Improving Coarse Grain Models of Protein Folding through Weighting of Polar-polar/hydrophobic-hydrophobic Interactions into Crowded Spaces

Hiram Isaac Beltrán
Universidad Autonoma Metropolitana Azcapotzalco
Salomón J. Alas-Guardado
Universidad Autónoma Metropolitana Unidad Cuajimalpa: Universidad Autonoma Metropolitana Cuajimalpa
Pedro Pablo Gonzalez Perez (pgonzalez@cua.uam.mx)
Universidad Autonoma Metropolitana Cuajimalpa https://orcid.org/0000-0001-7223-9035

Research Article

Keywords: HP model , Protein folding/structure , Polar contacts , Correlated networks , Convex function approach

Posted Date: June 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-566677/v1

Improving coarse grain models of protein folding through weighting of polar-polar/hydrophobic-hydrophobic interactions into crowded spaces

Hiram Isaac Beltrán,[1] Salomón J. Alas-Guardado,[2,*] and Pedro Pablo González-Pérez[3,*]

[1]*Departamento de Ciencias Básicas, Universidad Autónoma Metropolitana, Unidad Azcapotzalco, CDMX 02200, México.*
[2] *Departamento de Ciencias Naturales, Universidad Autónoma Metropolitana Unidad Cuajimalpa, CDMX 05300, México.*
[3]*Departamento de Matemáticas Aplicadas y Sistemas, Universidad Autónoma Metropolitana, Unidad Cuajimalpa, CDMX 05300, México.*

Hiram Isaac Beltrán
hibc@azc.uam.mx
orcid.org/0000-0002-1097-455X

[*]Salomón J. Alas-Guardado
salas@cua.uam.mx
orcid.org/0000-0001-8903-8766

[*]Pedro Pablo González-Pérez
pgonzalez@cua.uam.mx
orcid.org/0000-0001-7223-9035

[*]Corresponding author

ABSTRACT

In this piece of work were tested 7 Hydrophobic-Polar sequences in two types of 2D-square space lattices, homogeneous and correlated, the latter simulating molecular crowding included as a geometric boundary restriction. The optimization of the 2D structures was carried out using a variant of Dill's model, inspired by the convex function, which takes into account both the hydrophobic (Dill's model) and polar interactions, aimed to include more structural information to reach better folding solutions. While using correlated networks, the degrees of freedom in the folding of sequences were limited, and as a result in all cases more successful structural trials were found in comparison to the homogeneous lattice. In particular, the S_5 sequence turned out to be the most difficult sequence of the seven folded, this perhaps due to the intrinsic i) degrees of freedom and ii) motifs of the expected 2D HP structure. Regarding S_2 and S_6 sequences, although optimal folding was not achieved for neither of the two approaches, folding with correlated network approach not only produced better results than homogeneous space, but for both sequences the best values found with crowding were very close to the expected optimal fitness. The sequences S_1-S_4 and S_6 were better folded with medium lattice units for the correlated media, instead, S_5 and S_7 were better folded with a bit larger degree of

lattice unit, revealing that depending on the degrees of freedom and particular folding motifs in each sequence would require particular crowding to achieve better folding. Finally, we claim that in all folded sequences in crowded spaces achieve better results than homogeneous ones.

Keywords: HP model · Protein folding/structure · Polar contacts · Correlated networks · Convex function approach

Declarations

Funding

This research was supported by CONACyT (project 0222872 HIB and project A1-S-46202 SJAG) and Universidad Autónoma Metropolitana.

Conflicts of interest/Competing interests

The authors declare there are no competing financial interests. The authors declare they have no financial interests.

Availability of data and material

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Code availability

The bioinformatics framework that supported this study is available in:
http://bioinformatics.cua.uam.mx/site/

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Salomon J. Alas-Guardado, Pablo González-Pérez and Hiram Isaac Beltrán. Salomon J. Alas-Guardado participated in the experiment design and drafted the manuscript; Pedro Pablo González-Pérez coordinated the software engineering work of the Evolution bioinformatics platform, participated in the experiment design, conducted the in silico experiments, and drafted the manuscript. Hiram Isaac Beltrán conceived the design of the target 2D/3D structures and foldamers, participated in the experiment design and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Authors would like to thank the support provided by Oscar Sánchez Cortés, in the improvements of the *Evolution* bioinformatics platform.

Introduction

Protein folding problem [1-3] is one of the major tasks in biological sciences. In order to try to solve it one could get important clues or tracks by analyzing the (dis)functional/unfolded 3D structures of those macromolecules and also by sightseeing the properties of the individual aminoacids that conform them. But other clues could be acquired by analyzing the real environment that these biomolecules face off in living organisms, which indeed is very different from those present at lab scale, commonly set up for their study. In order to develop a more efficient protein folding strategy, the next issues must be taken into account as key roles in protein folding stability: i) polar-polar (P···P) interactions and ii) molecular crowding.

Physicochemical differences among amino acids, are mainly due to the nature of their side chains, developing important changes in conformation, dimensions and polarity. These three latter properties directly modulate the packing of those aminoacids included in proteins. Normally in this field, hydrophobicity is a major driving force in protein folding [4]. Nevertheless, itself hydrophobicity, or its counterpart, polarity, are both key role determinants of nature and strength of molecular interactions between aminoacids themselves molding (inner) the protein structure, as well as among aminoacids and media, again molding (outer) the protein structure. In this line, hydrophobic aminoacids (H) strongly interact between them by hydrophobic-hydrophobic (H···H) contacts to yield the called hydrophobic core. Besides, polar aminoacids (P) also interact strongly between them by P···P contacts, but they develop mainly electrostatic interactions [5]. In aqueous media those polar aminoacids tend to accommodate towards this polar media attracting water molecules and forming hydration cores surrounding polar protein surface. This simplification conducted to one of the simplest but interesting protein coding, the HP model [2,3,6].

In living organisms, the ideal diluted and well mixed lab conditions are not present, which indeed are wrong thoughtful [7]. Instead of that, there exist systems highly concentrated, not well mixed and highly tortuous, this could be a real definition of what should be thought about molecular crowding [8,9]. Therein exist water molecules, ions, metabolites, proteins, nucleic acids, lipids, carbohydrates, etc. all of them ranging concentrations as high as 400 $g \cdot L^{-1}$ [10-12]. In these molecular situations crowding and confinement take a crucial role in structure, folding stability, ligand binding/recognition, macromolecular interactions and assembly, as well as function of those biomolecules [13]. In particular, the crowding effect is noticeably in protein fragments possessing intrinsic disorder or moieties with high conformational variation due to function, e.g. mobility in ligand clefts, chaperones and transporter/motor proteins such as kinesins, and due to their own existence, this applies to the majority of proteins [9,14]. Moreover, protein-crowded systems are directly responsible of further selectivity and even specificity due to the enhancement of deeper thermodynamic holes in the potential energy surface [15]. Diverse studies have demonstrated that crowding effects are variable ranging from modest to drastic. Protein folding stability and binding stability belong to modest. Drastic effects rely in other biological functions, e.g. replication and transcription efficiencies, which could lead to diseases and diminished efficacy of drugs and related pharmaceuticals. No matter of the magnitude of such effects, the main clue herein is that the crowder agents have crucial consequences in many biological systems, all of them due to protein structure compactness as a major indicative [16]. Hence, crowding is able to propitiate that molecules present in living organisms would behave in very different ways depending on that type of crowding (different pHs, ions, molecules, concentrations, etc.), and also very different from that observed in lab scale conditions, orders of magnitude away from those conditions present in the interior of cells [7,11,12]. Therefore, many of the biochemical processes should be studied *in silico*/*in vitro* almost reaching those inner cell conditions and obtained results could bring aid in order to fully understand different shaping and function of those biomolecules. So no matter of which level of "coding" is the protein itself, it should be influenced due to crowding present in living organisms and thus should be taken into account being part of a complete folding model [8].

Hence, in this piece of work, we are exploring simple protein folding employing a HP coarse grained 2D lattice model, taking into account 1) both H···H and P···P interactions and 2) a molecular crowding mimetic with geometric restrictions within a correlated network space. Condition 1) is reached from a variation of Dill's HP model [2,3,6] inspired by the convex function, where now both H···H and P···P interactions are both considered in the optimization process of the 2D HP structures. On the other hand, condition 2) is based on the use of correlated networks [17-19], with different correlation lengths, in order to simulate the molecular crowding characteristic of the intracellular

medium. In particular, this study aims to investigate whether the double restriction imposed by 1) and 2) conditions generate more successful folded 2D HP structures.

Material and methods

Hydrophobic-Polar protein folding model based on the convex function

As can be seen below, expression (1) describes the HP protein folding model proposed by Dill [2,3,6], which has been one of the most widespread coarse-grained models to date for the study and exploration of HP sequence folding [6,20-26]. When the HP model is used, the amino acid sequence is expressed in the [H, P] alphabet. Using 2D/3D lattices for the movement and positioning of the H/P beads, the resulting 2D/3D HP structures emerge optimizing the number of H···H contacts, based on the thermodynamic principle that establishes the formation of a hydrophobic core embedded within the motifs formed by P amino acids, which are exposed towards the outside or interacting among them as P···P contacts. The optimization process of 2D/3D HP structures is commonly carried out through evolutionary computation techniques, such as genetic algorithms [17-19,27].

Based on expression (1), in this work we continue using a variant inspired by the convex function [28,29], in such a way that important substructures and motifs based on P···P interactions, e.g. hydrogen bonds and salt bridges, are taken into account during the structural optimization process. In other words, the new variant of the HP model – see expressions (2) to (4) – considers the contributions of the P···P interactions to achieve better folding stability in 2D/3D HP lattices.

Making a particular emphasis, the convex function has played an important role in the study of optimization problems [28-30] and as could be seen in expression (4), its key role in the proposed model is to tune the importance or weight to attribute to both H···H and P···P contacts simultaneously. In particular, each H···H interaction is weighted by the value α-1 while each P···P interaction is weighted by the value -α, being $0 \leq \alpha \leq 1$.

$$F = f(e) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} e_{ij} \Delta_{ij} \qquad (1)$$

where the weight attributed to the H···H, H···P, and P···P interactions is given by:

$$e_{HH} = -1.0;\ e_{HP} = 0.0;\ e_{PP} = 0.0$$

while Δ_{ij} is given by:

$$\Delta_{ij} = \begin{cases} 1, \text{if } i \text{ and } j \text{ are topological, but not sequence neighbors} \\ 0, \text{otherwise} \end{cases}$$

A real function of real variable $f(x)$: $\mathbb{R}^n \rightarrow \mathbb{R}$ is convex if and only if:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)\ \forall x, y \in \mathbb{R}^n, \forall \alpha \in [0,1] \qquad (2)$$

The x and y values in (2) could be considered as the total number of topological contacts P···P and H···H, respectively. Therefore, $F = f$ (e) in (1) is herein written as follows:

$$F = f(e) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} e_{HH} \Delta H_{ij} + \sum_{i=1}^{n} \sum_{j=i+1}^{n} e_{PP} \Delta P_{ij} \quad (3)$$

where the weight attributed to the H···H and P···P interactions is given by:

$$e_{HH} = \alpha - 1;\ e_{PP} = -\alpha;\ 0 \leq \alpha \leq 1$$

while ΔH_{ij} and ΔP_{ij} represent H···H and P···P interactions, respectively. That is:

$$\Delta H_{ij}, \Delta P_{ij} = \begin{cases} 1, \text{if } i \text{ and } j \text{ are topological, but not sequence neighbors} \\ 0, \text{otherwise} \end{cases}$$

From the two latter expressions, in a similar way as it is the right side of the convex function, the resulting model was finally obtained as expressed in (4):

$$F = f(e) = (\alpha - 1) \sum_{i=1}^{n} \sum_{j=i+1}^{n} \Delta H_{ij} + (-\alpha) \sum_{i=1}^{n} \sum_{j=i+1}^{n} \Delta P_{ij} \quad (4)$$

$$\Delta H_{ij}, \Delta P_{ij} = \begin{cases} 1, \text{if } i \text{ and } j \text{ are topological, but not sequence neighbors} \\ 0, \text{otherwise} \end{cases}$$

Therefore, in this work we use the model stated in (4) for testing α values that conduct to the formation of structures with better local minima, preserving hydrophobic core and now developing also PP substructures. The optimization process of 2D/3D HP structures is carried out by a genetic algorithm, whose fitness function is given precisely by expression (4). Further details can be referred in [17-19].

Correlated networks simulating the molecular crowding

In previous researches we used the Dual Site Bond Model (DSBM) to build correlated spaces, in order to mimic inhomogeneous media to fold 2D HP structures. Briefly, the correlated networks were achieved from Monte Carlo simulations and have been useful to simulate porous media, which show heterogeneity, tortuosity and fractal geometry [31-33]. The main idea herein is that these media could be used to perform and affect the folding of HP sequences, which in fact, this has been successfully observed [17-19]. It has been reported that the correlated networks are generated from two probability density functions, one depicts sites or nodes and the other denotes bonds. When both elements are overlapped (Ω) using a certain quantity of them, it is possible to obtain the correlation length (ξ), which indicates the degree of heterogeneity of the lattice. According to the analysis it has been observed that if $\Omega < 0.4$, i.e., sites and bonds are overlapped 40%, then $\xi = 0$, and the achieved lattices are totally heterogeneous, contrastingly if $\Omega > 0.4$ then $\xi > 0$. As $\Omega \to 1$, i.e., sites and bonds tend to overlap 100%, and $\xi \to \infty$, besides homogeneous patches are formed on the lattice. If $\xi = \infty$, then one homogeneous lattice is attained [33].

Next, to build the correlated networks with different correlation lengths, in this study different values of $\xi = 0.94$, 1.80, 3.34, 5.58, 11.70, 18.12, 28.24, and 37.98 lattice units were employed, and classical percolation spanning cluster were made up by this means, for which Hoshen-Kopelman algorithm was used [34]. The fractal dimension (d_f) was measured to each percolation cluster using the box-counting method [35]. One cluster grain of these structures at certain ξ was chosen to fold each one of the HP sequences proposed in this work. Details about the construction of the correlated networks and the corresponding percolation clusters could be further revised in references [17,19].

Methodological approach for *in silico* folding

In silico protein folding experiments were carried out on the Evolution bioinformatics platform (http://bioinformatics.cua.uam.mx/site/) [17,19], a protein folding simulation tool based on HP 2D/3D lattice models and evolutionary algorithms. As already mentioned, our goal is to explore simple protein folding employing a HP coarse grained 2D lattice model, considering 1) both H···H and P···P interactions and 2) a molecular crowding mimetic with geometric restrictions within a correlated network space; all the latter conditions are supported by the Evolution platform in its latest version. The *in silico* folding methodology carried out is shown in Fig. 1.

Fig. 1 Methodological approach for *in silico* protein folding experiments. Note that conditions 1) and 2) previously mentioned are reflected in the right branch of the workflow. T_{HH}=Total of expected H···H contacts and T_{PP}=Total of expected P···P contacts

Results and discussion

In Table 1 are gathered the most important experimental design variables developed in this work. Therein is shown the obtained results for the designed $S_{1\text{-}7}$ sequences, all of them treated with folding approaches 1 (*FA1*) and 2 (*FA2*). According to methodology, in both folding approaches the P···P contacts were rewarded, but in *FA2* molecular crowding [17,18] has also been included as a geometric boundary restriction. For each sequence, 10 batches of experiments were run per approach (*FA1* or *FA2*), each batch consisting of 10 trials. All the experiments were carried out considering an initial population made up of 600 2D HP structures randomly generated from the target sequence, maintaining the population size through the 400 runs of the optimization algorithm. All sequences were thoroughly studied by *FA1* and *FA2* approaches, nevertheless, just S_2, S_3, and S_7 are going to be discussed in more detail herein as a sampling or evidences of general results.

Table 1 Designed target HP sequences and their best score achieved when folding approaches 1 (*FA1*) and 2 (*FA2*) are used. That is, in both folding approaches PP contacts are also rewarded, but only in folding approach 2 the molecular crowding is taken as other restriction

S_{ID}	DTS	SL	T_{HH}	T_{PP}	α_B	OF	*FA1*	*FA2*	
							BFF	ξ_B	BFF
S_1	$H_5P_4H_5P_4H_5P_4H_5P_4H_5$	41	20	8	0.2	-17.6	-17.2	18.12	-17.2
S_2	$P_2H_5P_4H_5P_4H_5P_4H_5P_4H_5P_2$	45	24	12	0.1	-19.2	-18.3	18.12	-19.0
S_3	$H_9P_2H_2P_4H_2P_2H_2P_4H_2P_2H_2P_4H_2P_2H_2P_4H_2$	49	24	8	0.05	-23.2	-23.2	18.12	-23.2*
S_4	$H_6P_4H_6P_4H_6P_4H_6P_4H_6P_4H_6$	56	30	11	0.1	-28.1	-26.9	18.12	-27.9
S_5	$H_7P_5H_2P_2H_2P_2H_2P_5H_{10}P_5H_2P_2H_2P_2H_2P_5H_7$	64	33	8	0.1	-30.5	-27.9	28.43	-28.8
S_6	$H_2P_6H_2P_4H_4P_4H_4P_{12}H_4P_4H_4P_4H_2P_6H_2$	64	22	21	0.05	-21.95	-21.8	18.12	-21.9
S_7	$HP_2HP_6H_4P_6HP_2H_2P_2HP_6H_4P_6HP_2H$	48	17	18	0.1	-17.1	-16.0	28.43	-17.1

S_{ID}=Sequence identification; DTS= Designed Target Sequence; SL= Sequence Length; T_{HH}=Total of expected H···H contacts; T_{PP}=Total of expected P···P contacts; α_B=Best α value; OF= Optimal Fitness; BFF=Best Fitness Found; ξ_B = Best correlation length. S_1 and S_3: Sequences previously studied in [19].

*Note that although the expected optimum was achieved with both approaches, the folding in the correlated space produced a structure much closer to the expected 2D structure.

Analysis of the folding of the S_2, S_3 and S_7 sequences with *FA1* and *FA2* approaches

Figure 2 shows the behavior of the S_2 sequence, when the folding was carried out using the model given by the expression (4) in homogeneous (*FA1*) and correlated (*FA2*) spaces. As can be seen in Fig. 2a, the optimal folding of the S_2 sequence is a hydrophobic nucleus H_{25} (5×5 hydrophobic square) with short P_4 polar branches placed towards the outside of the hydrophobic structure. Fig. 2b shows the tendency graph of the folding carried out through 10 batches, each one made up of 10 experiments. The solid black and red circles represent the best folding achieved with approaches *FA1* and *FA2*, respectively. As can be seen in this graph, in 8 of the 10 batches of experiments *FA2* produced much better results than *FA1*, with batches 1 and 7 exhibiting the best folding, both with BFF = -19.0, with respect to optimal fitness OF = -19.2. Contrasting with the best folding produced by *FA1*, which was obtained in batch 3 with a BFF = -18.3 that nonetheless is farther away from the OF. The best 2D HP structures folded with *FA1* and *FA2* are illustrated in Figs. 2c (BFF = -18.3) and 2d (BFF = -19.0), respectively. Note in Fig. 2d the restrictions imposed by the correlated medium, employing the best correlated length (ξ_B) of 18.12 for S_2, limiting the degrees of freedom in the folding of the HP sequence, and in contrast to the folding in the homogeneous medium in an unrestricted lattice, equivalent to $\xi_B = \infty$, and where all the weight of the folding relies on the optimization through the evolutionary algorithm. As indicated in Table 1, for *FA1* and *FA2* the best folding was achieved with $\alpha = 0.1$ (see expression (4)), which in the case of *FA2* allowed obtaining both the hydrophobic core and almost reaching the completeness of the substructures polar expected, 10 of 12 P···P contacts. Values of $\alpha > 0.1$ did not allow to preserve the hydrophobic core in either of the two approaches, instead, a fast formation of unexpected PP substructures was achieved. Herein is worth to notice that lower and higher values of $\xi_B = 18.12$ did not achieved better results.

Fig. 2 S_2 HP 2D structures: (a) expected; (b) folding tendency graph; (c) *FA1* with $\alpha = 0.1$; (d) *FA2*, with $\alpha = 0.1$ and $\xi_B = 18.12$. In this sequence, mainly due to terminal P_2 branches, the hydrophobic core is restricted to form different degenerated structures, evidencing that the inclusion of those extra P···P contacts minimized degeneracy possibilities of S_2, due to the formation of the two lateral PP substructures. Nevertheless, of the clear gain observed between 1c and 1d, the final score achieved does not reach the optimum but just by a small amount due to the bending of one of the two terminal P_2 moieties. In (a), (c), and (d) the H and P residues are labeled in blue and red colors, respectively. The cluster grain is marked in black color circles in (d). For clarity in (d), the empty space of the correlated network is not shown here. In (b) dashed lines are only guides for the eye

The results of the folding of the S_3 sequence, using the *FA1* and *FA2* approaches, are shown in Fig. 3. As can be seen in Fig. 3a, the expected 2D HP structure for the S_3 sequence is a hydrophobic nucleus H_{25} (5×5 hydrophobic square) surrounded by a polar perimeter formed by 24 beads. The folding tendency graph for S_3 sequence is shown in Fig. 3b, where it could be seen that in 8 of the 10 batches of experiments *FA2* produced much better results than *FA1*, even reaching OF = -23.2 three times in batches 2, 8, and 9; while *FA1* produced the optimal fitness only once, occurring in batch 1. The optimal 2D HP structures folded with *FA1* and *FA2* are illustrated in Figs. 3c and 3d, both characterized by the OF = -23.2. Herein is worth to notice that although both structures reached the optimal folding (in terms of the optimal number of contacts H···H and P···P), only the 2D HP structure folded in the correlated space (Fig. 3d) matches with the expected 2D HP structure which is a symmetric 7×7 square (see Fig. 3a). Again, in this case, the restrictions imposed by a more crowded medium, employing the best correlated length ξ_B of 18.12, limited the degrees of freedom in the folding of the HP sequence, in contrast to the optimal folding achieved in the homogeneous medium (see Fig. 3c) which produced a degenerate optimal structure but with different shaping, which is more spread or elongated in one dimension due to this crowding absence. As shown in Table 1, for *FA1* and *FA2* the best folding was reached with $\alpha = 0.05$, which in the case of *FA2* allowed obtaining the expected 2D HP structured. On the other hand, values of $\alpha > 0.05$ did not allow to preserve the hydrophobic core in either of the two approaches. Regarding the correlation length, lower and higher values of $\xi_B = 18.12$ did not reached better results.

Fig. 3 S_3 HP 2D structures: (a) expected; (b) folding tendency graph; (c) *FA1* with $\alpha = 0.05$; (d) grain selection for folding; (d) *FA2*, with $\alpha = 0.05$ and $\xi_B = 18.12$. In this sequence the hydrophobic core is built by 5×5 square and both terminal H moieties rely or are embedded into this core. Nevertheless, of the expected degeneracy into this hydrophobic core, see a), c) and d), the main complexity in this sequence regards precisely to the accommodation of the P boundaries, where *FA1* yielded a more spread structure in c) but *FA2* generated the expected symmetrical 7×7 square in a more precise matching with expected S_3 mainly due to crowding imposed by the media. The colors labeled of the H and P residues and cluster grain are the same as Fig. 2. In (b) dashed lines are only guides for the eye

Lastly, folding results of S_7 sequence, using *FA1* and *FA2* approaches, are presented in Fig. 4. Fig. 4a shows the expected 2D HP structure for S_7 sequence, which corresponds to a hydrophobic 4×4 core yielding H_{16} concentric to a complex polar substructure formed by 32 amino acids. From which conjunction emerges a noteworthy rectangular symmetry that represents a 6×8 HP shape. The tendency graph of the folding carried out through 10 batches of experiments is shown in Fig. 4b. Note that in 8 of the 10 batches of experiments *FA2* get much better results than *FA1*, with batch 5 exhibiting the OF = -17.1 and values very close to OF in batches 4, 7 and 8, all with a BFF = -17.0. On the other hand, and as could be seen in Fig. 4b, the correct folding of S_7 sequence was not attained in the homogeneous medium, due to the degrees of freedom for both the polar and hydrophobic regions of the sequence itself. Note that in all batches of experiments, the BFF values produced by *FA1* fail to improve the worst BFF values produced by *FA2*. Figs. 4c and 4d show the best 2D HP structures folded with *FA1* (BFF = -16.0) and *FA2* (OF = -17.1), respectively. As could be seen in Fig. 4d, the folding with *FA2* was able to achieve the expected 2D HP structure (see Fig. 4a), when the best correlated length ξ_B of 28.43 was used. As specified in Table 1, for *FA1* and *FA2* the best folding for S_7 sequence was achieved with $\alpha = 0.1$. With values of $\alpha > 1$, a worsening in folding was noted with *FA1*, whereas with *FA2* it led to a breakdown of the hydrophobic core. Regarding correlation length, lower and higher values of $\xi_B = 28.43$ did not conducted to better results, being a precise crowding the required to correctly fold S_7.

Fig. 4 S_7 HP 2D structures: (a) expected; (b) folding tendency graph; (c) *FA1* with $\alpha = 0.1$; (d) *FA2*, with $\alpha = 0.1$ and $\xi_B = 28.43$. In this sequence, again, mainly due to P substructure loops, the hydrophobic core is one more time restricted to form different degenerated structures. Due to this particular sequence design, even the *FA1* provided almost the desired structure (b) reaching a $B_{FF} = -16.0$, correct folding was not achieved herein, due to the degrees of freedom for both the polar and hydrophobic regions of the sequence itself. Note that the desired structure is achieved (d) with the correlated space approach *FA2*, giving the requested $B_{FF} = -17.1$. Note that the polar regions develop twisted loops in all corners, and both terminal H moieties rely or embed into the hydrophobic core in a very close match as in S_3 sequence. The colors labeled of the H and P residues and cluster grain are the same as Fig. 2. In (b) dashed lines are only guides for the eye

Summarized results

The results of the folding of S_1 to S_7 sequences, using *FA1* and *FA2* approaches, are summarized in Fig. 5. As mentioned earlier, for each sequence, 10 batches of experiments were run per approach, each batch consisting of 10 trials. As can be seen in this figure, the S_5 sequence turned out to be the most difficult sequence of the seven folded sequences, this perhaps due to the intrinsic i) degrees of freedom and ii) motifs of the expected 2D HP structure. Note in Table 1 that the OF expected for this sequence was -30.5, however, the BFF in the homogeneous and correlated media were -27.9 and -28.8, respectively. Notwithstanding in these results again *FA2* folding produced better results than *FA1*. Another sequence whose folding turned out to be quite difficult was S_4 (see Fig. 5 and Table 1), it's OF was -28.1, however, the BFF by *FA1* and *FA2* approaches were -26.9 and -27.9, respectively, being the one employing correlated space a better folding approach. It seems that the more difficult folding of these S_4 and S_5 sequences, see Fig. 5, is due to the fact that they are stickier (enhanced amount of P···P and H···H interactions) and have deeper minima. Regarding S_2 and S_6 sequences, although optimal folding (OF = 19.2 for S_2 and OF = -21.95 for S_6) was not achieved for neither of

the two approaches, folding with *FA2* not only produced better results than *FA1*, but for both sequences the best values found with *FA2* were very close to the expected OF (BFF = -19.0 for S_2 and BFF = -21.9 for S_6).

Fig. 5 Folding tendency for sequences S_1 to S_7 in homogeneous and correlated media. OF = Optimal Fitness, BFF_HS = Best Fitness found in Homogeneous Space, BFF_CS = Best Fitness Found in Correlated Space. Dashed lines are only guides for the eye

Nevertheless, that for folding means in *FA2*, herein were tested different values of ξ = 5.58, 11.70, 18.12, 28.24, and 37.98. Precisely here is worth to mention that the 5.58 and 11.70 lattice units were too tortuous for this folding means and the 37.98 resulted very similar to a homogeneous space like *FA1*, due to the fact that the cluster grain size was very homogeneous. In S_1-S_4 and S_6 were better folded with ξ of 18.12, instead, S_5 and S_7 were better folded with ξ of 28.43, revealing that depending on the degrees of freedom and particular folding motifs in each sequence would require particular crowding to achieve better folding

Conclusions

Here were tested 7 HP-sequences in two types of 2D-square space lattices, homogeneous lattice and correlated lattices, the latter simulating molecular crowding. The optimization of the 2D-HP structures was carried out using a variant of Dill's HP model, inspired by the convex function, which takes into account both the H···H (Dill's model) and P···P contacts, to gather more structural information to reach better folding solutions in any given HP-sequence. When this model was used, H···H interactions were tuned as α-1 and P···P interactions as $-\alpha$, with $0 \leq \alpha \leq 1$. The molecular crowding simulated by the correlated lattices was included as a geometric boundary restriction, which allowed significantly limiting the degrees of freedom in the H-P folding of sequences and, as a result, in all cases more successful structural trials were found than those observed in the homogeneous lattice. Selected values of the correlation length played an important role to fold HP sequences, since for some of these values it was possible to achieve better folding.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181: 223-230.
2. Dill KA, Ozkan SB, Shell MS, et al. (2008) The protein folding problem. *Annual Review of Biophysics* 37: 289-316. http://dx.doi.org/10.1146/annurev.biophys.37.092707.153558.
3. Dill KA, MacCallum JL (2012) The protein-folding problem, 50 years on. *Science* 338: 1042-1046. http://dx.doi.org/10.1126/science.1219021.
4. Bolen DW, Baskakov IV (2001) The osmophobic effect: Natural selection of a thermodynamic force in protein folding. *Journal of Molecular Biology* 310: 955-963. http://dx.doi.org/10.1006/jmbi.2001.4819.
5. Leonhard K, Prausnitz JM, Radke CJ (2003) Solvent–amino acid interaction energies in 3-d-lattice mc simulations of model proteins. Aggregation thermodynamics and kinetics. *Physical Chemistry Chemical Physics* 5: 5291-5299. https://dx.doi.org/10.1039/b305414d.
6. Dill KA (1985) Theory for the folding and stability of globular proteins. *Biochemistry* 24: 1501-1509. http://dx.doi.org/10.1021/bi00327a032.
7. Luby-Phelps K (1999) Cytoarchitecture and physical properties of cytoplasm: Volume, viscosity, diffusion, intracellular surface area. In: Walter H, Brooks DE, Srere PA, editors. *International review of cytology*: Academic Press. pp. 189-221.
8. Wojciechowski M, Cieplak M (2008) Effects of confinement and crowding on folding of model proteins. *Biosystems* 94: 248-252. https://doi.org/10.1016/j.biosystems.2008.06.016.
9. Ostrowska N, Feig M, Trylska J (2019) Modeling crowded environment in molecular simulations. *Front Mol Biosci* 6: 86. https://dx.doi.org/10.3389/fmolb.2019.00086.

10. Zimmerman SB, Trach SO (1991) Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of escherichia coli. *Journal of Molecular Biology* 222: 599-620. https://doi.org/10.1016/0022-2836(91)90499-V.
11. Ellis RJ (2001) Macromolecular crowding: An important but neglected aspect of the intracellular environment. *Current opinion in structural biology* 11: 114-119. https://dx.doi.org/10.1016/s0959-440x(00)00172-x.
12. Ellis RJ, Minton AP (2003) Join the crowd. *Nature* 425: 27-28. https://dx.doi.org/10.1038/425027a.
13. Kuznetsova IM, Zaslavsky BY, Breydo L, et al. (2015) Beyond the excluded volume effects: Mechanistic complexity of the crowded milieu. *Molecules* 20. https://dx.doi.org/10.3390/molecules20011377.
14. Homouz D, Perham M, Samiotakis A, et al. (2008) Crowded, cell-like environment induces shape changes in aspherical protein. *Proceedings of the National Academy of Sciences* 105: 11754. https://dx.doi.org/10.1073/pnas.0803672105.
15. Bhattacharya A, Kim YC, Mittal J (2013) Protein-protein interactions in a crowded environment. *Biophys Rev* 5: 99-108. https://dx.doi.org/10.1007/s12551-013-0111-5.
16. Zhou HX (2013) Influence of crowded cellular environments on protein folding, binding, and oligomerization: Biological consequences and potentials of atomistic modeling. *FEBS Lett* 587: 1053-1061. https://dx.doi.org/10.1016/j.febslet.2013.01.064.
17. Alas SJ, González-Pérez PP (2016) Simulating the folding of hp-sequences with a minimalist model in an inhomogeneous medium. *Biosystems* 142: 52-67. https://doi.org/10.1016/j.biosystems.2016.03.010.
18. Alas SdJ, González-Pérez PP, Beltrán HI (2019) In silico minimalist approach to study 2d hp protein folding into an inhomogeneous space mimicking osmolyte effect: First trial in the search of foldameric backbones. *Biosystems* 181: 31-43. https://doi.org/10.1016/j.biosystems.2019.04.005.
19. Gonzalez-Perez PP, Orta DJ, Pena I, et al. (2017) A computational approach to studying protein folding problems considering the crucial role of the intracellular environment. *Journal of Computational Biology* 24: 995-1013. http://dx.doi.org/10.1089/cmb.2016.0115.
20. Gupta A, Maňuch J, Stacho L (2005) Structure-approximating inverse protein folding problem in the 2d hp model. *Journal of Computational Biology* 12: 1328-1345. http://dx.doi.org/10.1089/cmb.2005.12.1328.
21. Khodabakhshi AH, Maňuch J, Rafiey A, et al. (2008) Stable structure-approximating inverse protein folding in 2d hydrophobic-polar-cysteine (hpc) model. *Journal of Computational Biology* 16: 19-30. http://dx.doi.org/10.1089/cmb.2008.0096.
22. Hoque T, Chetty M, Sattar A (2009) Extended hp model for protein structure prediction. *Journal of Computational Biology* 16: 85-103. http://dx.doi.org/10.1089/cmb.2008.0082.
23. Hu J, Chen T, Wang M, et al. (2017) A critical comparison of coarse-grained structure-based approaches and atomic models of protein folding. *Phys Chem Chem Phys* 19: 13629-13639. https://dx.doi.org/10.1039/c7cp01532a.
24. Huang C, Yang X, He Z (2010) Protein folding simulations of 2d hp model by the genetic algorithm based on optimal secondary structures. *Computational Biology and Chemistry* 34: 137-142. https://doi.org/10.1016/j.compbiolchem.2010.04.002.
25. Shatabda S, Newton MAH, Rashid MA, et al. (2014) How good are simplified models for protein structure prediction? *Adv Bioinf* 2014: 867179-867179. http://dx.doi.org/10.1155/2014/867179.
26. Will S. Constraint-based hydrophobic core construction for protein structure prediction in the face-centered-cubic lattice. In: Altman RB, Dunker, A.K., Hunter, L., Klein, T.E., editor; 2002 2002; Singapore. World Scientific Publishing Co. pp. 661-672.
27. König R, Dandekar T (1999) Improving genetic algorithms for protein folding simulations by systematic crossover. *Biosystems* 50: 17-25. https://doi.org/10.1016/S0303-2647(98)00090-2.
28. Bertsekas DP (2009) Convex optimization theory. Belmont, Massachusetts, USA: Athena Scientific. 257 p.
29. Bertsekas DP, Nedic A, Ozdaglar AE (2003) Convex analysis and optimization. Belmont, Massachusetts, USA: Athena Scientific. 560 p.
30. Borwein J, Lewis AS (2010) Convex analysis and nonlinear optimization : Theory and examples. New York, NY: Springer.
31. Zgrablich G, Mayagoitia V, Rojas F, et al. (1996) Molecular processes on heterogeneous solid surfaces. *Langmuir* 12: 129-138. https://dx.doi.org/10.1021/la9408782.
32. Mayagoitia V, Rojas F, Kornhauser I, et al. (1997) Modeling of porous media and surface structures: Their true essence as networks. *Langmuir* 13: 1327-1331. https://dx.doi.org/10.1021/la950812m.

33. Hidalgo-Olguín DR, Cruz-Vázquez RO, Alas-Guardado SJ, et al. (2015) Lacunarity of classical site percolation spanning clusters built on correlated square lattices. *Transport in Porous Media* 107: 717-729. https://dx.doi.org/10.1007/s11242-015-0463-3.
34. Hoshen J, Kopelman R (1976) Percolation and cluster distribution. I. Cluster multiple labeling technique and critical concentration algorithm. *Physical Review B* 14: 3438-3445. https://dx.doi.org/10.1103/PhysRevB.14.3438.
35. Rothschild WG (1998) Fractals in chemistry. New York, NY [u.a.]: Wiley.

Figures

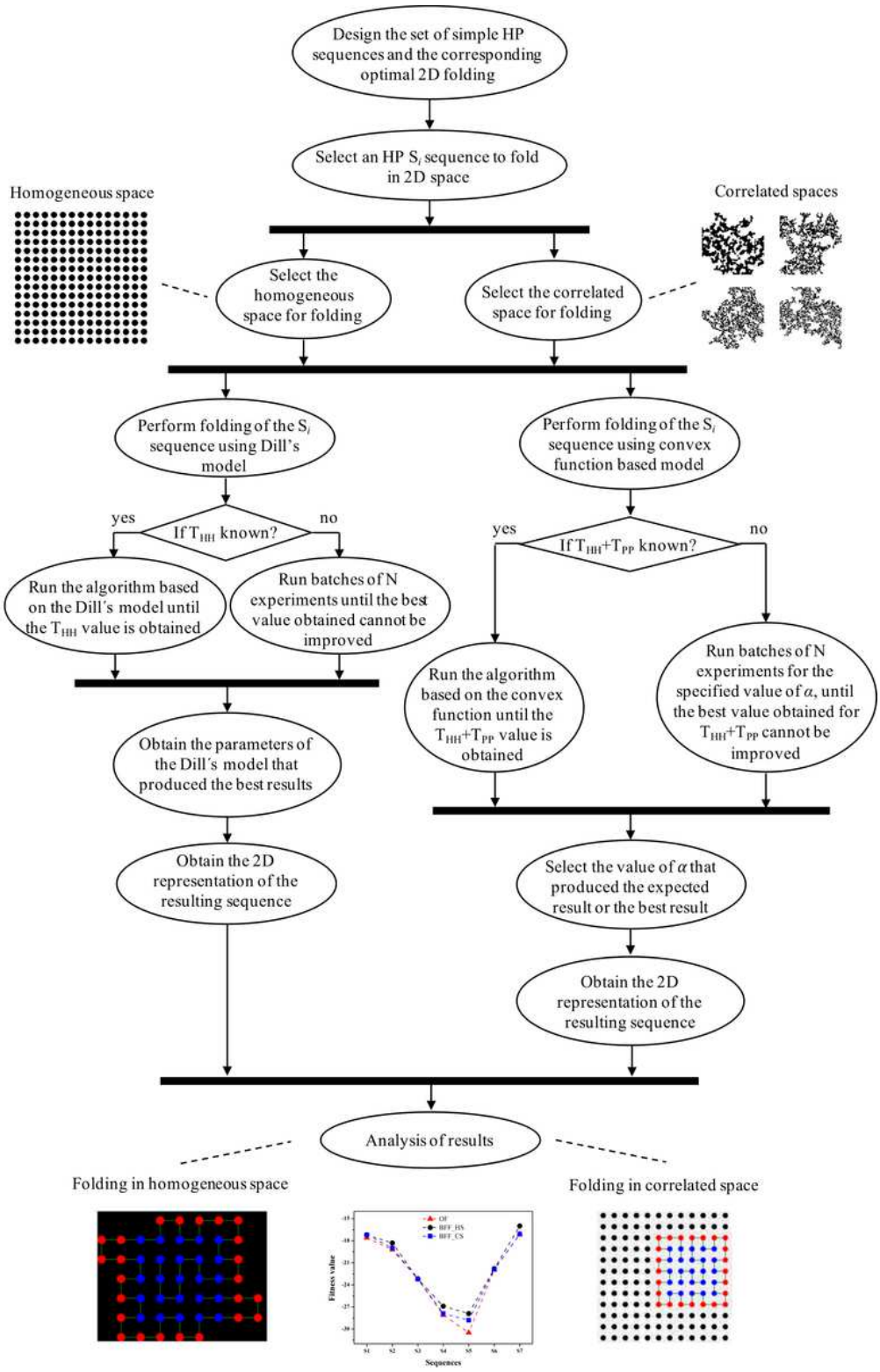

Figure 1

Methodological approach for in silico protein folding experiments. Note that conditions 1) and 2) previously mentioned are reflected in the right branch of the workflow. THH=Total of expected H···H contacts and TPP=Total of expected P···P contacts

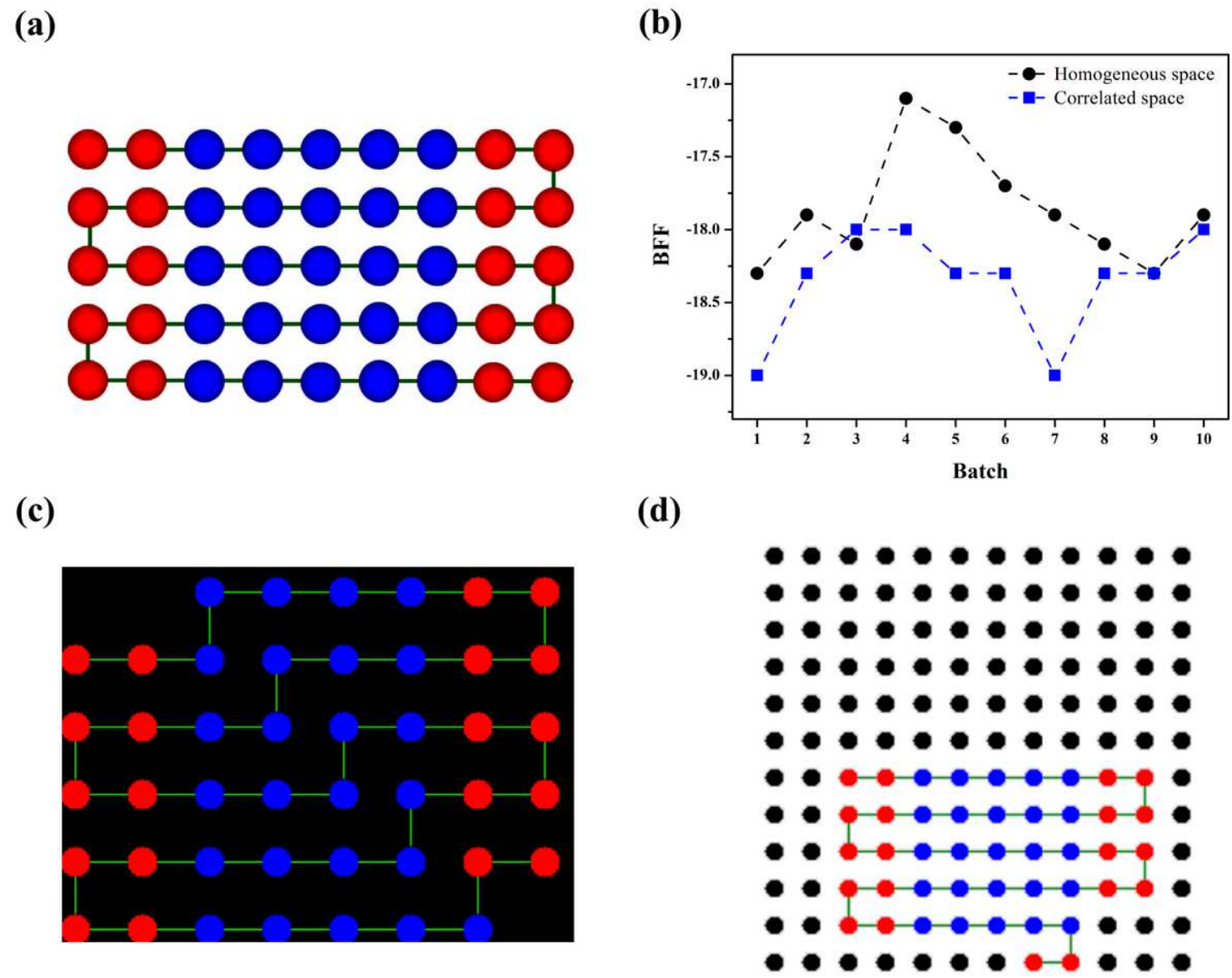

Figure 2

S2 HP 2D structures: (a) expected; (b) folding tendency graph; (c) FA1 with α = 0.1; (d) FA2, with α = 0.1 and ξB = 18.12. In this sequence, mainly due to terminal P2 branches, the hydrophobic core is restricted to form different degenerated structures, evidencing that the inclusion of those extra P□□□P contacts minimized degeneracy possibilities of S2, due to the formation of the two lateral PP substructures. Nevertheless, of the clear gain observed between 1c and 1d, the final score achieved does not reach the optimum but just by a small amount due to the bending of one of the two terminal P2 moieties. In (a), (c), and (d) the H and P residues are labeled in blue and red colors, respectively. The cluster grain is marked in black color circles in (d). For clarity in (d), the empty space of the correlated network is not shown here. In (b) dashed lines are only guides for the eye

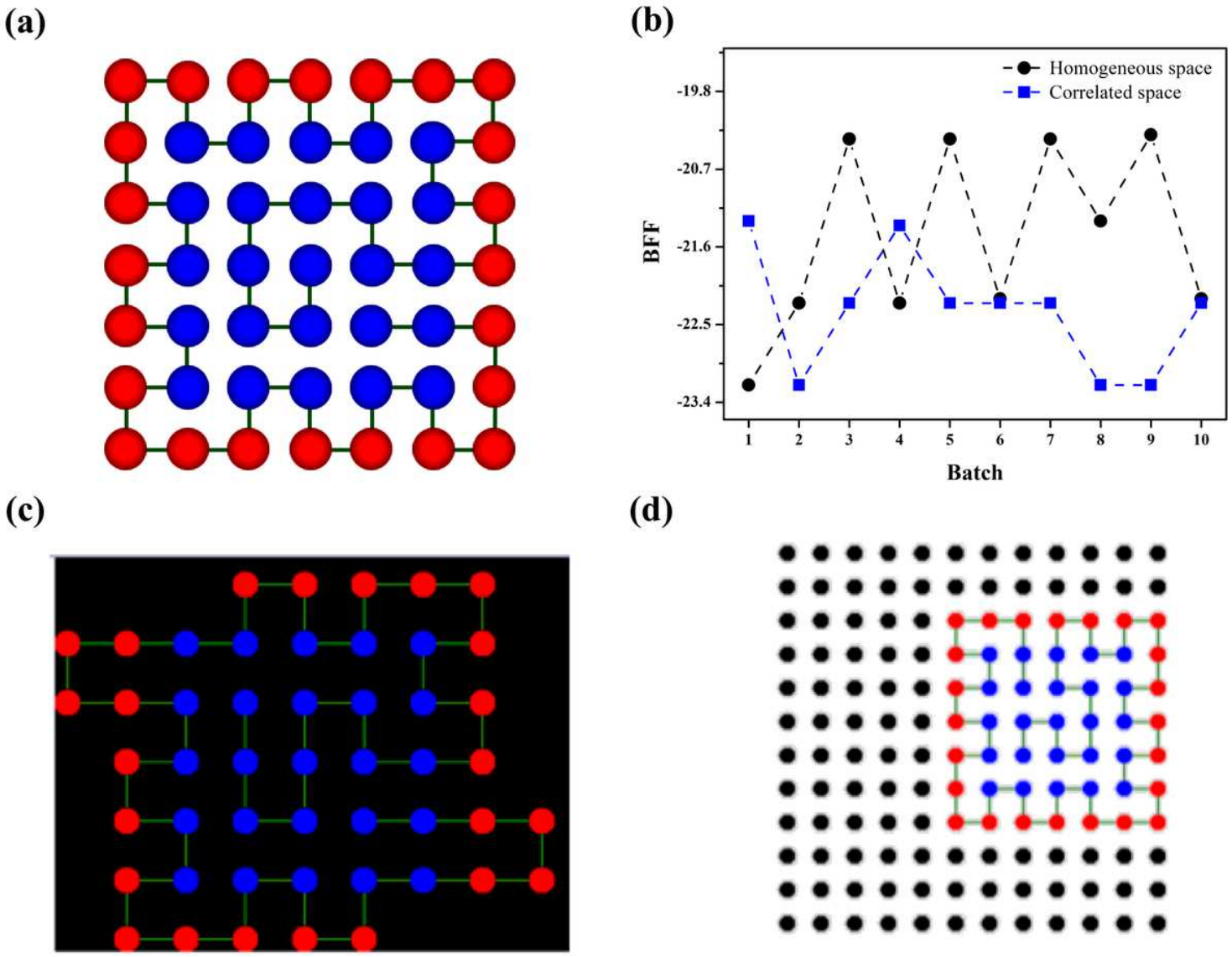

Figure 3

S3 HP 2D structures: (a) expected; (b) folding tendency graph; (c) FA1 with α = 0.05; (d) grain selection for folding; (d) FA2, with α = 0.05 and ξB= 18.12. In this sequence the hydrophobic core is built by 5×5 square and both terminal H moieties rely or are embedded into this core. Nevertheless, of the expected degeneracy into this hydrophobic core, see a), c) and d), the main complexity in this sequence regards precisely to the accommodation of the P boundaries, where FA1 yielded a more spread structure in c) but FA2 generated the expected symmetrical 7×7 square in a more precise matching with expected S3 mainly due to crowding imposed by the media. The colors labeled of the H and P residues and cluster grain are the same as Fig. 2. In (b) dashed lines are only guides for the eye

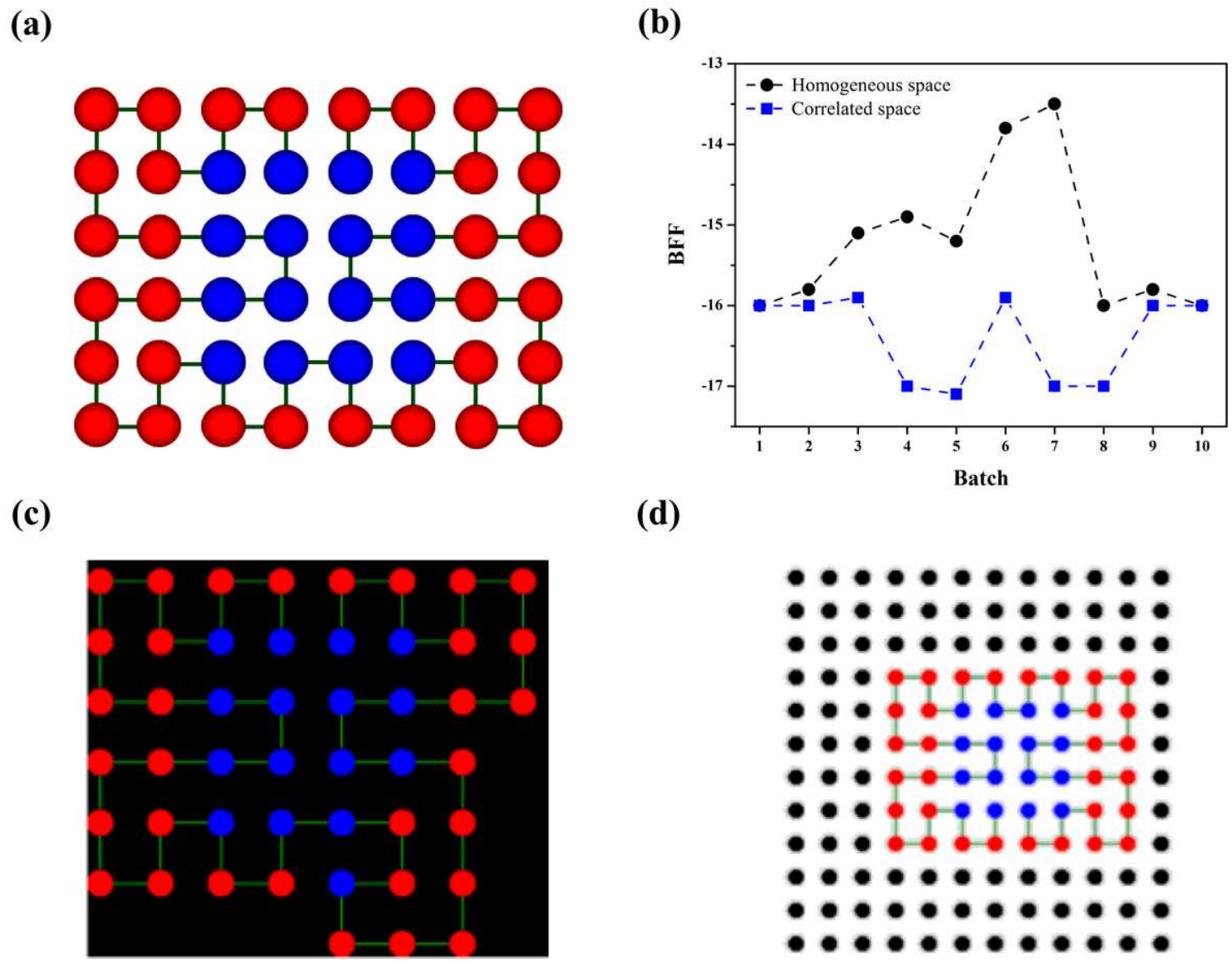

Figure 4

S7 HP 2D structures: (a) expected; (b) folding tendency graph; (c) FA1 with α = 0.1; (d) FA2, with α = 0.1 and ξB = 28.43. In this sequence, again, mainly due to P substructure loops, the hydrophobic core is one more time restricted to form different degenerated structures. Due to this particular sequence design, even the FA1 provided almost the desired structure (b) reaching a BFF = -16.0, correct folding was not achieved herein, due to the degrees of freedom for both the polar and hydrophobic regions of the sequence itself. Note that the desired structure is achieved (d) with the correlated space approach FA2, giving the requested BFF = -17.1. Note that the polar regions develop twisted loops in all corners, and both terminal H moieties rely or embed into the hydrophobic core in a very close match as in S3 sequence. The colors labeled of the H and P residues and cluster grain are the same as Fig. 2. In (b) dashed lines are only guides for the eye

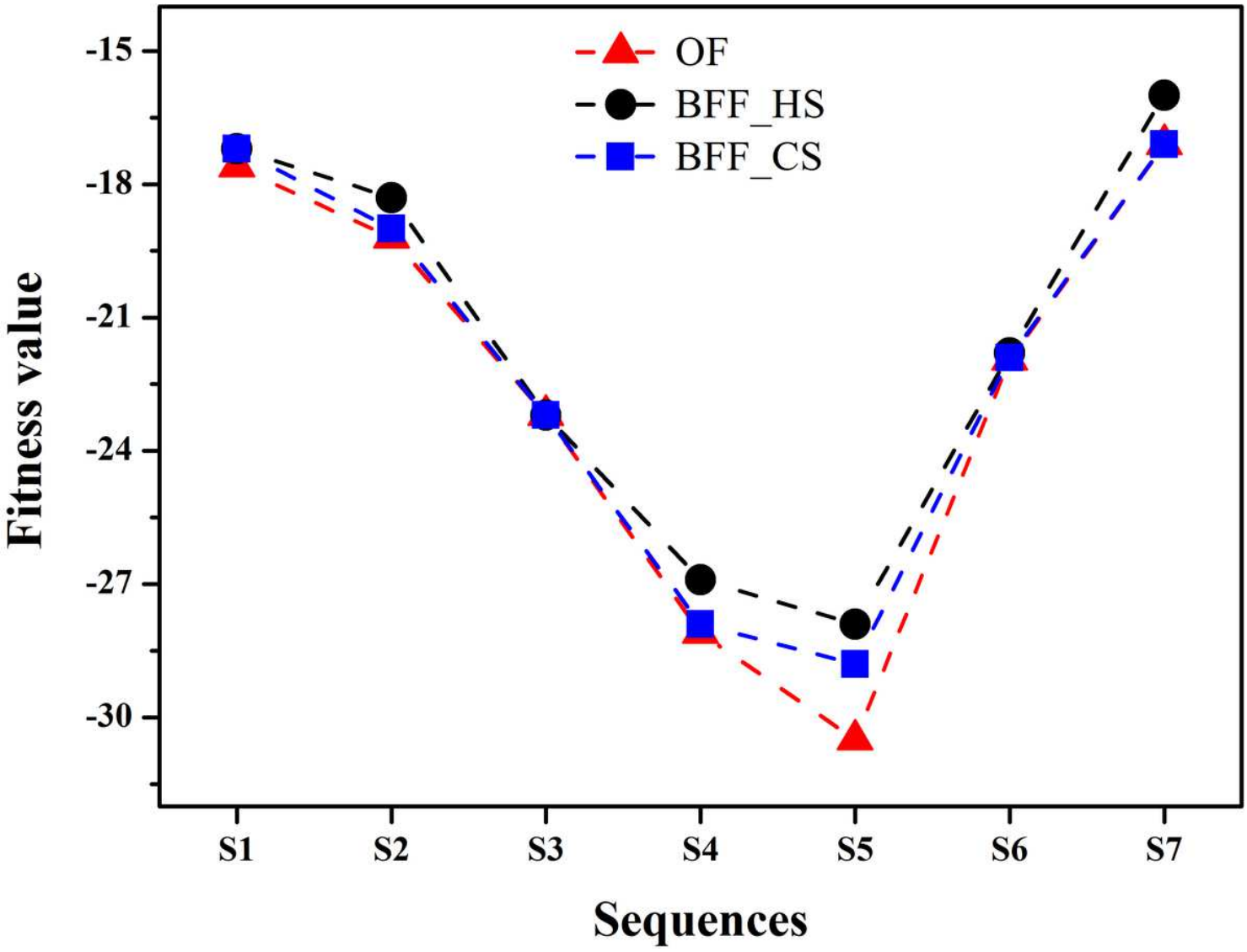

Figure 5

Folding tendency for sequences S1 to S7 in homogeneous and correlated media. OF = Optimal Fitness, BFF_HS = Best Fitness found in Homogeneous Space, BFF_CS = Best Fitness Found in Correlated Space. Dashed lines are only guides for the eye

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- JMMSupplementaryInfoPaperBeltranEtAl2021.docx
- GraphicalAbstractPaperBeltranEtAl2021.pdf