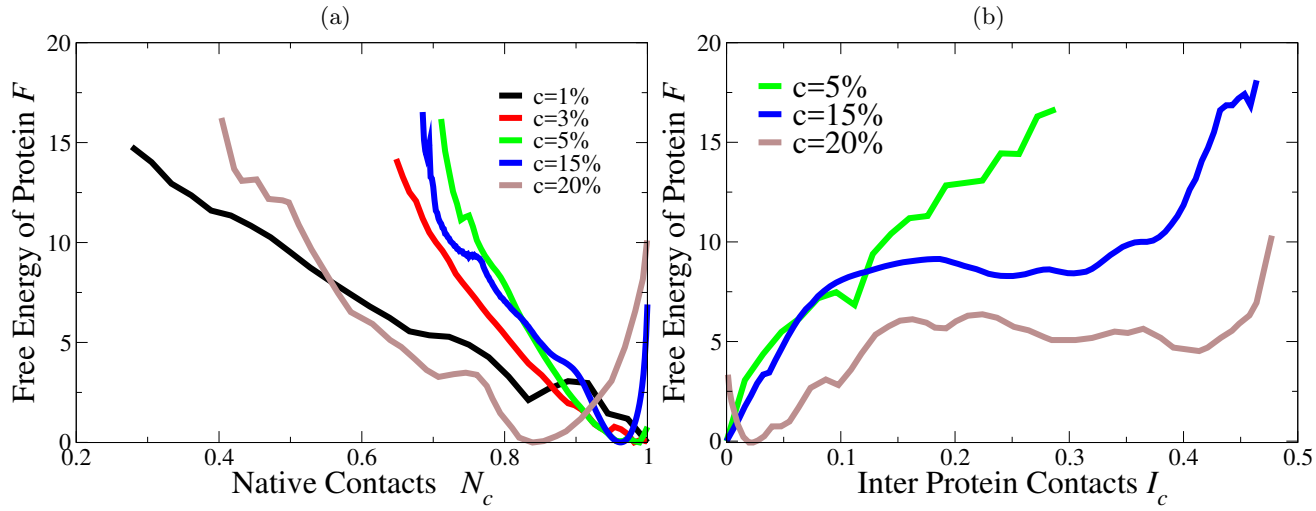

FIG. 11. a) Free energy profile of the protein F as function of the native contacts. b) Free energy profile of the protein F as function of the contacts between different molecules.

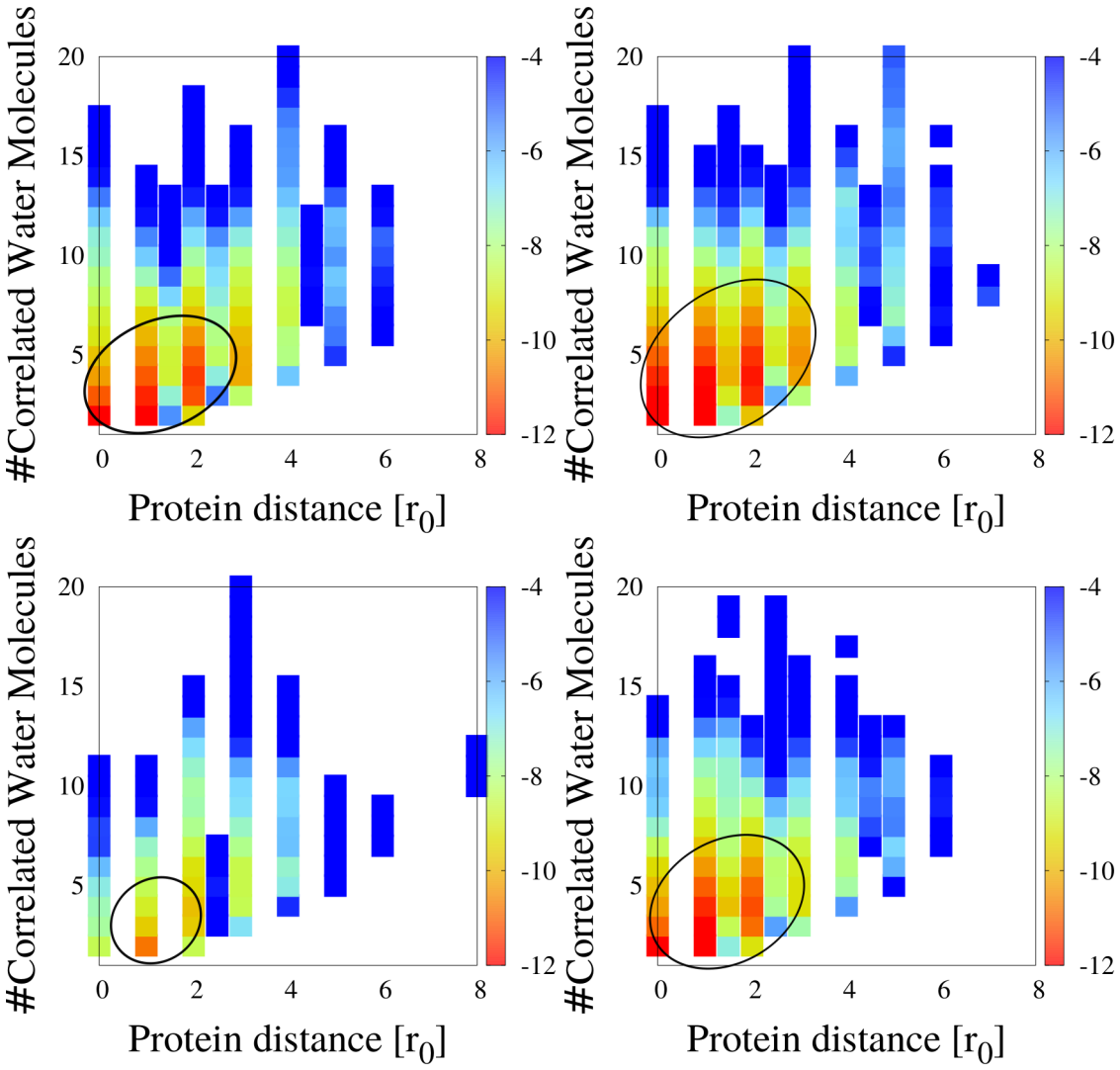

FIG. 12. Negative logarithm of the probability distribution of the clusters of statistically correlated water molecules in contact with two proteins, as function of the minimum protein distance and the number of water molecules belonging to the cluster. Following a percolation mapping shown in Ref. [81, 97, 98], two neighbour bonding variables σ_{ij} and σ_{ji}, such that $\sigma_{ij} = \sigma_{ji}$, belong to the same cluster with probability $p \equiv 1 - \exp(-\mathscr{J}/k_B T)$, where $\mathscr{J}$ is the specific interaction between σ_{ij} and σ_{ji}. In other words, a cluster represents a contiguous region of statistically correlated degrees of freedom of water, and its size is related to the statistical correlation length. On average, we assume that an entire water molecules belong to a cluster any four bonding indices (since any water molecules is described by four bonding indices). (a) Clusters' distribution between proteins A_0 folded. (b) Clusters between proteins A_0 unfolded. (c) Clusters between proteins C folded. (d) Clusters between proteins C unfolded. The proteins at distance 2.5 for A_0 and 5 for C have clusters and that is the distance at which they unfold. Lines are guides for the eyes showing the region with higher probability.

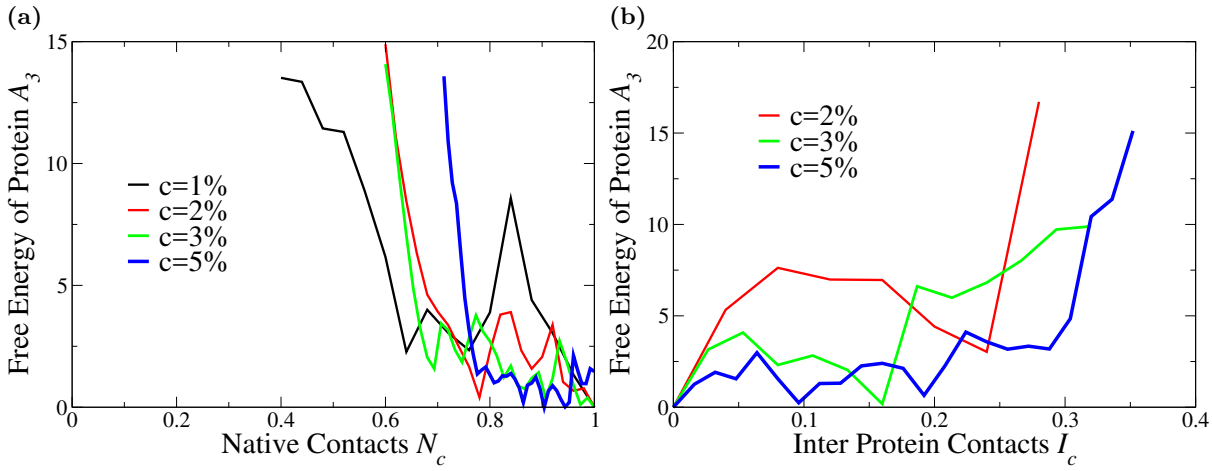

FIG. 13. Free energy profiles $F(N_c)$ (a) and $F(I_c)$ (b), function respectively of N_c and I_c for the proteins A_3. We designed the sequence of protein A_3 switching off all the water-water interaction terms in the hydration shell. Protein A_3 is not surprisingly less stable than the sequence designed with explicit water [34]. The data show the disappearance of the UNF state and the direct transition to the AGG state. Moreover, the $FOL \rightarrow AGG$ transition takes place at much lower concentrations with respect to the case where the hydration water is explicitly accounted for (in the present case as low as 2%). Hence, the hydration water acts as a barrier against the aggregation. We did not perform this analysis directly with A0 (or the other proteins designed in explicit solvent) because those sequences have been optimised in the full explicit water model and the aggregation would not have proven the role of the water.