J Biol Phys (2017) 43:15–30
DOI 10.1007/s10867-016-9431-6

ORIGINAL PAPER

Exploring the Levinthal limit in protein folding

Leonor Cruzeiro[1] · **Léo Degrève**[2]

Received: 29 March 2016 / Accepted: 5 September 2016 / Published online: 14 October 2016

Abstract According to the thermodynamic hypothesis, the native state of proteins is uniquely defined by their amino acid sequence. On the other hand, according to Levinthal, the native state is just a local minimum of the free energy and a given amino acid sequence, in the same thermodynamic conditions, can assume many, very different structures that are as thermodynamically stable as the native state. This is the Levinthal limit explored in this work. Using computer simulations, we compare the interactions that stabilize the native state of four different proteins with those that stabilize three non-native states of each protein and find that the nature of the interactions is very similar for all such 16 conformers. Furthermore, an enhancement of the degree of fluctuation of the non-native conformers can be explained by an insufficient relaxation to their local free energy minimum. These results favor Levinthal's hypothesis that protein folding is a kinetic non-equilibrium process.

Keywords Protein folding · Kinetic mechanism · Molecular dynamics

1 Introduction

According to Anfinsen's thermodynamic hypothesis, the native state of proteins is solely determined by their amino acid sequence [1]. Anfinsen also states that "... the three-dimensional structure of a native protein in its normal physiological milieu (...) is the one in which the Gibbs free energy of the whole system is lowest" [1]. In this perspective, the protein folding process is viewed as a progressive search for the global minimum of

✉ Leonor Cruzeiro
lhansson@ualg.pt

[1] CCMAR and FCT, Universidade do Algarve, Campus de Gambelas, Faro 8005-139, Portugal

[2] Grupo de Simulação Molecular, Departamento de Química, Faculdade de Filosofia,
Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes 3900,
14040-901, Ribeirão Preto, São Paulo, Brazil

the free energy. In the context of protein folding, Levinthal is usually quoted with regard to his demonstration that if proteins folded from an initial random structure by exploring all possible conformations they would take much longer to fold than they do [2] (the so-called Levinthal paradox). Within the thermodynamic hypothesis, a putative solution to this Levinthal paradox is the notion that the energy landscape of proteins is funnel shaped so that, in their progress to the global minimum at the bottom of the funnel, proteins only have access to an ever-decreasing number of conformations [3, 4]. On the other hand, Levinthal's solution to his supposed paradox is rather different, namely, it is the proposal that the native state of proteins is merely a *local* free energy minimum and that protein folding is a kinetic process in which proteins follow specific conformational pathways [5]. In spite of attempts to conciliate these two views [4, 6] the fact is that, according to Anfinsen's thermodynamic hypothesis, each protein can only assume *one* stable well-defined three-dimensional structure and, according to the Levinthal limit, proteins can have *many different structures*, as thermodynamically stable as the native state. The aim of this work is to explore this Levinthal limit and investigate the causes of the stability of the different conformers of the same protein.

Making the same protein assume different folds is a very difficult task to do experimentally. On the other hand, computationally, it is possible to force a given amino acid sequence into many different folds and use molecular dynamics simulations (MDS) to study their relative stability. In an early study [7], four proteins were selected and, for each one, three decoys were built by using the backbone folds of the other proteins. MDS showed that the decoys had average energies and average fluctuations similar to those of the corresponding native state, so that, in a blind choice, each decoy would be as probable as the native state. In this study, the only distinguishing feature was the resistance to a heat pulse that was greater for the native state than for the decoys [7]. Two limitations of the latter study were (1) that the duration of the MDS was at most 50 nanoseconds (ns) and (2) that those simulations were made in the absence of explicit water. To improve upon those limitations, in [8], MDS with a duration of at least 500 ns, in the presence of explicit water molecules, were performed, which indicate that proteins can have many non-native states that are dynamically as stable as the native state. Indeed, during the 500 ns, none of the nine non-native states probed in [8] showed any tendency to evolve towards the native state, as should be expected from Anfinsen's thermodynamic hypothesis. However, an intriguing result of the study in [8] was also that all of the non-native conformers exhibited a degree of fluctuation greater than that of the corresponding native structures. This raises the question of whether this greater degree of flexibility can be due to a difference between the nature of the stabilizing interactions of non-native states versus the native states. Thus, here we analyze the contributions of the different types of interactions to the total potential energy of native and non-native structures of the same protein.

2 Methods

We use the same four proteins as previously [8]. According to the CATH protein structure classification scheme [9], all the current protein structures in the Protein Data Base (PDB) [10] fit into just four classes. Since each of the four proteins used in this work belongs to a different class, in spite of their relatively small number, they cover all the known protein structure classes. Indeed, protein PDB1BDD [11] belongs to the mainly-α class, protein PDB1J08 [12] belongs to the mainly-β class, protein PDB1IGD [13] belongs to the α/β and protein PDB1AAP [14] belongs to the "few secondary structures" class. The protein

1BDD includes one histidine, which can be either protonated or unprotonated. Here we are concerned with the protonated protein.

As explained in the introduction, our purpose is to test whether we can differentiate between the native and non-native states of the same protein via the type of interactions that stabilize them. To this end, for each of the four proteins, three alternative, non-native conformations were generated by threading the sequence of the first protein onto the fold of the other three, using the Leap module of the AMBER package [15] to add the coordinates of the residue atoms. For instance, one mainly-β conformation for the mainly-α protein 1BDD was produced by taking the coordinates of the backbone atoms of the native state of the mainly-β protein 1J08 and adding the coordinates for the residues of 1BDD to that backbone. In this way, for each of the four proteins, three non-native conformations were generated. All such sixteen initial structures (four native, taken from the PDB [10], plus three non-native structures for each protein) are displayed in Fig. 1.

Since we want to compare the interactions in native and non-native structures, ideally, we should obtain their initial coordinates in the same manner, i.e., all from experimental measurements. However, this is impossible to do for the non-native states and the consequence is that, while for the native states of the four proteins the coordinates for both the backbone atoms and the side chain atoms are taken from the PDB [10], for the non-native conformations, as explained in the previous paragraph, the coordinates of the side chains are added in a more artificial manner. To minimize this difference that exists between native and non-native structures from the very start, we also created four additional "native" structures for which the coordinates of the backbone atoms came from the PDB files [10], but for which the coordinates of the side chains were added by the LEAP module of the AMBER package [15]. These latter "native" structures, whose generation is closer to that of the non-native structures, are designated as LEAP structures in the next section. In summary, in this work, for each protein, we will analyze the trajectories from *five* different initial structures: the full native PDB structure, the LEAP native structure and the three non-native structures. Figures 2–6, however, display results from the trajectories of the PDB and of the three non-native conformations, the four main conformers of each protein.

All protein structures were placed in an explicit water bath and sodium ions were added to make the whole system electrically neutral. All simulations were performed with GROMACS [16–19], using the Gromos96 43A1 force field [20] together with the SPC/E water potential [21]. The systems constituted by the protein and the added water molecules were all first subjected to energy minimization and resulting structures were inserted as initial conditions for molecular dynamics at constant temperature and pressure (the NPT ensemble), with T=300 K and p=1 atm. For all simulations, the time step for the integration was 2 femtoseconds (fs) and the total integration time was 0.5 microseconds (μs).

Figure 2 shows the evolution of the protein-only potential energies of the four main conformers of the four proteins (the black curve is for the PDB native conformer of each protein). Protein-only means that the interactions of the protein atoms with the solvent, as well as solvent–solvent interactions, are not included. This figure shows that, in all trajectories, the potential energy has converged after 250 nanoseconds (ns). It also shows that, although the native conformer is that which tends to have the lowest protein–protein potential energy, there are other conformers that can have similar energies. For example, for the mainly-α 1BDD protein, the conformer with the α/β fold of 1IGD has a similar protein–protein potential energy as the native 1BDD (compare the yellow curve with the black in the top plot); for the mainly-β 1J08 protein, the conformer with the mainly-α fold of 1BDD has a similar protein–protein potential energy as the native 1J08 (compare the blue curve with the black in the second plot from the top); and for the α/β 1IGD protein, the conformer

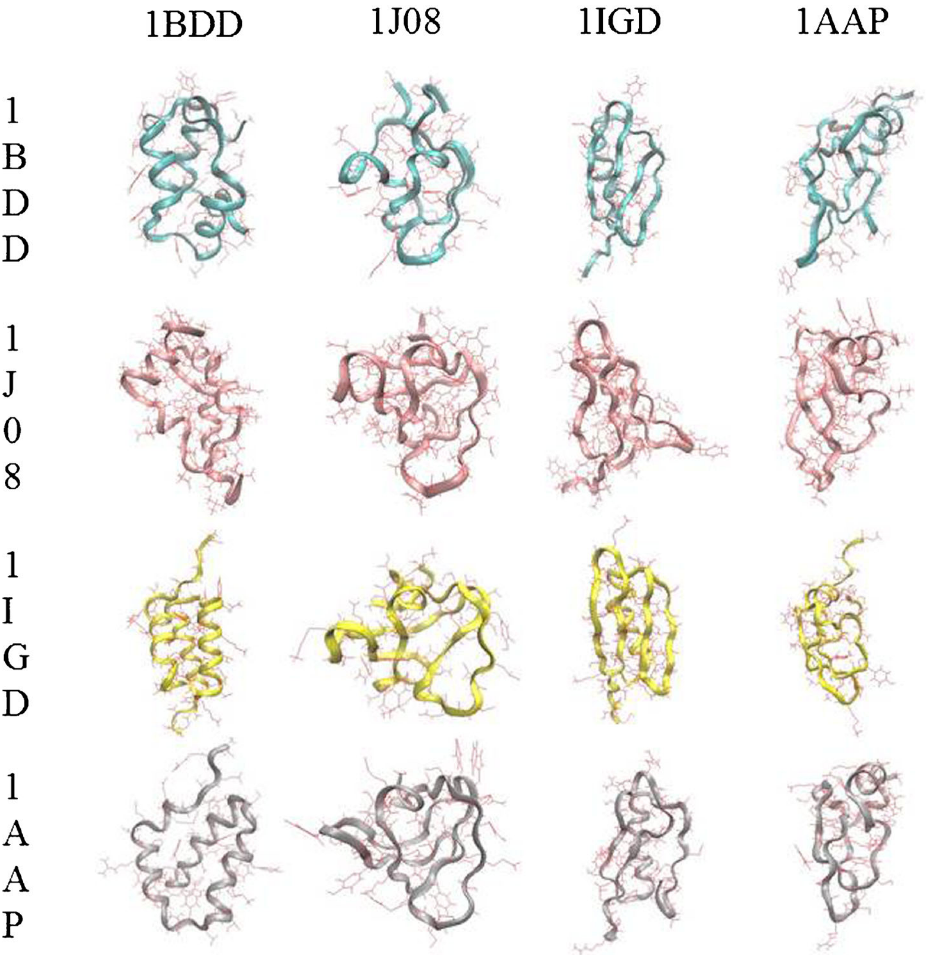

Fig. 1 Initial four main conformers of the four proteins, displayed in a matrix-like fashion. For each protein, the native states are displayed along the diagonal of this "matrix". The *label at the beginning of each row* identifies the protein and the *labels above each column* identify the protein whose fold was imposed on the non-native structures of the other proteins in that column. For instance, going down the first column we have first the native structure of protein 1BDD, and the next three are non-native structures obtained by imposing the backbone fold of 1BDD onto proteins 1J08, 1IGD, and 1AAP, respectively. The first row displays the native and three non-native structures for the mainly-α 1BDD protein (*all with a cyan backbone*); the second row displays the native and non-native structures for the mainly-β 1J08 protein (*all with a pink backbone*), the third row displays the native and non-native structures for the α/β 1IGD protein (*all with a yellow backbone*), and the fourth row displays the native and non-native structures for the few secondary structure 1AAP protein (*all with a grey backbone*). The side chains are represented by *red lines*. The protein pictures were prepared with the software Visual Molecular Dynamics (VMD) [22]

with the mainly-α fold of 1BDD has a similar protein–protein potential energy as the native 1IGD (compare the blue curve with the black in the third plot from the top). Figure 2 thus suggests that, energy-wise, there is nothing particularly special about native conformers, i.e., the alternative conformations which correspond to very artificial structures for those

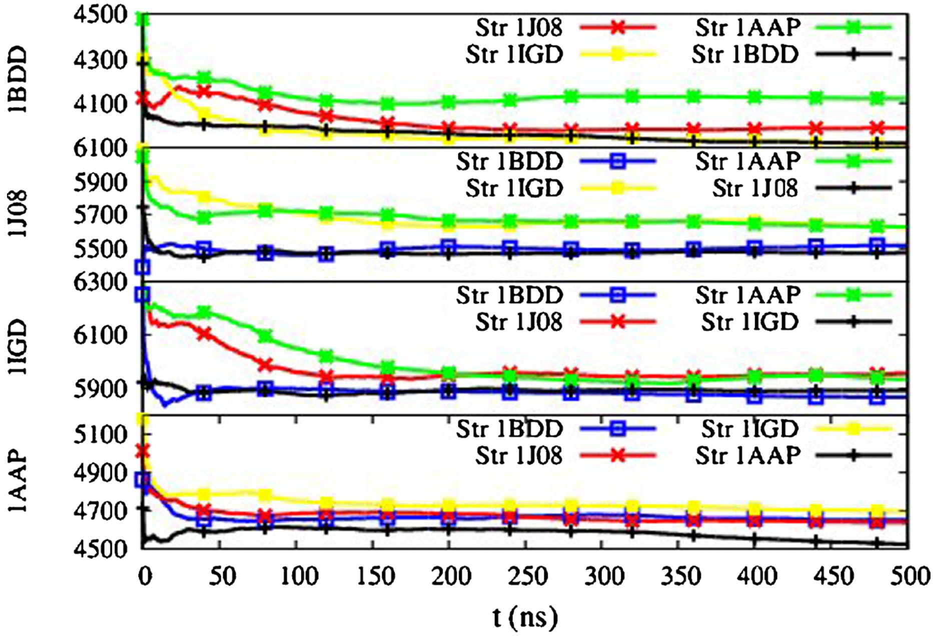

Fig. 2 Time dependence of the cumulative average of the potential energy when only interactions between protein atoms are taken into account. In each plot, the trajectories of the four main conformers of a given protein, i.e., starting from the native PDB structure and from the three non-native conformations for that protein (see text), were used. The order of the plots is the same as the order of the rows/proteins in Fig. 1. The *labels on the vertical axis of each plot* identify the protein, and the *labels within each plot* identify the initial structure. For instance, in the top plot, Str 1J08 means that the initial structure of the trajectory in red was obtained by imposing the backbone fold of protein 1J08 on protein 1BDD, i.e., that the initial structure for this trajectory was that seen in the first row, second column of Fig. 1. The curves for the trajectories that started from the native structures are in *black* in all plots. All energies are in kJ/mol

particular proteins nevertheless are stabilized to the same extent as the native state. However, while the protein–protein potential energy depends on the overall balance between all the attractive and repulsive interactions that stabilize each conformation of the four proteins, it may be that the non-native conformations are stabilized by one kind of interactions while the native state is stabilized by a different kind of interactions. Thus, in the next section the different contributions to the total potential energy, both from the internal interactions and from the interaction with the solvent, are analyzed in detail for the four main conformers of the four proteins.

3 Sources of stabilizing interactions in native and non-native conformations

In this section, we are concerned with one main question, namely, what types of interactions stabilize the different protein conformers? Figure 3 displays the variation with time of the different contributions that make up the potential energy related to protein–protein interactions. As for the total potential energy, the values of the energies can vary from protein to protein, but the difference between maximum and minimum values in the axes was kept the

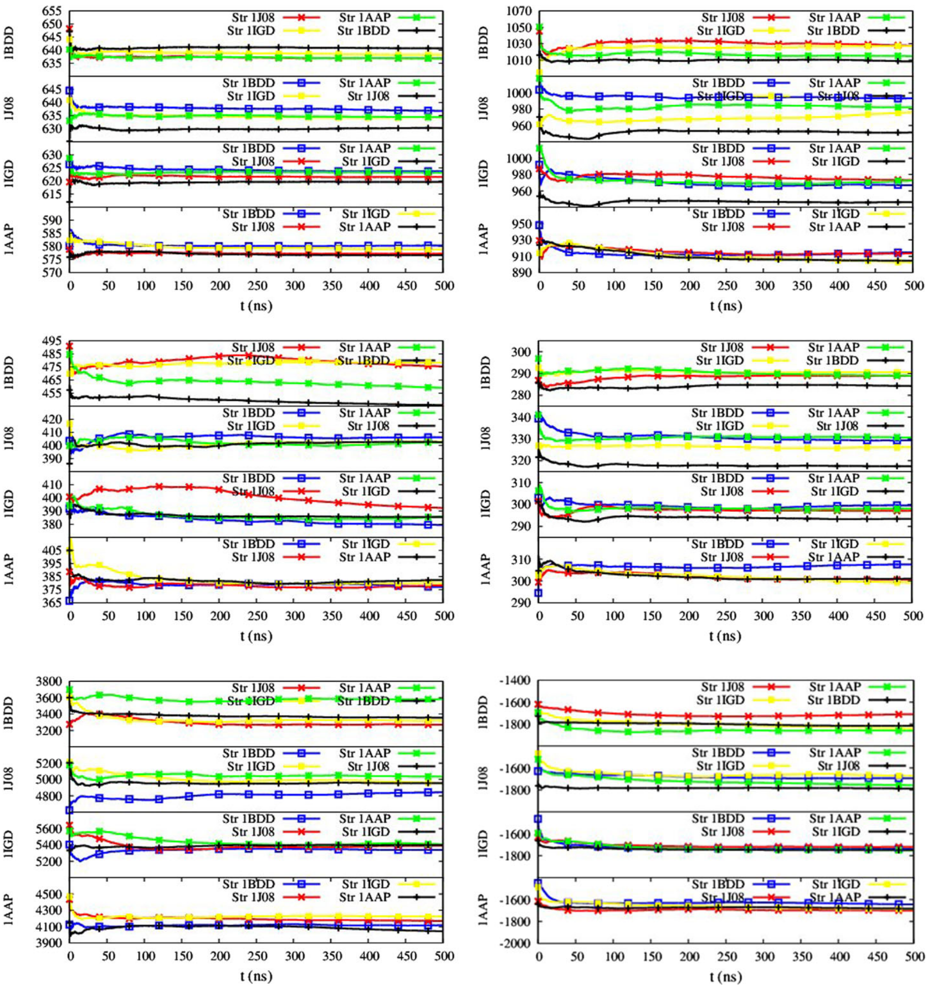

Fig. 3 Time dependence of the cumulative averages of bond (*top left*), angle (*top right*), proper dihedral (*middle left*), improper dihedral (*middle right*), Coulomb (*bottom left*), and Lennard–Jones (*bottom right*) energies of protein atoms only. The organization of each energy plot is as in Fig. 2. All energies are in kJ/mol

same for all proteins, to make the different plots more comparable. A close inspection of the curves indicates that all conformations of the four proteins are stabilized by the same type of interactions as the corresponding native states. Indeed, where some interactions, like the bond energy (top left), the improper dihedral (middle right) and Coulomb interactions (bottom right) are well conserved in the native, they are also well conserved in the non-native conformations, and conversely, the interactions that show greater changes in the non-native, also show similar characteristics for the native conformation. Figure 3 also shows that, for all proteins, the weakest contribution to the total interaction energy comes from the dihedral terms (improper and proper), which are followed, in an increasing sense, by the bond and angle energies. On the other hand, the strongest contribution is electrostatic (see bottom left plot of Fig. 3), followed, in absolute terms, by the Lennard–Jones interaction (in the bottom right plot of Fig. 3). Remembering that the non-native structures are stabilized by amino acid pairs that are very different from those that stabilize the corresponding native structures, this shows, that in spite of the local differences, when we consider the complete protein, the contribution of the different types of interactions to the total potential energy is very similar for the four different conformers of the same protein.

Thus, taken together, Fig. 3 *shows that the interactions that stabilize/destabilize the 16 native and non-native conformers are very similar in all of them.*

Hydrogen bonds are one of the key interactions that define protein structure and the Lennard–Jones potential energy, which represents them, is thus one of the most important contributions. In Fig. 4 we dissect further the Lennard–Jones contribution by plotting the average Lennard–Jones interaction energies between all the atoms that are represented in the united atom model of GROMOS96 43a1 [20] are displayed. The averages are made over at least 2500 conformations, sampled with a frequency of 0.2 ns in a 500-ns-long trajectory. The stronger the attractive (negative) interaction between two residues, the darker the dots in the triangles of Fig. 4, and the weaker the interaction, the more to the red the corresponding dots will be. Of the six types of interactions displayed in Fig. 3, we have selected the

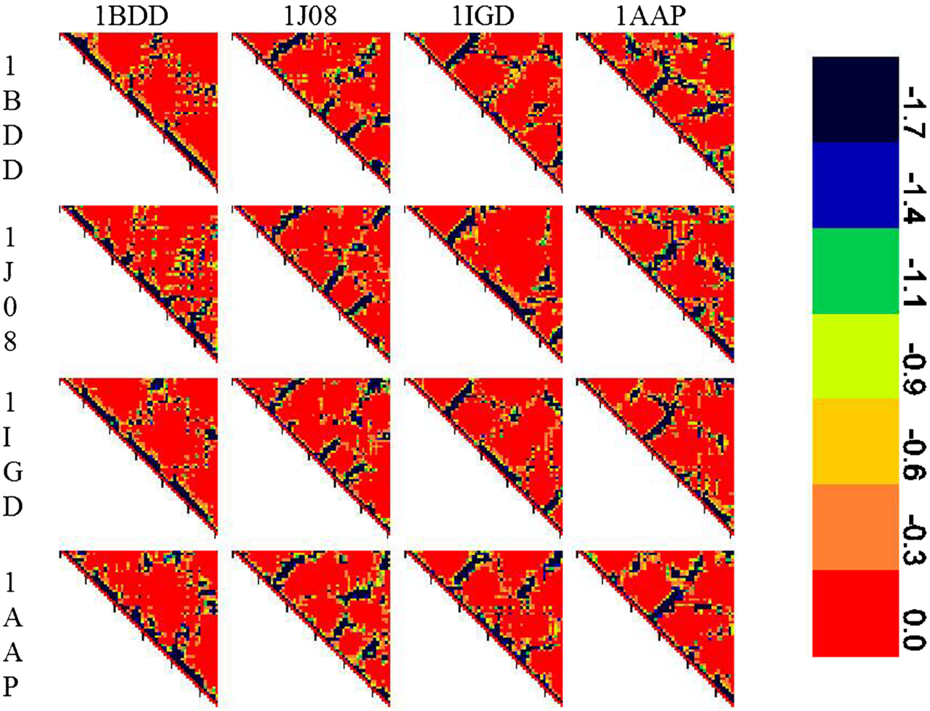

Fig. 4 Average Lennard–Jones interaction energies between all the atoms of the proteins, residue by residue (see text). Each *triangle* is obtained from a trajectory starting from one the 16 main conformers (see Fig. 1) and shows the average Lennard–Jones interaction between residues i and j, where both indices go from 1 to the total number of amino acids in the protein structure. Since this corresponds to a symmetric matrix, only the upper part is shown. The darker a spot in a triangle is, the greater (in intensity) the interaction between the corresponding residues is. The organization of the triangles is the same as in Fig. 1, i.e., the *label in each line identifies the protein* and *the label of each column identifies the structure imposed on the initial conformation*. The energy scale is given on the right-hand side, where the values are in kJ/mol. The *black sticks* along the diagonal mark the residue number in multiples of 10, starting with the first amino acid in the primary sequence

Lennard–Jones interaction also because it is that which is mostly correlated with the secondary and tertiary structure of the proteins, with α-helices leading to dark stripes along the third and fourth parallels to the diagonal of the triangles, and β-sheets showing up as dark stripes perpendicular to the diagonal. Both of these features are of course related to the hydrogen bonded networks that stabilize those two secondary structures. For instance, we can identify the three α-helices of protein 1BDD by the three dark patches, one larger, one smaller, followed by another larger one, parallel to the hypotenuse of the triangle in the first row, first column, of Fig. 4. Also, we can identify the β-sheets in protein 1J08 by the dark stripes perpendicular to the hypotenuse of the triangle of the second row, second column, of Fig. 4. *This figure thus provides a glimpse into the secondary structure and into the topology of the different conformers*. It tells us how well, or how not so well, the non-native conformers of the other proteins retain the backbone fold that was imposed initially on them. For instance, the three 1BDD helices are clearly visible in the non-native conformer of protein 1IGD (third row, first column, of Fig. 4), but less so in the non-native conformer of proteins 1J08 (second row, first column) and 1AAP (fourth row, first column). On the other hand, inspection of the triangles under the second column shows that the β-sheet structure of protein 1J08 is reasonably preserved by the three non-native conformers of the other proteins (first, third and fourth rows, first column). Also, of the four proteins, protein 1IGD is that whose non-native conformers resemble the corresponding native templates the most (compare the residue interactions of the triangles in the third row of Fig. 4 with the interactions of the corresponding native templates in the other rows). However, even if protein 1IGD seems to have the greater physical plasticity of the four proteins, i.e., it seems to be particularly capable of keeping very foreign folds, Fig. 4 does indicate that *all of the four different amino acid sequences of the four proteins can be made to fold into very different shapes, which are stabilized by interactions very similar to those that stabilize the corresponding native states*. The general conclusion is thus that, from the energetic point of view, the non-native conformers cannot be distinguished from their native states.

The energies presented thus far do not include the interaction of protein with the solvent, but protein–water interactions are known to be very important for the stability of the folded state. In Fig. 5, the total potential energy for the protein–protein and protein–solvent interactions is displayed. Comparing Figs. 2 and 5 we notice that, in the latter, the native conformers are no longer those that have the lowest energy values. In fact, two of the proteins (1BDD and 1IGD) possess one conformer that has a total potential energy that is lower than the native, another protein (1J08) has two such conformers and for the protein 1AAP all three alternative conformers have total energies either similar or lower than the native! Thus, taking the protein–solvent interactions into account, we still conclude that the same protein can have more than one conformer that is very different from its native structure and yet as stable as the native.

4 Flexibility of native and non-native conformers

So far, we have concluded that not only the average energies and the type of interactions that stabilize the non-native conformers but also their average stability are very similar to those of the native states. In this section, we re-visit the topic of the relative flexibilities of non-native versus native states [8].

We use the root mean square deviation (RMSD) between two conformations of the same protein as a measure of the structural "distance" between them. Here we want to determine

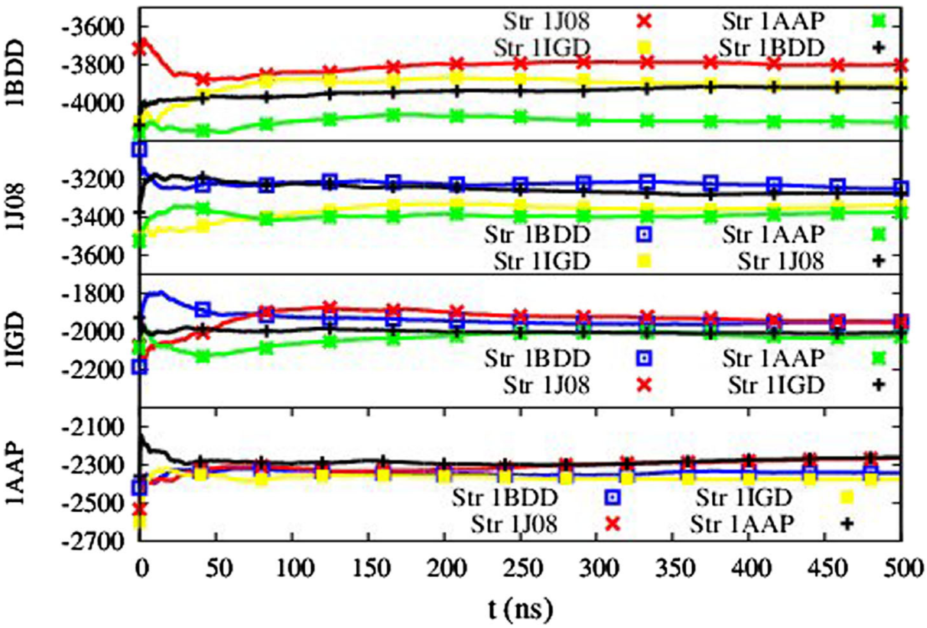

Fig. 5 Time dependence of the cumulative average of total potential energy, i.e., including not only interactions between all the atoms in the protein but also interactions between protein atoms and the solvent. The organization of each plot is as in Fig. 2. All energies are in kJ/mol

the structural distances between the conformations sampled in the trajectories of all the conformers. While Fig. 4 shows that, on average, the non-native conformers keep the fold that was initially imposed on them, one curious observation in [8] was that they also displayed a greater flexibility than the corresponding native state, as measured by RMSD. This finding is illustrated by Fig. 6. Each square in Fig. 6 is associated to a single trajectory, namely that which started from the initial structure that is the same position in Fig. 1. In each trajectory we selected 250 conformations, separated by 2 nanoseconds (ns), thus spanning the full 500-ns time period. In order to evaluate the conformational space covered in a trajectory we calculate the RMSD deviation of each conformation in the trajectory with all the other conformations in that trajectory. This leads to a symmetric matrix because the RMSD of conformation i ($i = 1, \cdots, 250$) with respect to conformation j ($j = 1, \cdots, 250$) is of course equal to the RMSD of conformation j with respect to conformation i (notice that indeed all squares in Fig. 6 are symmetric; we could also have plotted only the upper half but, unlike Fig. 4, that would make Fig. 6 less clear).

The first line in each of the 16 squares of Fig. 6 is the deviation of each conformation in the trajectory with respect to the initial conformation; the second line is the deviation of each conformation in the trajectory with respect to the second conformation; and so forth. The diagonal in each of the 16 squares of Fig. 6 represents the RMSD of conformation i ($i = 1, \cdots, 250$) with respect to the same conformation i and is of course zero (and thus represented by a black dot). The first parallel to the diagonal represents the RMSD deviations between one conformation and the next one in time (i.e., the RMSD between two conformations separated by 2 ns), the second parallel to the diagonal represents the RMSD between one conformation and its second neighbor in time (i.e., the RMSD between

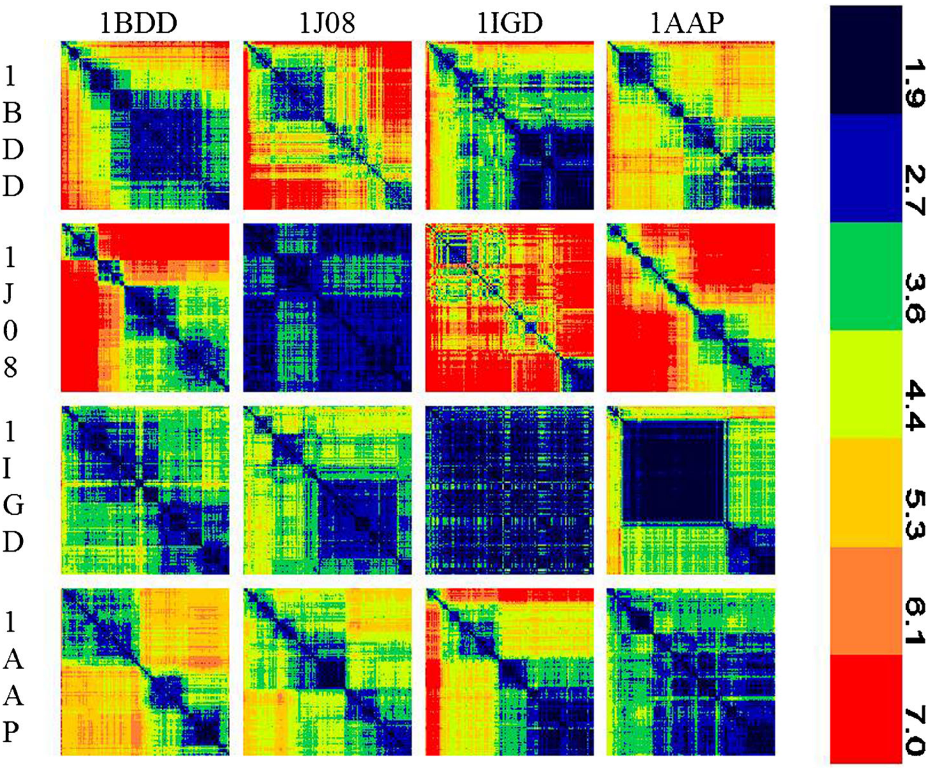

Fig. 6 Each of the 16 squares represents the RMSD of all 250 conformations selected from a single trajectory with respect to one another, i.e., the *first line in each square* represents the RMSD between all conformations in a trajectory with respect to the first conformation in that trajectory; the *second line* represents the RMSD of all conformations with respect the second conformation in that trajectory; and so on. The organization of the squares is the same as that of the triangles in Figs. 1 and 4. The RMSD scale is given on the right-hand side, where the values are in Å. The smaller the RMSD, the more to the blue a point in a square is, and the more structurally similar the two structures being compared are, and the larger the RMSD, and the red color is proportional to the structural difference between the conformations

two conformations separated by 4 ns), and so forth. The value of each RMSD is given by its color and the color scale is on the right-hand side of the figure, in Å. The more to the blue a spot is, the more similar the two structures being compared are and the more red a spot is, the greater the differences between the two structures being compared. In Fig. 6, the more blue a whole square is, the closer in structure are the conformations assumed by the protein throughout the trajectory and the more rigid the corresponding protein structure is. On the other hand, the more red a whole square is, the more flexible that other protein structure is.

The first line in each square tells us how much the protein structure has deviated from the initial structure that is displayed in Fig. 1. Looking at the squares along the diagonal of Fig. 6, which come from the trajectories that started from the native PDB structures of the four proteins, we see that, with the exception of protein 1BDD, the conformations sampled by the proteins in these trajectories have deviated at most 4 Å from the initial structure. This shows that the average initial structure is kept throughout, as is expected from the stability of native states. In each square of Fig. 6, the darker square patches centered along the diagonal, represent subsets of conformations that are similar to each other in that time interval, and signal structural sub-clusters that arise during that trajectory. A square with a smaller number of structural sub-clusters, and with smaller RMSD's between the conformations in the sub-clusters, that is, with a small number of darker patches centered along the diagonal, is associated with a less flexible structure. Inspection of Fig. 6 shows that the native state of proteins 1IGD and 1J08 are the least flexible, as expected for structures with β-sheets, and that the native state of protein 1BDD is very flexible. On the other hand, the native state of protein 1AAP has a degree of flexibility that is intermediate between those two. Also, comparing all native trajectories with all non-native trajectories we notice that, with the exception of protein 1BDD, the latter are always more flexible than the former.

In order to understand better the source of the greater flexibility of the non-native structures we generated the native LEAP structure, as explained in Section 2. The native LEAP structures have the same backbone as the native PDB structures that we have been considering so far, but they also share with the non-native initial structures the fact that the coordinates of the side chain atoms are added with the Leap module of AMBER. Therefore, while the PDB structures are obtained from proteins at equilibrium, the LEAP structures mix an experimental backbone structure with side chain orientations that may not be fully equilibrated. In this way, they are closer to the initial non-native structures for which neither the backbone, nor the side chains come from equilibrated structures (see Section 2). We wish to know how the flexibility of the conformations covered in the LEAP trajectories compares with that of the native and non-native trajectories.

Figure 7 plots the RMSD time profile for the native and LEAP structures and the cross RMSD's between the native and LEAP structures. The squares in the first column in Fig. 7 are the same as the diagonal squares in Fig. 6. The squares in the second column of Fig. 7 are also RMSD deviations of the conformations covered in a single trajectory but, in this case, the trajectories are those that started from the LEAP native structures. On the other hand, the four squares in the third column are obtained by using *two* different trajectories. Indeed, each spot $(i, j = 1, \cdots, N)$ in each square of the CROSS column represents the RMSD between conformation i of the native LEAP trajectory of one protein with respect to the conformation j of the native PDB trajectory of that protein. Notice that the squares under the CROSS column are *not* symmetric because spot (j, i) in each square of that column is the RMSD between conformation j of the native LEAP trajectory of one protein with respect to the conformation i of the native PDB trajectory of that protein, which is different from the RMSD plotted in spot (i, j). More specifically, the first line in each square of column CROSS is the RMSD of all the conformations sampled in the LEAP trajectory with respect to the first conformation in the PDB trajectory; the second line is the RMSD of all the conformations sampled in the LEAP trajectory with respect to the second conformation in the PDB trajectory; and so on. In fact, while the squares under the two first columns of Fig. 7 represent the structural overlap of conformations spanned within a single trajectory, *the squares in the CROSS column represent the structural overlap between the conformations covered in the PDB trajectories and the conformations covered in the LEAP trajectories*.

Figure 7 shows that the conformations in the LEAP trajectories tend to deviate more from the initial one than the conformations in the PDB trajectories. This is clearly seen for protein 1AAP, whose LEAP RMSD's is visibly more to the red than the PDB (full native) square, but it is also true for the protein 1J08 (notice that the first line of the RMSD square in the second row, second column of Fig. 7 shows that the conformations of the LEAP 1J08 deviate consistently more than 2.7 Å from the initial structure than the PDB structure; also the dark patch centered on the diagonal, at the bottom shows that the protein stabilizes in

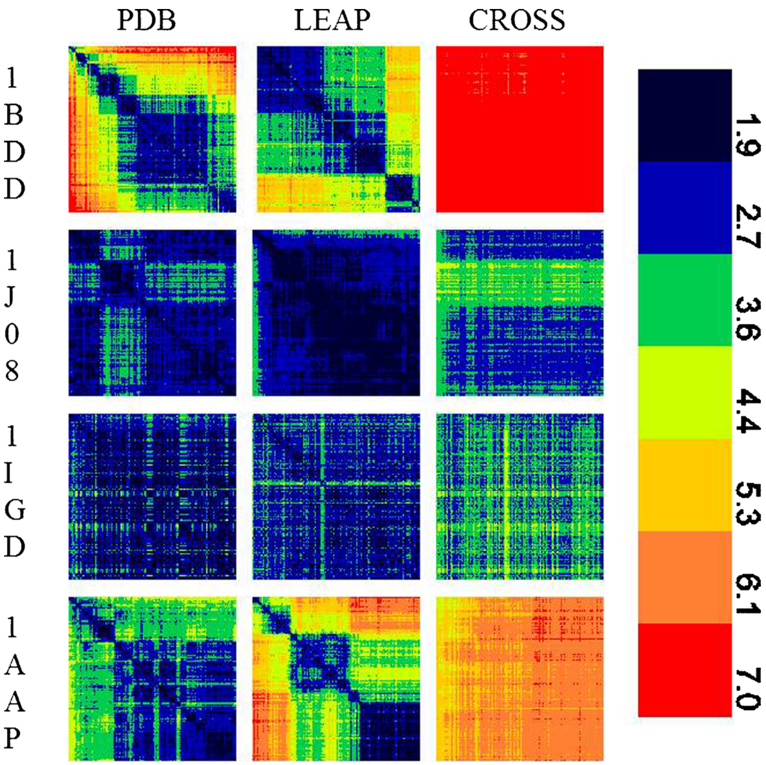

Fig. 7 The *squares* in columns PDB and LEAP are the RMSD time profiles for the trajectories that start from the native and the LEAP structures, respectively (the *squares* under PDB are thus the same as in the diagonal of Fig. 6). On the other hand, the *squares under CROSS* represent the structural overlap between the conformations sampled in the PDB trajectory and those sampled in the LEAP trajectory of the same protein (see text). The proteins are identified by the label at the beginning of each row. The RMSD scale is the same as in Fig. 6 and is repeated at the right with values in Å

a different structural sub-cluster). The LEAP conformation of protein 1IGD, whose native structure is the least flexible of all native structures (compare the squares in the diagonal of Fig. 6), is also generally more flexible than its PDB (full native) conformation, although less markedly so, and the only exception is the 1BDD protein, whose LEAP conformation appears *less* flexible than its full native conformation. We thus conclude that, in general, the native LEAP structures tend to be more flexible the native PDB structures.

In the CROSS column, the structural distance between the conformations sampled in the LEAP trajectories and in the PDB trajectories is evaluated. The greater the difference between the conformations adopted by the same protein in the PDB and LEAP trajectories, the more to the red the RMSD squares in the CROSS column will be. A comparison of the last column of Fig. 7 with the two first ones shows that, generally, the structures sampled in each trajectory are structurally closer to one another than the structures sampled in the different trajectories. This is particularly so for the mainly-α protein 1BDD whose bright red color means that *all* conformations sampled in the LEAP trajectory are more than 7.0 Å away from *all* the conformations sampled in the PDB trajectory. It is also true for few secondary structures protein 1AAP, but also affects the α/β 1IGD protein and the mainly-β 1J08 protein, which, like all proteins with β-sheets, tend to be more rigid. Our simulations show that such structural differences have not been eliminated after 500 ns. This means that just changing the initial orientations of the residues can lead to changes in the backbone fold of a protein and move it away from the full native fold.

The general conclusion is that changing the initial orientations of the residues can explain the enhanced flexibility of the non-native structures with respect to the PDB native structures in two ways. First, less than optimum orientations for a given structure will tend to destabilize that structure. From the results in Fig. 7, this has happened to the LEAP trajectories of proteins 1AAP and 1J08. Secondly, the initial orientation of residues can induce local changes in the backbone fold of the protein and thus correspond to a structure that starts further away from a local free energy minimum. This happened to the LEAP trajectories of all four proteins, but particularly so to proteins 1BDD and 1AAP.

5 Discussion

Since 1994, the successive critical assessments of protein structure prediction (CASP) experiments have provided ample evidence that the same protein can assume many different structures, all with similar potential energies [23]. This is the main reason for the difficulty in applying the thermodynamic hypothesis [1] to determine the three-dimensional structure of proteins from their sequence alone. In previous MD studies [7, 8], we have shown that the different non-native conformations a single protein may assume can also have stabilities that are very similar to that of the native state. There are also cases in which direct experimental evidence for a kinetic control of protein folding has been found [24–26]. The aim here was to investigate the potentiality of two other features, namely, the nature of the stabilizing interactions of the different conformations and their degree of flexibility, in facilitating the identification of the native structure from among a set of different structures of the same protein. To that end, our analysis has centered on the energetics of the different structures. Specifically, we wanted to ascertain whether the very alien non-native states built for this study are stabilized by interactions that are essentially different from those that stabilize the native state. In Section 3, we have looked into the nature of the stabilizing interactions of non-native structures and the general conclusion from Figs. 2, 3 and 5 is that, in spite of their radically different structures, the non-native conformers of the four proteins selected are stabilized by the same type of interactions as the corresponding native states. Furthermore, Fig. 5 shows that all the four proteins used here possess non-native conformers that have a total potential energy that is lower than that of the native state, as is found in the CASP experiments [23]. This means that, also from the point of view of the nature of their stabilizing interactions, when we consider the full proteins, at least, the non-native structures are *indistinguishable* from the native states. In [7, 8] we had already excluded other criteria, such as the stability, to differentiate between native and non-native states of the same protein and here we exclude this one as well.

In Section 4 we considered another possible distinguishing feature apparently identified in a previous study [8], namely, an enhanced degree of fluctuation of the non-native structures with respect to that of the native state. For that, we compared the trajectories obtained when starting from two different native structures. One, designated as the PDB trajectory, whose starting structure made use of all the PDB coordinates and a second trajectory, designated as the LEAP trajectory, whose starting structure had the same backbone coordinates as for the PDB trajectory but with different coordinates for the side-chain atoms (see Section 2). Figure 7 shows that mis-orientations of the side-chains can make the resulting native LEAP conformers display a greater flexibility with respect to the native PDB conformers. Our results also indicate that the orientations of the residues are important for the definition of the final structure and that finding the fully relaxed residue orientations in an equally fully relaxed backbone fold may be a rate-limiting step in folding. One open question is if these simulations were continued for much longer, e.g., for milliseconds or seconds, in order to provide them with the ample time that the structures obtained experimentally have to relax their backbone fold and their side-chain orientations, would the non-native states of these proteins have a degree of flexibility similar to that of the corresponding PDB native states? The results in Section 4 suggest that the answer may be yes and that the apparent lower degree of fluctuation of the conformations sampled in the PDB trajectories may be due to that they start closer to a local free-energy minimum. In this case, *the lower degree of flexibility cannot be used as a distinguishing feature of the native state itself.*

In summary, our conclusion is that neither the total potential energy, nor the degree of flexibility, nor the nature of the stabilizing interactions can be used to distinguish the native state from possible non-native structures of the same protein. In fact, if we were to pick an unknown native structure out of the four conformers considered here, using only the information from our simulation data, we would only have a 25% chance of selecting the right structure, because each conformer would be equally probable, and, of course, this probability will decrease as the number of structures we consider increases. Thus, on the whole, our results provide support to Levinthal's suggestion [5] that the native state is just one of the many kinetically accessible structures each protein can have.

It may be argued that we base our conclusion above on simulations with just four proteins and that the results might be different if we had used different proteins. There may well be proteins that will only be stable in the native conformation. Such proteins should be particularly easy to fold in a computer, thus, the difficulty in folding proteins in the CASP exercises [23] suggests that, if they exist, they are not very common. The four proteins we have chosen are globular proteins with stable native states (something that the MDS which start with the PDB structures confirm). Furthermore, each of the proteins belongs to one of the four structural CATH classes [9] and thus, although just four, they represent the entire set of structural classes identified in the protein data bank [10]. For each of these four proteins, we have built three non-native conformers, as shown in Fig. 1. These non-native conformers are as structurally different from the native state as is conceivable since, for each, a given protein assumes the fold of a foreign class of proteins. If such alien protein structures can be stable it is reasonable to assume that there exist many other structures, closer to these native and non-native ones, which will also be stable. Thus, in spite of the small number of proteins used, we think the protocol we have followed provides our results with a certain degree of generality. The probability, that just these four proteins will present a special behavior that would lead to our conclusions, can be considered close to zero reinforcing our conclusions.

It may also be said that, although the four proteins selected cover the four existing structural classes, they do not cover many of the known folds and that they are very small, but whatever their shape or size, all proteins share one feature: their three-dimensional structure is not stabilized by a few very strong interactions (as in a solid or in small molecules where the strong interaction is the covalent bond) but by many weak interactions. The consequence is that each protein can have many shapes all with the same global sum of attractive and repulsive interactions. This is the reason why the four different conformers of the each of the four proteins studied have similar overall energetics. Therefore our prediction is that in further simulations, with larger and more varied proteins, the degeneracy problem that we have identified in the four small proteins selected will be much greater. i.e., larger proteins have much larger conformational spaces and thus many more conformations that have globally the same energetics as the native conformation. The fact that, in cells, the folding of large proteins is often aided by chaperones corroborates this prediction.

One other question that may be asked is: if the same protein can assume many different average structures that are all stable, as we assert, how is it that, in cells, most proteins always assume the same average structure, known as the native state? Anfinsen's thermodynamic hypothesis [1] and the funnel model [3, 4] according to which folding is an equilibrium process in which a protein's free energy is minimized, cannot explain this. Or better, within Anfinsen's thermodynamic hypothesis each protein should only have one stable state, which is its native state. On the other hand, within Levinthal's kinetic hypothesis [5] this is readily explained. Indeed, if the structure a protein has as it leaves the ribosome (the structure of the nascent chain) is always the same, and if the pathway the protein follows after being synthesized is also always the same, this protein will always reach the same final structure, no matter how many other stable structures it may possess in the same thermodynamic conditions. In previous publications [27, 28], it has been suggested that the nascent chain is always helical and that the transient (kinetic, deterministic) forces that define the pathway come from vibrational excited states (the VES hypothesis). In short, from the perspective of a kinetic mechanism for folding, the native state is just the structure that is more kinetically accessible in the normal cellular environment. In this way, all the other protein structures, even if they are more stable than the native, will not arise.

Following Anfinsen's thermodynamic hypothesis, protein folding is thought to be a process in which the protein, whatever its initial structure, will reach the native state by progressively minimizing its energy. This is the strategy that is often used to determine the native state in a computer. On the other hand, the results here and in previous works [7, 8] suggest that this strategy cannot succeed. Indeed, such a strategy presumes that to each amino acid sequence there corresponds only one well-defined minimum energy structure, while our results indicate that there can be many more than one. As explained in the previous paragraph, an alternative strategy to obtain the three-dimensional structure that a protein assumes in a cell (the native structure), is to determine first, the structure of the nascent chain and secondly, the changes that this initial structure suffers (the pathway). From the point of view of the Levinthal limit, the fact that, in cells, the native states are usually well defined means that those two questions must have a unique answer.

Acknowledgments LC received national funds from FCT - Foundation for Science and Technology, Portugal, through the project UID/Multi/04326/2013. LD thanks the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), and to the Conselho Nacional de Desenvolvimento Cientía co e Tecnológico (CNPq) for financial support.

References

1. Anfinsen, C.: Principles that govern the folding of protein chains. Science **181**(4096), 223–230 (1973)
2. Levinthal, C.: How to Fold Graciously. In: Debrunner, J.T.P., Munck, E. (eds.) Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, Monticello, Illinois, Vol. 22, pp. 22–24. University of Illinois Press (1969)
3. Bryngelson, J.D., Wolynes, P.G.: Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. U. S. A. **84**(21), 7524–7528 (1987). doi:10.1073/pnas.84.21.7524
4. Dill, K., Chan, H.S.: From Levinthal to pathways to funnels. Nature Struct. Biol. **4**, 10–19 (1997)
5. Levinthal, C.: Are there pathways for protein folding? J. Chim. Phys. **65**, 44–45 (1968)
6. Lazaridis, T., Karplus, M.: "New view" of protein folding reconciled with the old through multiple unfolding simulations. Science **278**, 1928–1931 (1997)

7. Cruzeiro, L., Lopes, P.A.: Are the native states of proteins kinetic traps? Mol. Phys. **107**(14), 1485–1493 (2009)
8. Cruzeiro, L., Degrève, L.: What is the shape of the distribution of protein conformations at equilibrium? J. Biomol. Struct. Dyn. **33**(7), 1539–1546 (2015). doi:10.1080/07391102.2014.966148. http://www.ncbi.nlm.nih.gov/pubmed/25229986
9. Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., Thornton, J.: Cath- a hierarchic classification of protein domain structures. Structure **5**, 1093–1108 (1997)
10. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The protein data bank. Nuc. Acid. Res. **28**, 235–242 (2000)
11. Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y., Shimada, I.: Three-dimensional solution structure of the b domain of staphylococcal protein a: Comparisons of the solution and crystal structures. Biochemistry **31**, 9665–9672 (1992)
12. Fazi, B., Cope, M., Douangamath, A., Ferracuti, S., Schirwitz, K., Zucconi, A., DG, D., Wilmanns, M., Cesareni, G., Castagnoli, L.: Unusual binding properties of the SH3 domain of the yeast actin-binding protein Abp1: Structural and functional analysis. J. Biol. Chem. **277**, 5290–5298 (2002)
13. Gallagher, T., Alexander, P., Bryan, P., Gillilan, G.: Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. Biochemistry **33**, 4721–4729 (1994)
14. Hynes, T.R., Randal, M., Kennedy, L.A., Eigenbrot, C., Kossiakoff, A.A.: X-ray crystal structure of the protease inhibitor domain of Alzheimer's amyloid beta-protein precursor. Biochemistry **29**, 10,018–10,022 (1990)
15. Case, D., Cheatham, T.I., Darden, T., Gohlke, H., Luo, R., Merz, K.J., Onufriev, A., Simmerling, C., Wang, B., Woods, R.: The Amber biomolecular simulation programs. J. Computat. Chem. **26**(16), 1668–1688 (2005)
16. Lindahl, E., Hess, B., van der Spoel, D.: Gromacs 3.0: a package for molecular simulation and trajectory analysis. J. Mol. Mod. **7**, 306–306 (2001)
17. Apol, E., Apostolov, R., Berendsen, H., van Buuren, A., Bjelkmar, P., van Drunen, R., Feenstra, A., Groenhof, G., Kasson, P., Larsson, P., Meulenhoff, P., Murtola, T., Pll, S., Pronk, S., Schulz, R., Shirts, M., Sijbers, A., Tieleman, P., Hess, B., van der Spoel, D., Lindahl, E.: Gromacs user manual, version 4.5 (2010). www.gromacs.org
18. Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E.: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. **4**, 435–435 (2008)
19. van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J.C.: Gromacs: fast, flexible, and free. J. Comp. Chem. **26**, 1701–1701 (2005)
20. van Gunsteren, W., Mark, A.: Validation of molecular dynamics simulation. J. Chem. Phys. **108**, 6109–6116 (1998)
21. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., Hermans, J.: Interaction models for water in relation to protein hydration. In: Intermolecular Forces (pp. 331-342), Vol. 14, The Jerusalem Symposia on Quantum Chemistry and Biochemistry (Ed. B. Pullman). Reidel, Dordrecht, The Netherlands (1981)
22. Humphrey, W., Dalke, A., Schulten, K.: Vmd: visual molecular dynamics. J. Mol. Graphics **14**, 33–38 (1996)
23. CASP: Critical Assessment of Protein Structure Prediction. In: Predictioncenter.org (2015)
24. Baker, D., Sohl, J., Agard, D.: A protein-folding reaction under kinetic control. Nature **356**, 263–265 (1992)
25. Gettins, P.: Serpin structure, mechanism, and function. Chem. Rev. **102**, 4751–4803 (2002)
26. Sohl, J., Jaswal, S., Agard, D.: Unfolded conformations of alpha-lytic protease are more stable than its native state. Nature **395**, 817–819 (1998)
27. Cruzeiro, L.: Protein folding. In: Springborg, M. (ed.) Chemical Modelling, pp. 89–114. Royal Society of Chemistry, London, UK (2010)
28. Cruzeiro, L.: The VES hypothesis and protein conformational changes. Z. Phys. Chem. **230**, 743–776 (2016)