Conference paper

Leonor Cruzeiro*

The VES KM: a pathway for protein folding in vivo

https://doi.org/10.1515/pac-2019-0301

Abstract: While according to the thermodynamic hypothesis, protein folding reproducibility is ensured by the assumption that the native state corresponds to the minimum of the free energy in normal cellular conditions, here, the VES kinetic mechanism for folding in vivo is described according to which the nascent chain of all proteins is helical and the first and structure defining step in the folding pathway is the bending of that initial helix around a particular amino acid site. Molecular dynamics simulations are presented which indicate both the viability of this mechanism for folding and its limitations in the presence of a Markovian thermal bath. An analysis of a set of protein structures formed only of helices and loops suggests that bending sites are correlated with regions bounded, on the N-side, by positively charged amino acids like Lysine and Histidine and on the C-side by negatively charged amino acids like Aspartic acid.

Keywords: ICPOC-24; protein dynamics; protein folding; thermal baths; vibrational excited states.

Introduction

Proteins are very large molecules. Even a small protein, with 60 amino acids, will have approximately a 1000 atoms linked by covalent bonds to one another. In order to function properly, each protein, upon being synthesized in a cell, must assume a well defined average structure known as its native state. Given that proteins are flexible molecules one question, known as the protein folding problem, is how the native state is selected from among the immense number of other potentially accessible structures. Back in the late 1960's, when the studies on the protein folding problem started, two main hypotheses were proposed. One hypothesis, put forward by Levinthal [44, 45], was that the native state of protein is a kinetic trap and that folding reproducibility in cells is achieved because the folding process is a non-equilibrium process characterized by a specific pathway. A second hypothesis, put forward by Anfinsen [3], was that the native state is the global free energy minimum for each protein in the normal cellular environment and that folding is a thermodynamic equilibrium process of free energy minimization. Here, each amino acid sequence corresponds to only a single thermodynamically stable state (the native state), and folding reproducibility is achieved because, no matter which pathway the protein follows, it will always necessarily end at that state. Anfinsen's experiments of reversible unfolding of two proteins [3], which suggested that they can resume the native state even when starting from a completely unstructured state, favoured the thermodynamic hypothesis. Furthermore, the fact that proteins synthesized by a solid phase method can have an activity similar to that of the native state [35, 48] provided extra support. Thus, so far, Anfinsen's thermodynamic hypothesis has guided most of the research on the protein folding problem.

Article note: A collection of invited papers based on presentations at the 24th IUPAC International Conference on Physical Organic Chemistry (ICPOC 24) held in Faro, Portugal, 1–6 July 2018.

***Corresponding author: Leonor Cruzeiro,** CCMAR and Physics, FCT, Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal, e-mail: lhansson@ualg.pt. https://orcid.org/0000-0001-7958-6435

In spite of the progress in the understanding of the protein folding problem, clouds have been building (see [16, 19]). One was the discovery of proteins that misfold in the absence of mutations [52]. It is known that misfolded proteins tend to aggregate and it was first thought that only the proteins associated with misfolding diseases formed aggregates. It was a surprise to find that other proteins, like myoglobin, which are completely unrelated to such diseases, can also form aggregates [32]. In fact, it is now thought that all proteins can form aggregates [28] and the evidence is that the aggregated state is their true global free energy minimum (see figure 1 in [36]). Although there are cases of protein aggregation in cells, soluble proteins do manage to avoid the aggregated state most of the time and instead reach the native state, in spite of the fact that the latter is not the global free energy minimum and thus just a kinetic trap, as first proposed by Levinthal [44].

Other cases of proteins that can have more than one stable state in a non-aggregated condition are serpins [34], the α-lytic protease [58] and metamorphic proteins [59], as well as the proteins whose irreversible unfolding is unrelated to aggregation [50, 54]. There is also computational evidence of proteins that have many other thermodynamically stable conformations apart from the native one [16, 20–22], and at least 30 000 compact structures have been found in a systematic conformational search of a polypeptide formed by just sixty valines [14]. Taken together, these findings indicate that most proteins can assume many different, thermodynamically stable, structures. However, these results do beg the question of how proteins, upon synthesis in cells, almost always assume their native structure.

Although Levinthal did provide evidence for at least one protein which folds in cells via a kinetic process [44], he did not make any suggestions as to a description of a concrete pathway, and may even have been thinking in terms of a process of energy minimization [45]. In the next section a concrete pathway for folding in vivo, designated as quantum Vibrational Excited State Kinetic Mechanism, or VES KM for short in previous publications [16, 18, 19], is reviewed. In Section “Influence of the thermal bath” a study of the effect of thermal agitation on the VES KM is presented and in Section “Statistical analysis of native structures” a statistical analysis based on known protein structures is made in the light of the VES KM. The article ends with a discussion of these results and of the support they may provide for the VES KM.

The VES KM

Two requisites are needed for folding reproducibility if the native state is a kinetic trap. One is what Levinthal proposed: there is a pathway for folding, i.e. the trajectory followed by a given protein from synthesis to the native state must be always the same. The second requisite is that the starting structure must also be the same every time. If the nascent chain structure is not always the same, then following the same trajectory (defined for instance as a sequence of structural transformations to be operated on the initial structure) will generally not take a protein to the same state. If the native state is just one of the many kinetic traps that each protein can have, to ensure that every protein always (or most of the time) reaches the native state, the nascent chain must have a well defined structure and the pathway to the native state must always be the same.

The VES KM [16–19] proposes that the structure all proteins have as they come out of the ribosome is helical and, given its preponderance, it is reasonable to assume that for most proteins this helix is a α-helix. This defines the starting structure. Regarding the pathway, the VES KM proposes that the first operation on the initial helix is its bending at a particular amino acid site. The idea is that the bending site is specified by the local amino acid sequence, thus ensuring that it is always the same for each protein.

Once the bending has been completed the helical hairpin that results may continue to evolve in two ways. If the side chains that are brought together by the bending lead to a global interaction (protein plus environment) that is sufficiently attractive (negative) to compensate for the possible decrease in entropy due the formation of a more compact state, the helical hairpin thus formed will be thermodynamic stable and further evolution will consist only of minor adjustments of the backbone and side chains. If the global interaction is repulsive, or not sufficiently attractive, the structure will be unstable and part or all of the two helices will unfold, leading to the formation of β-sheets and/or loops.

Synthesis, especially for larger proteins, is known to have pauses [30, 31, 40, 41, 60, 62, 64]. The pathway outlined above applies to each protein region immediately after synthesis. However, the structure of the parts of the protein synthesized after the first will both influence and be influenced by the structure already assumed by the first.

This folding pathway has several implications. The first implication is that the ribosome is not just a synthesizing machine, it is also a chaperone which controls the structure of the nascent chain, something for which there is experimental evidence [30, 31, 40, 41, 60–62, 64]. The second implication is that the amino acids at the bending sites control the native structure. It is known that two proteins with more than 30 % sequence identity have a great probability of having similar folds and that this value decreases as the sequence size increases [33, 49]. In contrast, sometimes it takes only a single mutation to completely change a protein structure [1]. These experimental findings can be explained if a few amino acids control the native structure. In the VES KM, those amino acids are most probably one or more of those found at the bending site. Changing the position of the bending site leads to a different native state.

A third implication of the VES KM is that the bending of the helix must be sufficiently fast for the initial helix to avoid structure randomization due to thermal agitation, i.e. while conformational changes are many times viewed as the result of a large number of small changes whose cumulative effect becomes apparent only after microseconds or more, here the bending is assumed to be a concerted motion of a large number of atoms (a portion of the initial helix) in the nanosecond timescale, something more akin to normal modes [46], essential modes [2] or functional modes [37]. The viability of such fast collective motions has been confirmed both from experiments [13] and from simulations [18, 19].

Two crucial questions for the VES KM are: where does the energy for the bending of the nascent helix come from and how is it guaranteed that this energy is always delivered at the same site? The answer to those questions is the VES hypothesis [15]. According to the VES hypothesis the energy necessary for biological processes mediated by proteins is first stored and transported in the form of quantum vibrational excited states. In particular, McClare [47] and later Davydov and Scott [26, 55] proposed that the amide I band might be the vehicle for this quantum energy. The amide I vibration, which consists essentially of the stretching of the carbonyl of the peptide group [42], is particularly interesting since its one quantum state is resonant with the bending mode of water and its two quanta state is resonant with the OH stretch vibration of water.

Indeed, one possibility of creating amide I vibrations in proteins is via transfer from VES in water molecules [15]. On the other hand, in deuterated water, the vibrational modes have lower frequencies and will be off-resonance with the amide I in proteins, something which can be partly compensated by deuteration of the protein itself. Therefore, from the point of view of the VES KM, deuterated water will tend to block or slow down conformational changes and protein folding.

Another source of VES in proteins, first proposed by Davydov [26] who was interested in muscle contraction, is the hydrolysis of adenosinetriphosphate (ATP), which releases an energy equivalent to two quanta of amide I. Also, a similar energy input can be delivered to the protein by the binding of one calcium ion [9]. Extending this idea, it is reasonable to assume that the binding of many other ligands to a protein can result in the local creation of quantum VES in the protein. There are thus many possible sources of VES in proteins, some coming from their own enzymatic functions and others coming from the environment.

How can amide I excitations concentrate at a specific site in a protein? Amide I vibrations have an associated transition dipole moment and once created at a particular peptide group in a protein they can be transferred from carbonyl to carbonyl via dipole-dipole interactions [18, 23–26, 55]. Computer simulations which preserve the quantum nature of the amide I and the classical nature of the motion of peptide groups showed that, at finite temperature, the amide I vibrations hop stochastically from one amino acid site to another [18, 23–25]. In crystals, quantum excitations tend localize at defects, or at disordered regions, and they can also be enhanced by local electric fields [56, 57]. In the VES KM the idea is that eventually (in the picosecond timescale) the amide I vibrations will localize at a section of the initial helix that maximizes one or more of these physical characteristics. This pinpoints the location of the bending site in each nascent chain, i.e. a fourth implication of the VES KM is that the bending site is determined in a quantum process of energy propagation in the initial helix.

The VES hypothesis explains how relatively small amounts of energy, such as the amide I vibration, can be stored and transported in a protein without being dissipated on the way. It does not explain conformational changes. In order to have a complete picture of protein folding according to the VES KM it is necessary to explain how the VES energy is used to promote the bending of the initial helix. Quantum VES have a lifetime in the picosecond timescale [6, 29, 63]. Computer simulations have shown that in this short time span they can go from one end of a helix to another, many times over [18, 23–25]. Thus, VES can gather at a bending site in the picosecond timescale. Then, they decay. According to the VES KM, when they decay, the energy that was stored in VES is transferred, in the form of kinetic energy, to the amide group where VES was located, i.e. it is when VES decay that the amide group receives a strong kick. The fifth and final implication of the VES KM is that this kick is the initiator of the folding trajectory and that its first effect is to bend the initial helix around the bending site. Computer simulations have shown that a kick at a single amide group can indeed bend a helix (see [18, 19] and the next section for some constraints on this process).

Influence of the thermal bath

In previous studies a protein whose native state is as close to a helical hairpin as possible was selected, namely, protein PDB2HEP [4], a small all-α protein with just 42 amino acids and 692 atoms. Its native structure, obtained from the Protein Data Bank (PDB) [7] and energy minimized with the AMBER force field [10], is displayed on the left panel of Fig. 1. The right panel shows the putative structure of its nascent chain according to the VES KM, i.e. it shows protein PDB2HEP in the form of a α-helix.

In previous studies [18, 19] the effect of kicks to amide groups at different locations of the putative initial helix of protein PDB2HEP was investigated. The aim was to determine a possible location for the initial

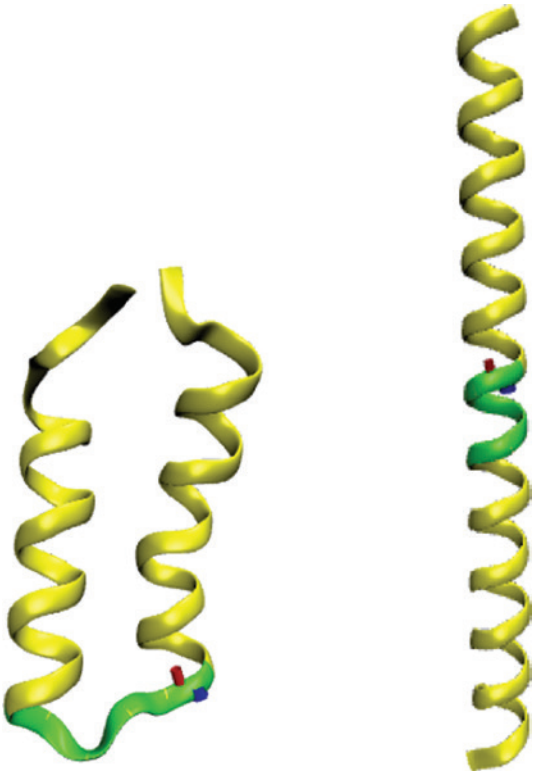

Fig. 1: The native structure of protein PDB2HEP [4], as obtained from the PDB [7] (left panel). The α-helical conformation assumed to be the corresponding nascent chain in the VES KM (right panel). The amino acids Gly20-Val21-Ile22-Thr23-Glu24 that constitute the link between the two helices are displayed in green. The amide group CO23NH24 where particular initial momenta were added initially (see text) are displayed as thick bonds in red and blue, respectively.

bending site of this protein. The simulations indicated that a kick in the amide group constituted by the carbonyl of Threonine (T) 23 and the amine of Glutamic acid (E) 24 (highlighted, respectively, in red and blue in Fig. 1 and referred to as CO23NH24) was the most efficient initiator of a trajectory capable of reaching the native basin.

In order to make the conditions as similar as possible for all the initial conditions tested, the previous simulations [18, 19] were made in the microcanonical ensemble, i.e. in the absence of thermal baths. However, proteins must fold and work at normal temperatures and their dynamics must survive the fluctuations associated with a thermal bath. Thus, in this section a preliminary study of the effects of thermal fluctuations is made using the following Langevin equations:

$$m_i \frac{d\vec{r}_i}{dt} = -\vec{\nabla} V(\{\vec{r}_i\}) + \vec{F}_i(t) - \gamma \frac{d\vec{r}_i}{dt} \tag{1}$$

Without the last two terms on the right-hand-side, (1) is just Newton's equation of motion for all the atoms in the protein, m_i and $\vec{r}_i$ being, respectively, the mass and position of atom i at time t and $V(\{\vec{r}_i\})$ being AMBER's potential energy function [10].

The second term and third terms on the right-hand-side of (1) transform Newton's deterministic equations of motion into Langevin equations when the stochastic forces, $\vec{F}_i(t)$, and the friction parameter, γ, also designated by collision frequency, obey the fluctuation-dissipation relation:

$$\ll \vec{F}_i(t)\vec{F}_j(t') \gg = 2\sqrt{m_i m_j}\,\gamma k_B T \delta_{ij} \delta(t-t') \tag{2}$$

T and k_B being, respectively, the target equilibrium temperature and Boltzmann's constant, and $\ll \cdots \gg$ standing for average over time.

For a sufficiently long trajectory, the latter relation between the stochastic forces and the damping term ensures that the temperature T is reached and then maintained on average throughout the trajectory. However, the last two terms do imply choices about the nature of the thermal bath. The delta functions mean that the thermal kicks in atom i are uncorrelated with those in atom j and also that the thermal kicks in atom i at one time instant are uncorrelated with the thermal kicks in the same atom at another time. Therefore, the thermal bath represented by (2) tends to randomize the velocities of all the atoms, and the larger γ is, the faster the randomization. Preventing such randomization effects requires, on the one hand, the knowledge of the protein relaxation time and, on the other hand, substitution of the Dirac delta function in time by a suitable memory kernel, both of which are unknown at present. Instead, in Fig. 2 we investigate how large γ must be in order to eliminate the conformational changes that are observed in the absence of a thermal bath. Other details of the simulations are as follows. The sander module of the AMBER package [10] was used in all the molecular dynamics simulations described in this section. The AMBER force field applied was ff99SB and all simulations were done without explicit water molecules and with solvation effects represented by a generalized Born/surface area model [parameter gbsa = 1 in the input file for sander (see the amber manual of amber 9 [10])]. The Langevin terms [the last two terms in Eq. (1)] were implemented by setting parameter ntt = 3 in the input file.

In all the trajectories presented here, the initial structure for protein PDB2HEP is essentially the same, namely, the α-helix displayed on the right panel of Fig. 1. As a measure of how close to the native basin the protein gets in each trajectory, the Root Mean Square Deviation (RMSD) of each conformation with respect to the native state (seen in the left panel of Fig. 1) is computed. For each value of γ, six trajectories are displayed, each in a different color. In each of the six trajectories, the protein receives the same energy input through a kick to the amide group CO23NH24. The main difference between the initial conditions of the six trajectories is in the velocities of the atoms that are not kicked, which were randomly selected from a Boltzmann distribution with $T = 298$ K. In Fig. 2 trajectories with the same color have exactly the same initial structure and velocities. This allows for a study of the effect of γ on each trajectory, separately. The values of γ used are given in the y-label of Fig. 2. $\gamma = 0$ means that Langevin terms were not included.

The top plot displays the trajectories resulting from a kick to amide group CO23NH24, in the absence of a thermal bath (i.e. these simulations are made in the microcanonical ensemble). Considering that the interval

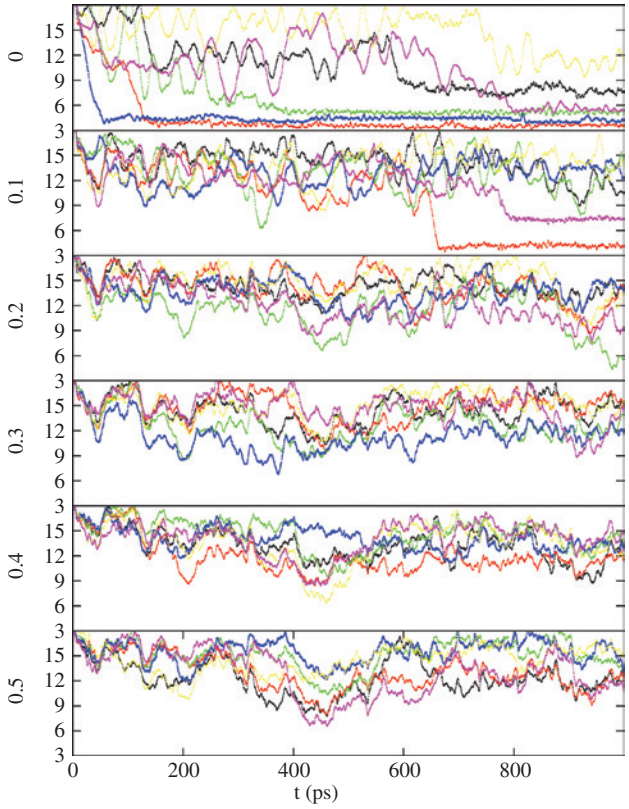

Fig. 2: The RMSD of each structure with respect to the native structure is displayed, for each value of the friction parameter γ (given in the *y*-axis in units of ps^{-1}). $\gamma = 0$ means that Langevin terms were not included. The initial structure for the protein was essentially the α-helix displayed on the right panel of Fig. 1. Also, all initial structures received the same energy input in the form of a kick to amide group CO23NH24. The difference in the six initial conditions are the velocities of the unkicked atoms which are randomly selected from a Boltzmann distribution with $T = 298$ K. In all the plots, the lines with the same color have exactly the same initial coordinates and velocities for all the atoms.

of 3 Å–6 Å in RMSD represents the native basin, or its vicinity, this plot shows that in these conditions, the protein has a 67 % chance of reaching the native state under 1 ns. In fact, two of the trajectories reach the native state even under 200 ps and remain stable after that. The following plots indicate that the introduction of a thermal bath tends to reduce the number of helical structures that transform into the native helical hairpin in such a time frame. For $\gamma = 0.1$ ps^{-1} one trajectory reaches the native basin and another approaches it, and for $\gamma = 0.2$ ps^{-1} only one trajectory reaches the native basin. For γ above 0.2 ps^{-1} none of the trajectories reaches the native basin in one nanosecond. It should be noted that the default value of γ in Amber is 5.0 ps^{-1}, at least ten times larger than the values used in Fig. 2.

Statistical analysis of native structures

According to the thermodynamic hypothesis the native state of proteins can arise from a completely unstructured initial conformation and there have been many studies to try to characterize the amino acid propensities and patterns that favour α-helix and β-sheet formation [5, 11, 38, 51, 53]. On the other hand, following the VES KM described in Section "The VES KM", the most important step in defining the native structure of a protein is the gathering of amide I vibrations at a particular amide group site where their decay leads to the bending of the initial (nascent) helix. This step is that which is effectively important for the selection of the native basin from among all the other potential conformational basins that can exist for the same protein in normal cellular conditions. From this point of view, the question that interests us is whether we can use the known native structures to characterize the amino acid sequences that constitute the bending site.

The simulations in Section "Influence of the thermal bath" indicate that the process of bending can be over in the nanosecond or even sub-nanosecond timescale, but the full folding trajectory, which involves the reaction of the protein to the bending (i.e. adjustment of all the protein atoms to the new backbone fold which may even lead to further helical unfolding and β-sheet or coil formation) can take microseconds or more to be completed. If the structure of the bending region does not change throughout the latter trajectory, the bending sites will correspond to some of the unstructured regions found in the native state, namely, turns, loops, and similar. However, the bending sites are not the only source of unstructured regions in the native state. Indeed, other unstructured regions may form further along the trajectory as the system drifts towards a lower energy state after the bending. Furthermore, the amino acids located at or near the bending site may themselves later be re-integrated into neighbouring helices or β-strands. Thus, the dynamical evolution following the initial bending can blur the location of its original site.

In order to try to minimize all the uncertainties mentioned in the previous paragraph, a restricted class of proteins was selected which, according to VES KM, has a greater probability of having evolved less from the putative helical hairpin that arises after the initial bending, namely, those proteins with only helices and loops (designated here as all-α proteins). Furthermore, in that protein set, only loops made of just four amino acids were considered as possible bending sites (designated hereafter as T4's). Thus, in the protein set selected T4's are non-helical bits surrounded by helices.

One reason for considering only T4's is that this is the size of the loop in protein PDB2HEP which involves amino acids Gly20-Val21-Ile22-Thr23 according to program Dictionary of Protein Secondary Structure (DSSP) that determines secondary structures from atomic coordinates [39]. In Fig. 1 the link between the two helices, shown in green, also includes Glu24, but according to DSSP (and also from visual inspection) this amino acid has a conformation characteristic of a helix.

A second reason for considering only T4's is that, in a α-helix, the C=O group of amino acid i is hydrogen-bonded to the HN group of amino acid $i+4$, which forms an amide group with the C=O group of amino acid $i+3$, making four the shortest number of amino acids at an initial bending site, according to the VES KM.

The file ss.txt, which includes all the protein sequences available from the PDB [7] together with their corresponding secondary structures as defined by DSSP [39], was downloaded from http://www.rcsb.org/pdb/files. A set of 10 751 all-α proteins was selected from it of which only 2397 proteins have T4's. The latter proteins have 8447 T4's and of the potential $20^4 = 8000$ different amino acid sequences, only 3430 were found.

Figure 3 shows the average amino acid composition of T4 loops, compared to the full amino acid composition of the 2397 proteins. This average amino acid composition for the full protein is calculated by counting the number of each amino acid (of the 20 most common ones) in each protein, dividing by the total number

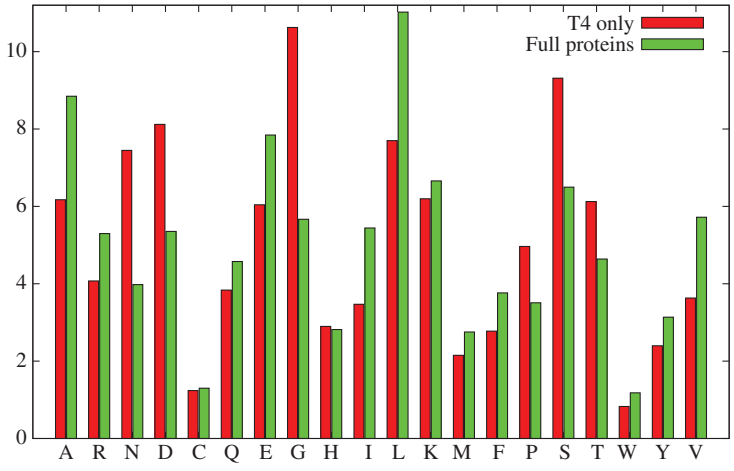

Fig. 3: Average amino acid composition for the whole protein and for T4's, expressed in the form of percentages. The green histogram is the average percentage of each amino acid in the full protein and the red histograms is their average percentage in T4's (see text).

of amino acids in that protein and averaging over all the proteins in the sample. In a similar way, the average amino acid composition of the T4's is done by counting the number of each amino acid in each T4, dividing by four times the total number of T4's found for each protein and averaging over the proteins with T4's.

In Fig. 3 these amino acid compositions are presented in the form of percentages, so that summing over all the amino acids gives 100 %. The green histograms in this figure show that some amino acids, like Leucine (L), Alanine (A), E, Lysine (K), Serine (S) and Valine (V) are more common in the all-α proteins, as happens in other proteins. However, this figure also shows that the amino acid composition of T4's differs significantly from the amino acid composition of the full protein. Indeed, L and A appear much less in T4's than in the full protein, and other amino acids, like Glycine (G), Serine (S), Aspartic acid (D), Asparagine (N) and T are much more frequent than in the full protein. While G, the smallest amino acid, is known to be associated with disordered regions of proteins, the greater presence of S, D, N and T in T4's may be due their polar or charged nature. Other possible explanations will be discussed in Section "Discussion and Conclusion".

From the point of view of the VES KM, even more important than the amino acid composition is the arrangement of those amino acids in the four positions of T4's. In Fig. 4 the amino acid distribution of each of the four positions of T4's is displayed. These distributions are calculated by counting the number of each amino acid in each position of the T4, dividing that number by the total number of T4's in that protein and averaging over all the proteins in the sample. The resulting amino acid frequencies are again presented in the form of percentages so that summing over all amino acids leads to 100 % in each position.

Figure 4 shows that the average amino acid composition varies very markedly from one position to another. Position 4 (histogram in violet) is that with the least uniform distribution, characterized by percentages of 16 for S's and D's compared to the values of 6.5 and 5.3, respectively, in the full protein (see Fig. 3). On the other hand, the most common amino acids in the whole protein, L and A, feature comparatively very little in position 4 of the T4's. Other amino acids with a strong presence in position 4 are N and T. G, on the other hand, is more common in the 2nd and 1st positions and very curious is also the fact that Proline (P) is completely absent from the 1st position. An interpretation of these findings in the light of the VES KM is given in Section "Discussion and Conclusion".

The T4's found in native states may arise because of the initial bending of the helix, but they can also be formed later on in the dynamical trajectories. In order to try to distinguish the two, the T4 sequences that appear also in structures that are either fully helical, or partly helical and partly loop, were separated from those sequences that appear only in T4's. The idea is that these latter T4's have a greater probability of having been created by the initial bending. Thus, the average amino acid composition was calculated for the 2092 T4's that are found in helices and mixtures and for the 1338 that are not. The results are displayed in Fig. 5.

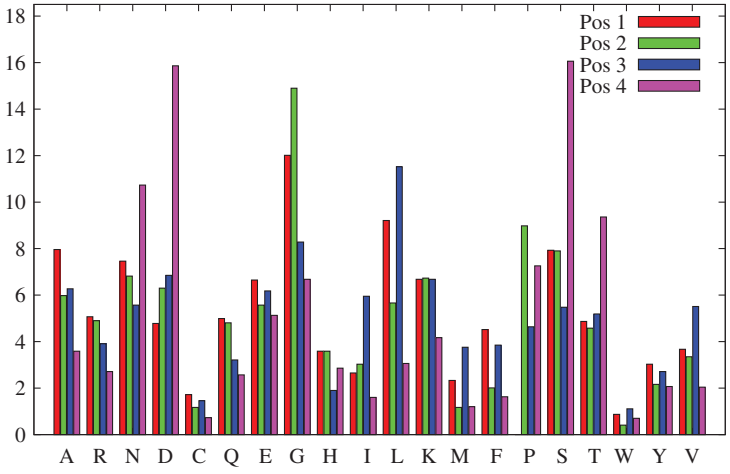

Fig. 4: Average amino acid composition in each of the four positions of all T4's, expressed in the form of percentages. Red is the histogram for position 1, green for position 2, blue for position 3 and violet for position 4.

Figure 5 shows that the marked differences already seen in Fig. 3 between the composition of T4's and that of the full protein are further exacerbated in the sequences that are found only in T4's. Indeed, T4's not found in helices or mixtures have even less L and A than the other T4's and more G, D and N (compare the respective blue and red histograms). T4's not found in helices or mixtures also have much more P, a little less T and approximately the same amount of S as the other T4's.

Since the average amino acid composition of the T4's that do not appear in helices and in structures that are partly helical, partly loop, is different from that of the T4's that do, it is also interesting to see what the positional averages of the former are like. Figure 6 portrays this new distribution.

One feature that stands out when comparing Figs. 4 and 6 is the greatly enhanced presence of P in positions 2, 3 and 4. Indeed, in position 2, P has now become the most prevalent amino acid. Other features of the positional amino acid distribution of T4's have become more marked, namely, D's and N's have become more frequent in position 4, while the numbers of S and T have been reduced. Also reduced in T4's not found in helices and mixtures is the number of L and A. On the other hand, a common feature between Figs. 4 and 6 is that P is absent from position 1.

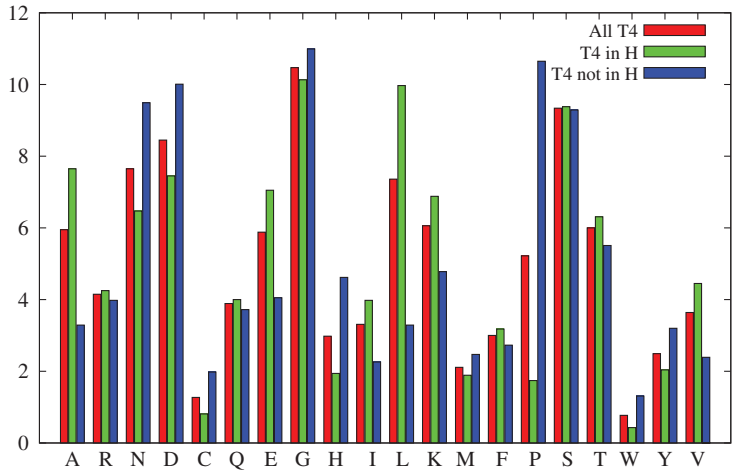

Fig. 5: Average amino acid composition for all T4's (red), T4's found in helices or in structures that are part helix, part loop (green), and the remaining T4's (blue), expressed in the form of percentages.

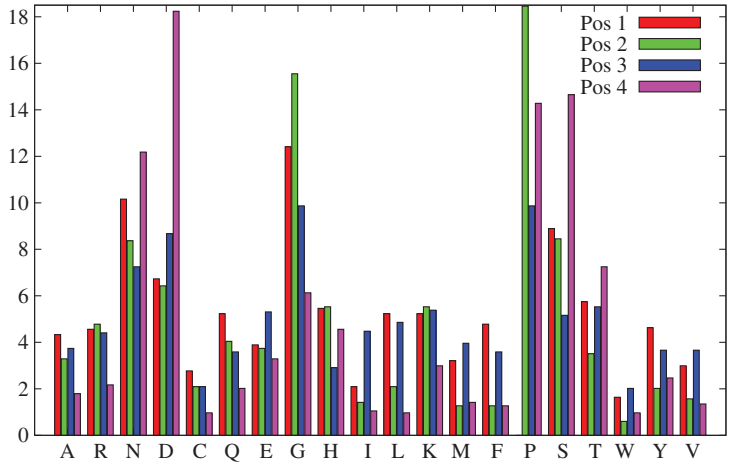

Fig. 6: Average amino acid composition in each of the four positions of T4's that are not found either in helices, nor in structures that are part helix, part loops, expressed in the form of percentages. Red is the histogram for position 1, green for position 2, blue for position 3 and violet for position 4.

Discussion and Conclusion

According to the VES KM, the first step in protein folding is the bending of the nascent helix at a specific amino acid site. The bending is the result of a kick whose energy comes from the decay of VES. In Section "Influence of the thermal bath" the influence of a Markovian thermal bath on the dynamical evolution immediately after the kick was investigated. The simulations show that such thermal baths, represented by (2), tend to reduce the folding efficiency of the initial kick, when compared to the dynamical evolution in the microcanonical ensemble (see Fig. 2). Although short, if the thermal bath is Markovian, these simulations do set constraints on the value of γ [see equations (1) and (2)] that is compatible with the VES KM, namely, γ should be less that 0.1 ps^{-1} (comparing to the standard values in Amber [10], which are 2–5 ps^{-1}, these are at least ten times smaller). On the one hand, these results may be interpreted as implying that the VES KM is not viable at biological temperatures. On the other hand, if protein conformational changes are indeed concerted motions of a large number of atoms triggered by a local action, as concepts like normal modes [46], or essential modes [2] or functional modes [37] imply, the limitation may instead be on the values of γ and/or on the model used for the thermal bath, which produces a randomization of the velocities that is too fast and does not respect protein's longer relaxation times. Indeed, those values of γ correspond to mode lifetimes of the order of 1 ps, far too short for conformational changes.

While the protein response time to a perturbation such as the kick simulated in Section "Influence of the thermal bath" is not known, even small proteins like protein PDB2HEP [4] can have folding times of microseconds. In many of the trajectories displayed in Fig. 2, the comparative large value of the RMSD is due to that the protein structure remains an open structure throughout. In general terms, one can expect that an isolated long helix, such as that portrayed in Fig. 1, will tend to bend and form a more compact structure that is stabilized by a greater number of attractive interactions, even if this compact structure is different from the native one. Here the aim was just to find the influence of a thermal bath on the trajectory in the short time after the kick. Although the simulations in the absence of a thermal bath show that the helical hairpin that characterizes the native state of protein PDB2HEP can form under one nanosecond, for a complete assessment of the effect of the kick and of thermal baths on the corresponding structural outcomes, longer simulations are needed. Such studies will be presented in forthcoming publications.

According to the thermodynamic hypothesis the reason some sequences lead to α-helices and other to β-sheets or turns is that the amino acids in those sequences have higher propensities for those particular secondary structures. But the fact that it is possible to form fibrils and aggregates from all proteins [28] shows that any amino acid sequence can be turned into sheets which have a free energy lower than that of a population of native structures (see figure 1 in [36]). This fact is in better agreement with the VES KM according to which most amino acid sequences can assume many different stable structures.

According to the VES KM, the ubiquitous presence of helices in native structures is due to that a helix is the default structure of all nascent chains, with other secondary structures arising only when the following helical hairpin is partially or totally unstable. In this view, the crucial information needed to determine the native state from the sequence is the knowledge of the location and amino acid composition of the bending sites. The putative bending sites, i.e. the T4's analysed in Section "Statistical analysis of native structures", are chain reversal regions which are known to be particularly well conserved in proteins [12]. This indicates that they are indeed important for structure and function, as assumed by the VES KM.

Following the thermodynamic hypothesis it is often argued that the main driving force in protein folding are hydrophobic interactions [27]. Considering that T4's tend to be at the surface of proteins, their amino acid composition (see blue histogram in Fig. 5) is in general agreement with this idea since the hydrophobic amino acids Isoleucine (I), V, L, Phenylalanine (F), Cysteine (C), Methionine (M) and A, here quoted in order of decreasing hydrophobicity [43], are not very frequent in T4's. Figure 5 shows that C and M are present in greater amounts in T4's found in helices than in T4's not found in helices, but their presence is not very representative. On the other hand, it would be expected that the most hydrophilic amino acids R, K, E and Q (here quoted in order of decreasing hydrophilicity [43]) should be present in greater relative amounts in T4's than in helices, but that is not the case (compare the red and blue histograms in Fig. 3 and the blue and green in Fig. 5).

The correlation we can establish is between the mass and the frequency of appearance of hydrophilic amino acids in T4's in the sense that those with smaller masses are more frequent, i.e. of the two amino acids that can be negatively charged, D (133 Da) is much more prevalent than E (147 Da), even if E is present in greater amounts in the full protein (compare the green histograms for D and E in Fig. 3). Similarly, of the three amino acids that can be positively charged, Lysine (K) (146 Da) is slightly more prevalent than H (155 Da) (more so than their relative weights in the full protein would predict) and both are more prevalent than R (174 Da), even if the latter's frequency in the full protein is greater than that of H. This is curious because the smaller the mass, the faster the response to the kick, i.e. a dependence on the mass is an indication of the importance of dynamics, rather than of thermodynamics. Finally, it is difficult to explain how such an undifferentiated cause as hydrophobic interactions can lead to the strong positional dependence of the amino acid frequency seen in Fig. 6.

Let us then consider a possible explanation for the positional distribution of amino acids in T4's which is characterized by a strong presence of D in the 4th position (as well as S and T) and a more pronounced presence of K and H, followed by R, in the first two or three. It has also been known for the past 30 years that, in the native state, D and also E – negatively charged – are preferentially placed at the N-end of helices, while R, K and H – positively charged – are preferentially found at the C-end [53]. The explanation offered for the positional preferences of D and E is that the three unpaired amine groups at the N-end of an α-helix can be stabilized by the side chains of hydrogen acceptors such as D and E [51]. From this perspective, the role of D at position 4 of the T4's would be to stabilize the N-end of the second of the two helices that form the helical hairpin after the initial bending. While D can indeed have such a role, this does not explain why E, which is present in larger amounts at the level of the whole protein, is less frequent at that position. Furthermore, such an explanation is not applicable to the prevalence of K, H and R at the C-end of helices because the carbonyl groups of the amino acids that are there tend to form hydrogen bonds directly to the backbone of the neighbouring loops. Below it is suggested instead that these positional favouritisms are the result of electric field effects on the VES KM.

In α-helices the C=O groups are approximately aligned with the helical axis and their permanent dipoles point from the C- to the N-terminal. Thus, each α-helix has a "macroscopic" electric dipole which points from the C- to the N-end. Negative charges, like D (and possibly S and T?), at the N-end, and positive charges, like K, H and R, at the C-end, create electric fields that counteract the "macroscopic" dipole field. Considering two helices separated by a T4 as in Fig. 1, the positively charged K's, H's, or R's that are at the C-end of the first helix are, in the nascent chain, at the N-end of the part of the helix to be transformed into a T4. And the negatively charged D's that are at the N-end of the second helix are, in the nascent chain, at the C-end of the helix to be transformed into a T4, i.e. the same charges that counteract the dipole field of the two helices in the native state reinforce that electric field in the T4 region of the nascent helix.

Electric fields affect both the energy of vibrational states and their absorption intensity, a phenomenon known as vibrational Stark effect [8]. Also, simulations in which a term that can represent a local electric field was included have led to an enhancement of the quantum number of VES [56, 57]. Higher excited VES lead to strong kicks when they decay and in this way local electric fields can help promote the bending of the initial helix and other conformational changes. Thus, according to the VES KM, the same charges that enhance the probability of bending at a given region of the nascent chain (the T4 region) are used to stabilize the two helices that arise after the bending.

Another puzzling feature of the positional amino acid distributions in T4's is the fact that P is never found in position 1 of T4's, something that is in marked contrast to its high probability of appearing in positions 2, 4 and 3 (see Fig. 6). As can be seen from this figure, there are many cases of amino acids with small positional probabilities, but none of them is rigorously zero, as happens to P in position 1 of T4's. Such a striking observation must be the result of a very strong constraint.

The absence of P in position 1 of loops has been noted before [38]. P is disruptive to a helix because its side chain binds into the backbone, creating a tertiary amide. α-helices are stabilized by hydrogen-bonds, and since a P at site i prevents the formation of a hydrogen-bond with amino acid at i–4, P introduces a local instability in a α-helix and is known as a helix breaker. Therefore, a simple explanation for the absence of P's from position 1 might be that P always disrupts the α-helix and drags the amino acid before it into the

disruption, i.e. into the T4. However, if this were true, P's and the amino acids to the N-side of P's should always be in a loop and there shouldn't be any P's in helices, while Fig. 5 shows that, although the presence of P's in T4 sequences found in α-helices is markedly lower than in T4's, helices do survive a 2 % presence of P's. This number should be compared to the exactly 0 % value in position 1 of T4's.

From the point of view of VES, a P at site *i* is disruptive for two more reasons. One is that the amide I excitation in the tertiary amide of P is uncoupled from the rest of the other amide groups in the helix, making it essentially a local vibrational mode [42]. This diminishes the energy stored in it and consequently weakens the strength of a potential kick if VES were to decay at a P site, something that precludes the conformational change [19]. A second reason for P to be disruptive to the VES KM is that, by binding its side chain directly into the backbone, P makes the decay site heavier than that of a primary amide, leading to a faster dissipation of the initial kick. The overall conclusion is that P's should not be kicking sites. Thus, in the T4's that are bending sites and which happen to include P's, the initial kick must be in any of the other non-P sites, either to the N-side or to the C-side of P. But if the kick site was to the C-side of P, then it should be possible to have P's in position 1 of T4's as long as any of the other four positions was not a P. The fact that there are no P's in position 1 suggests that kicks are in amino acids sites to the N-side of P.

However, this explanation, appealing as it may be for supporters of the VES KM, only applies if we assume that all T4's found in the statistical analysis of Section "Statistical analysis of native structures" are bending sites. For those T4 sequences that are also partially or wholly localized in helices in the native structure, this means that, having been a bending site to start with, they must have been re-integrated into helices later on in the folding trajectory. This is not impossible especially considering that the proteins selected, with just helices and loops, are arguably the set that will have evolved the less from the helical hairpin that arises from the initial bending. Still, it seems more likely that some of the T4 sequences that are also found in helices are not initial bending sites, in which case the absolute zero probability for P to be in position 1 of T4's remains a mystery.

On the other hand, while it is not yet possible to identify the kicking site, a general conclusion of this work is that the bending sites are associated with enhanced electric fields and thus, a prediction of the VES KM is that changes in amino acids that affect local electric fields will have a strong effect on the folding of the corresponding proteins. In particular, sequences bounded on the N-side by K and H (or R), and/or by D (or S, or N) on the C-side, are the most probable candidates for the initial bending sites of the VES KM.

Acknowledgements: This study received Portuguese national funds from FCT – Foundation for Science and Technology through project UID/Multi/04326/2019. The author also acknowledges the Laboratory for Advanced Computing at University of Coimbra (http://www.lca.uc.pt) for providing HPC computing resources that have contributed to the research results reported in Section "Influence of the thermal bath" of this paper.

References

[1] P. A. Alexander, Y. He, Y. Chen, J. Orban, P. N. Bryan. *Proc. Natl. Acad. Sci. USA* **106**, 21149 (2009).
[2] A. Amadei, A. Linssen, H. Berendsen. *Proteins: Struct. Funct. Genet.* **17**, 412 (1993).
[3] C. Anfinsen. *Science* **181**, 223 (1973).
[4] J. Aramini, S. Sharma, Y. Huang, G. Swapna, C. Ho, K. Shetty, K. Cunningham, L. Ma, L. Zhao, L. Owens, M. Jiang, R. Xiao, J. Liu, M. Baran, T. Acton, B. Rost, G. Montelione. *Proteins* **72**, 526 (2008).
[5] R. Aurora, G. Rose. *Prot. Sci.* **7**, 21 (1998).
[6] R. H. Austin, A. Xie, L. van der Meer, M. Shinn, G. Neild. *J. Phys.-Cond. Matt.* **15**, S1693 (2003).
[7] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne. *Nuc. Acid. Res.* **28**, 235 (2000).
[8] D. M. Bishop. *J. Chem. Phys.* **98**, 3179 (1993).
[9] G. Careri, U. Buontempo, F. Galluzzi, A. Scott, E. Gratton, E. Shyamsunder. *Phys. Rev. B* **30**, 4689 (1984).
[10] D. Case, T. I. Cheatham, T. Darden, H. Gohlke, R. Luo, K. J. Merz, A. Onufriev, C. Simmerling, B. Wang, R. Woods. *J. Computat. Chem.* **26**, 1668 (2005).
[11] P. Y. Chou, G. D. Fasman. *Biochemistry* **13**, 222 (1974).
[12] P. Y. Chou, G. D. Fasman. *Biophys. J.* **26**, 385 (1979).

[13] H. S. Chung, K. McHale, J. M. Louis, W. A. Eaton. *Science* **335**, 981 (2012).
[14] P. Cossio, A. Trovato, F. Pietrucci, F. Seno, A. Maritan, A. Laio. *PLoS Comput. Biol.* **6**, e1000957, 1 (2010).
[15] L. Cruzeiro. *J. Phys. Org. Chem.* **21**, 549 (2008).
[16] L. Cruzeiro. "Protein folding", in *Chemical Modelling*, M. Springborg (Ed.), p. 89, Royal Society of Chemistry, London, UK 7 (2010).
[17] L. Cruzeiro. *Bio-Algorithms and Med-Systems* **10**, 117 (2014).
[18] L. Cruzeiro. *Z. Phys. Chem* **230**, 743 (2016).
[19] L. Cruzeiro. "Protein folding in vivo: from Anfinsen back to Levinthal", in *Nonlinear Systems, Vol. 2. Understanding Complex Systems*, J. Archilla, F. Palmero, M. Lemos, B. Sánchez-Rey, J. Casado-Pascual (Eds.), pp. 3–38, Springer, Cham (2018).
[20] L. Cruzeiro, L. Degrève. *J. Biomol. Struct. Dyn.* **33**, 1539 (2015).
[21] L. Cruzeiro, L. Degrève. *J. Biol. Phys.* **43**, 15 (2016).
[22] L. Cruzeiro, P. A. Lopes. *Mol. Phys.* **107**, 1485 (2009).
[23] L. Cruzeiro-Hansson. *Europhys. Lett.* **33**, 655 (1996).
[24] L. Cruzeiro-Hansson. *Phys. Lett. A* **223**, 383 (1996).
[25] L. Cruzeiro-Hansson, S. Takeno. *Phys. Rev. E* **56**, 894 (1997).
[26] A. Davydov. *J. Theor. Biol.* **38**, 559 (1973).
[27] K. A. Dill. *Biochemistry* **29**, 7133 (1990).
[28] C. M. Dobson. *Nature* **426**, 884 (2009).
[29] J. Edler, P. Hamm. *J. Chem. Phys.* **117**, 2415 (2002).
[30] J. P. Ellis, C. K. Bakke, R. N. Kirchdoerfer, L. M. Jungbauer, S. Cavagnero. *ACS Chem. Biol.* **3**, 555 (2008).
[31] M. S. Evans, I. M. Sander, P. L. Clark. *J. Mol. Biol.* **383**, 683 (2008).
[32] M. Fändrich, M. A. Fletcher, C. M. Dobson. *Nature* **410**, 165 (2001).
[33] A. Fersht. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*, W. H. Freeman and Company, New York (1999).
[34] P. Gettins. *Chem. Rev.* **102**, 4751 (2002).
[35] B. Gutte, R. B. Merrifield. *J. Biol. Chem.* **246**, 1922 (1971).
[36] F. U. Hartl, M. Hayer-Hartl. *Nature Struc. Mol. Biol.* **16**, 574 (2009).
[37] B. L. d. G. Jochen S. Hub. *PLoS Comput. Biol.* **5**, e1000480 (2009).
[38] C. R. K. Gunasekaran, H. A. Nagarajaram, P. Balaram. *J. Mol. Biol.* **275**, 917 (1998).
[39] W. Kabsch, C. Sander. *Biopolymers* **22**, 2577 (1983).
[40] C. M. Kaiser, D. H. Goldman, J. D. Chodera, I. Tinoco Jr., C. Bustamante. *Science* **334**, 1723 (2011).
[41] C. Kimchi-Sarfaty, J. M. Oh, I. W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, M. M. Gottesman. *Science* **315**, 525 (2007).
[42] S. Krimm, J. Bandekar. *Adv. Prot. Chem.* **22**, 181 (1986).
[43] J. Kyte, R. Doolittle. *J. Mol. Biol.* **157**, 105, 1982.
[44] C. Levinthal. *J. Chim. Phys.* **65**, 44 (1968).
[45] C. Levinthal. In *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House*, J. T. P. DeBrunner, E. Munck (eds.), Monticello, Illinois, volume 22, 22 (1969).
[46] M. Levitt, C. Sander, P. Stern. *Int. J. Quant. Chem.* **10**, 181 (1983).
[47] C. McClare. *Ann. N.Y. Acad. Sci.* **227**, 74 (1974).
[48] R. B. Merrifield. *Protein Sci.* **5**, 1947 (1996).
[49] L. Meyerguz, J. Kleinberg, R. Elber. *P. Natl. Acad. Sci. USA* **104**, 11627 (2007).
[50] R. K. Mitra, S. S. Sinha, S. K. Pal. *Langmuir* **23**, 10224 (2007).
[51] L. Presta, G. Rose. *Science* **240**, 1632 (1988).
[52] S. B. Prusiner. *Science* **216**, 136 (1982).
[53] J. S. Richardson, D. C. Richardson. *Science* **240**, 1648 (1988).
[54] J. M. Sanchez-Ruiz. *Biophys. Chem.* **148**, 1 (2010).
[55] A. Scott. *Phys. Rep.* **217**, 1 (1992).
[56] P. A. S. Silva, L. Cruzeiro-Hansson. *Phys. Lett. A* **315**, 447 (2003).
[57] P. A. S. Silva, L. Cruzeiro-Hansson. *Phys. Rev. E* **74**, 021920 (2006).
[58] J. Sohl, S. Jaswal, D. Agard. *Nature* **395**, 817 (1998).
[59] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, B. F. Volkman. *P. Natl. Acad. Sci. U.S.A.* **105**, 5057 (2008).
[60] K. G. Ugrinov, P. L. Clark. *Biophys J.* **98**, 1312 (2010).
[61] N. R. Voss, M. Gerstein, T. A. Steitz, P. B. Moore. *J. Mol. Biol.* **360**, 893 (2006).
[62] D. N. Wilson, R. Beckmann. *Curr. Opin. Struc. Biol.* **21**, 274 (2011).
[63] A. Xie, L. van der Meer, W. Hoff, R. H. Austin. *Phys. Rev. Lett.* **84**, 5435 (2000).
[64] G. Zhang, Z. Ignatova. *Curr. Opin. Struc. Biol.* **21**, 25 (2011).