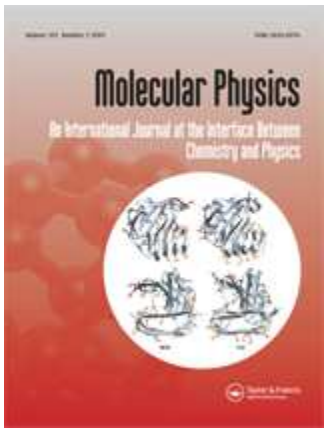
# Are the native states of proteins kinetic traps?

Leonor Cruzeiro, Paulo Afonso Lopes




# Are the native states of proteins kinetic traps?

| | |
|---|---|
| Journal: | *Molecular Physics* |
| Manuscript ID: | TMPH-2008-0190.R1 |
| Manuscript Type: | Full Paper |
| Date Submitted by the Author: | 27-Feb-2009 |
| Complete List of Authors: | Cruzeiro, Leonor; University of Algarve, Physics<br>Lopes, Paulo; Universidade Nova de Lisboa, Informática |
| Keywords: | decoys, Protein folding, Energy landscape, Kinetic mechanism, VES Hypothesis |
| | |
| Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online. | |
| paper3.tex | |

# Are the native states of proteins kinetic traps?

Leonor Cruzeiro

*CCMAR and FCT,*

*Universidade do Algarve,*

*8005-139 Faro, Portugal*[*]

Paulo A. Lopes

*CITI, Departamento de Informática,*

*Faculdade de Ciências e Tecnologia, FCT,*

*Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*[†]

(Dated: February 27, 2009)

## Abstract

Four proteins are selected to represent each of the four different CATH classes and, for each protein, three decoys are built with structures that are totally alien to the native state. The decoys are scored against the native state with the help of the AMBER force field, using three measures: the average energy, the average fluctuation and the resistance to a heat pulse. **Two sets of simulations are performed, one with explicit solvent and another set with implicit solvent. The overall conclusion is that of these three measures that which is most successful in picking out the native states is the last one since the native structures take a consistently longer time to be destabilized in this manner. But the general conclusion is also that none of measures is completely effective in discriminating all the decoys, a result that supports other studies according to which the native state is reached by a kinetic step.**

PACS numbers: 87.15.Aa; 87.14.Ee; 87.15.He

Keywords: Decoys, Protein folding, Energy landscape, Kinetic mechanism

## I. INTRODUCTION

An outstanding question in Biology and Medicine, known as the protein folding problem, is how a given sequence of amino acids, in cells, most of the times assumes the native structure[1,2]. An important concept is that of the free energy landscape and the current working hypothesis is that this landscape is funnel-shaped[3–6] and that the native structure corresponds to its global minimum[2–6]. However, a question arises about the theoretical consistency between the funnel hypothesis and the interactions that stabilize protein structure. These interactions are reasonably well represented by potentials such as these[7]:

$$V = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \tag{1}$$
$$+ \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] +$$
$$+ \sum_{i<j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

where bond stretching and bond bending (the first two sums) are harmonic, rotations around a bond are described by a truncated Fourier series (third sum) and nonbonded interactions are modelled by the Lennard-Jones potential and Coulomb interactions due to the partial charges on each atom (the last sum). A few systematic studies of the shapes of the energy landscape of small polypeptides and water clusters using these kind of potentials have been attempted, which show both funnelled and multi-funnelled free energy landscapes[8–11], with the local topography of the energy landscape being related to the conformation of the molecule[9]. Furthermore, a 4 $\mu$s study of the free energy landscape of a 16 amino acid beta-hairpin in an implicit water bath led to three well defined non-native basins with free energies comparable to that of the native basin[12]. On the other hand, the conformational space of even small proteins, remains too large to be probed in a systematic manner, even with the most powerful computers. More recently, tests of force fields given by (1) have been made by evaluating their ability to select the native fold from among a number of decoys[13]. In the latter studies, the decoys are usually artificial protein conformations corresponding to local energy minima structurally similar to the native state[14]. Instead, here, decoys structurally very far from the native basin are prepared and the ability of an energy function such as (1) to select them out, when compared to the native structure, is investigated. To that end, three measures are consid-

ered, namely, the average energy, the degree of fluctuation and the resistance to a heat pulse.

## II. METHODS.

Four proteins are selected which, using the nomenclature of the Protein Data Bank (PDB)[15], are: 1QLX (104 amino acids)[16], 1I0S (161 amino acids)[17], 1AAP (56 amino acids)[18] and 1IGD (61 amino acids)[19]. These proteins have different sizes, different biological origins and different functions. While the first is a fragment of the human prion[16], the second is an oxireductase from archae[17], the third is the protease inhibitor domain of Alzheimer's amyloid $\beta$-protein[18] and the fourth is a immunoglobulin binding domain of streptococcal protein G[19]. The main criterion for their selection was to have one representative of each of the four main classes of proteins identified in the CATH hierarchical structural classification scheme[20]: mainly $\alpha$ (1QLX), mainly $\beta$ (1I0S), essentially structureless (1AAP) and $\alpha/\beta$ (1IGD). For each protein, three decoys are built by threading the native fold, or part of the native fold, of each of the other three proteins, onto the amino acid sequence of that protein, as explained in detail below.

The coordinates for the atoms in the native structures of the four proteins selected were taken from the PDB[15] and their structures were energy minimized with the AMBER ff99 force field[7], to relieve any steric or otherwise strongly unfavorable interactions. For each protein, decoy structures were then built by taking the energy minimized native structures as templates, and forcing the sequence, or part of the sequence, of each protein to have the backbone fold, or part of the backbone fold, of each of the other three proteins. For example, the initial coordinates for the structure in the first row, second column of figure 1 were obtained by imposing the backbone fold of the first 104 amino acids of 1I0s onto the backbone of the 104 amino acids of 1QLX and the initial coordinates for the second row, first column were obtained by imposing the backbone fold of the 104 amino acids of 1QLX onto the backbone of the first 104 amino acids of 1I0s. These and the other decoy structures thus generated were first relaxed, in order to eliminate all the steric interactions such a procedure leads to, and, after relaxation, they were energy minimized[21]. With this protocol, both the native structures and the decoys are initially in the vicinity of a local

minimum of the AMBER force field but the decoys have structures that are drastically different from those of the corresponding native state.

While implicit solvent models are often used in order to minimize the number of degrees of freedom in a protein-water system, it is generally acknowledged that a better representation of a solution environment is with explicit water molecules. Thus, in a first set of simulations (see section III A) the 16 energy minimized structures described above were all solvated in water using the box option of the leap program of AMBER[7] and the resulting systems were energy minimized, with the TIP3P water model and keeping the protein fixed. Then the NAMD program[22], with the AMBER ff99 and TIP3P force fields, was used to heat each of the systems to 298 K and to equilibrate them at that temperature for 0.6 ns. A representative statistical ensemble, at 298 K, for each of the 16 systems thus constructed, was obtained by storing snapshots every picosecond from a further 0.2 ns molecular dynamics (MD) run. This time interval is short but, on the other hand, it should be noted that the smallest system in these simulations has 12234 atoms and the largest has 33451 atoms. **Furthermore, to compensate for the shortness of the explicit water simulations, a second set of simulations was performed, with a duration of up to 50 nanoseconds and with an implicit water environment, using the AMBER ff99SB force field[24]. The initial structures for the implicit solvent simulations were obtained by energy minimization, with the AMBER ff99SB force field, of the final structures from the explicit water simulations, followed by a 2 nanosecond heating from 10 K to 298 K. The structures at the end of the heating procedure were then equilibrated at 298 K for up to 50 nanoseconds and snapshots were stored every 10 picoseconds.**

## III. RESULTS

### A. Explicit Water simulations

Figure 1, which was made with the program VMD[23], shows the native folds of the four proteins and also the twelve decoys, at the end of the MD sampling run. The native

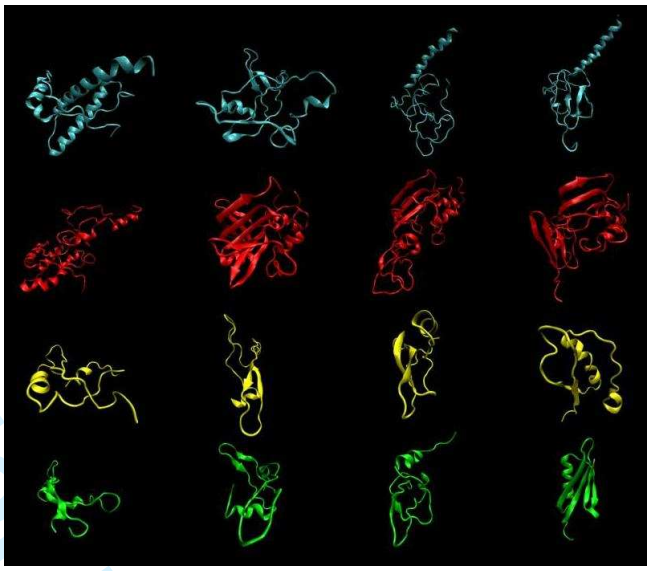


FIG. 1: (Colour online) Protein structures at the end of the 0.8 ns equilibration period at 298 K. All proteins in the same row (with same colour) have the same amino acid sequence. The four native structures are displayed along the diagonal. The first row has the structures for protein 1QLX (cyan), the second for 1I0S (red), the third for 1AAP (yellow) and the fourth is for 1IGD (green). Along each column, the decoy structures are obtained by threading the backbone fold, or part of the backbone fold, of the native structure in that column on to the backbone of the other proteins. This figure was made with VMD[23].

TABLE I: **RMSD with respect to native (Å)**

| **Seq** | **mainly** $\alpha$ **Str 1QLX** | **mainly** $\beta$ **Str 1I0S** | **disordered Str 1AAP** | $\alpha/\beta$ **Str 1IGD** |
|---|---|---|---|---|
| **1QLX** | **0** | **14.72** | **16.02** | **16.99** |
| **1I0S** | **19.91** | **0** | **15.36** | **12.61** |
| **1AAP** | **11.34** | **11.69** | **0** | **9.40** |
| **1IGD** | **15.50** | **13.39** | **11.42** | **0** |

structures for the four proteins are found along the diagonal of figure 1 and each row includes the native fold plus its three decoy structures, all in the same colour. All decoys in the same column were generated to have at least part of the fold of the native structure in

that column. Close inspection of figure 1 shows that, even after heating to, and equilibrating at, 298 K, the alternative structures retain the overall backbone folds that were imposed on them initially, even if these correspond to protein structures in a class that is very unnatural for the amino acids sequences concerned. A quantitative estimate of the structural distance between the decoys and the native structures in figure 1 can be gauged by the root mean square deviation (RMSD) per atom between the coordinates of the decoys, with respect to the corresponding native structure. Considering only the non-hydrogen backbone atoms in each structure this is given in table I. RMSDs per atom between two conformations of the same protein generally increase with the number of atoms, a rule that is statistically verified by the data in table I. Although the larger proteins (1QLX and 1I0S) in this study tend to have larger RMSDs than the smaller proteins (1AAP and 1I0s), there are some deviations from the latter rule because, with the protocol described above, in these larger proteins, part of the decoy structure is left as in the native fold, which is the reference for the values displayed in table I. On the whole, the RMSD values in table I are 2-4 times larger than those of the decoys usually used to test force fields, which have values of 4-6 Å with respect to the native structure[13,14].

If the native states of proteins correspond to well-defined global energy minima, a good force field should give free energies for the decoys that are smaller than the free energies of the corresponding native states. The data in tables II and III was calculated from the last 0.2 ns part of the MD simulation. The average energies of four native structures plus the 12 decoys are displayed in table II, in which the data is organized in the same manner as in figure 1. All systems in a given row of table II are exactly the same, i.e., not only do they have the same protein and ions but also the same number of water molecules, namely, the N water molecules closest to the protein. The number N was chosen as that for which the interaction energy between the protein and water reached an average saturation value. For the 1QLX and 1I0S proteins N is 4000 and for the smaller proteins 1AAP and 1IGD N is 2000. In each cell of table I, the first number is the total energy of the system constituted by the protein plus ions plus N water molecules and is dominated by the water-water interactions. The second number in each cell is the total energy of the protein, including the intra-protein interactions (third number), the protein-ion interactions (fourth number) and the protein-water interactions (fifth number). Inspection of table II shows that some of the decoys

have equivalent, or even lower, potential energies than the native structure. For example, imposing part of $\beta$-fold of 1I0S on the naturally mainly $\alpha$ structure of 1QLX leads to a decoy that has an average potential energy of $-6709$ kcal/mol, lower than the potential energy of the native structure of 1QLX which is $-6428$ kcal/mol, and imposing the essentially disordered structure of the native fold of 1AAP on to the first 56 amino acids of 1I0S leads to a decoy with an average energy of $-7437$ kcal/mol, approximately equal to that of the native fold of 1I0S, of $-7438$ kcal/mol. Furthermore, in the case of the protein 1AAP, all three decoys have energies that are lower than the native fold and in the case of 1IGD, two of the alternative structures have lower energies.

While some of the decoys have potential energies comparable to those of the native states, the important thermodynamic quantities are their relative free energies. The more open and the less compact a structure is, the greater its entropy will be, because the more it can fluctuate. Inspection of figure 1 already shows that the decoys used in this study are generally less compact than the corresponding native structure. In order to have a quantitative insight into the entropy associated with each structure, the root mean square deviations per atom of each of the structures with respect to its thermal equilibrium average structure are presented in table III (where only non-hydrogen backbone atoms have been used). As expected, the data indicate that the native structures fluctuate less than the alternative structures in all cases, something that reinforces the thermodynamic viability of the decoys. On the other hand, when we consider the average fluctuation of each conformation as a separate measure, we find that it constitutes the best measure in this study to discriminate the decoys against the native structures, as its value for the native states is, in all cases, smaller than that of the decoys, although the difference which varies from 0.1 to 0.4 Å, is in some cases within the uncertainty that goes from 0.12 and 0.20. The decoy with the closest value to its native is that in which the native fold of 1AAP is threaded on the first 56 amino acids of 1I0S (second row, third column in table III).

To partially compensate for the shortness of the simulation and as a further qualitative measure of the relative structural stability of the decoys with respect to the native states, all systems were heated from the equilibrium value of 298 K to a final value of 698 K, at the rate of 2 K per ps. Figure 2 shows the variation of the RMSD per atom of each

TABLE II: **Average Potential Energies (kcal/mol)**

| Seq | mainly $\alpha$ Str 1QLX | mainly $\beta$ Str 1I0S | disordered Str 1AAP | $\alpha/\beta$ Str 1IGD |
|---|---|---|---|---|
| 1 | -38773[a] ± 150 | -38067[a] ± 168 | -38650[a] ± 140 | -38224[a] ± 148 |
| Q | -6428[b] ± 65 | -6709[b] ± 73 | -6333[b] ± 78 | -6242[b] ± 88 |
| L | -2619[c] ± 50 | -1854[c] ± 47 | -2573[c] ± 56 | -2531[c] ± 58 |
| X | -97[d] ± 32 | -363[d] ± 57 | -177[d] ± 25 | -138[d] ± 41 |
| | -3712[e] ± 74 | -4492[e] ± 76 | -3584[e] ± 80 | -3573[e] ± 118 |
| 1 | -38404[a] ± 161 | -38796[a] ± 152 | -38173[a] ± 164 | -38089[a] ± 147 |
| I | -7332[b] ± 80 | -7438[b] ± 85 | -7437[b] ± 87 | -7088[b] ± 90 |
| 0 | -1534[c] ± 67 | -1948[c] ± 72 | -1562[c] ± 58 | -1533[c] ± 80 |
| S | -443[d] ± 36 | -220[d] ± 42 | -597[d] ± 40 | 3[d] ± 35 |
| | -5355[e] ± 83 | -5270[e] ± 113 | -5278[e] ± 101 | -5558[e] ± 110 |
| 1 | -18493[a] ± 128 | -18663[a] ± 113 | -18140[a] ± 135 | -18305[a] ± 137 |
| A | -3334[b] ± 71 | -3452[b] ± 51 | -3078[b] ± 54 | -3413[b] ± 52 |
| A | -671[c] ± 53 | -535[c] ± 40 | -924[c] ± 29 | -622[c] ± 40 |
| P | -523[d] ± 28 | -775[d] ± 42 | -249[d] ± 48 | -815[d] ± 35 |
| | -2140[e] ± 93 | -2141[e] ± 51 | -1904[e] ± 57 | -1975[e] ± 51 |
| 1 | -18353[a] ± 128 | -18478[a] ± 103 | -18281[a] ± 122 | -18315[a] ± 109 |
| I | -3252[b] ± 58 | -3299[b] ± 53 | -2943[b] ± 55 | -3060[b] ± 58 |
| G | -466[c] ± 35 | -323[c] ± 37 | -572[c] ± 46 | -598[c] ± 38 |
| D | -202[d] ± 45 | -591[d] ± 24 | -86[d] ± 24 | 23[d] ± 36 |
| | -2584[e] ± 65 | -2385[e] ± 56 | -2285[e] ± 84 | -2485[e] ± 66 |

[a] Total energy of the system including protein, ions and the N closest water molecules (see text)).

[b] Total energy of the protein, including [c] all the atom-atom interactions in the protein plus [d] the ion-protein interactions and [e] the protein-water interactions.

TABLE III: **Average fluctuations in the presence of explicit water molecules(Å)**

| **Seq** | **mainly $\alpha$ Str 1QLX** | **mainly $\beta$ Str 1I0S** | **disordered Str 1AAP** | **$\alpha/\beta$ Str 1IGD** |
|---|---|---|---|---|
| **1QLX** | **0.85 ± 0.06** | **1.26 ± 0.17** | **1.26 ± 0.22** | **1.16 ± 0.16** |
| **1I0S** | **1.15 ± 0.15** | **0.89 ± 0.09** | **0.98 ± 0.12** | **1.08 ± 0.16** |
| **1AAP** | **1.23 ± 0.13** | **1.20 ± 0.13** | **0.84 ± 0.07** | **1.08 ± 0.15** |
| **1IGD** | **1.17 ± 0.18** | **1.24 ± 0.20** | **1.16 ± 0.19** | **0.92 ± 0.12** |

structure during this extra heating procedure, with respect to the corresponding initial structure. The plots are organized per sequence, in the same order as in the previous figure and tables, and, in each plot, the solid line is for the native state. A general trend is that native structures take longer to deviate from the initial structure, something that is in agreement with the fact that they fluctuate less while at thermal equilibrium at 298 K. However, some of the decoys have a very similar behaviour to their corresponding native state. Indeed, figure 2 indicates that imposing part of the fold of 1I0S onto 1QLX leads to a decoy that holds out for as long as the native state of 1QLX (thick dotted line in the top plot of figure 2) and imposing the fold of 1AAP onto 1IGD leads to a decoy that holds out for as long as the native state of 1IGD (thick dotted line in the bottom plot of figure 2), during this heating procedure. Also, other general trends are that all structures show the same average structural stability until 50 ps (when the temperature has increased to 398 K) and that the greatest divergence with respect to the initial structure takes place at 100 ps for the native states and decoys of the smaller proteins 1AAP and 1IGD (when the temperature has increased to 498 K) and at 150 ps, when the temperature is 598 K, for the native states and decoys of the larger proteins 1QLX and 1I0S.

### B. Implicit water simulations

**The simulations in the previous section were performed with the explicit inclusion of water molecules which constitutes the most realistic model for a**

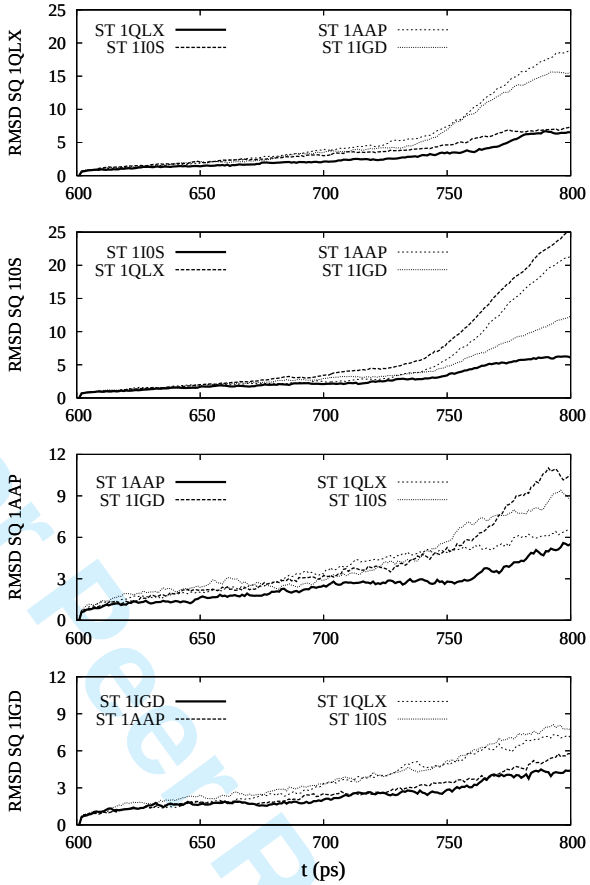



FIG. 2: Variation of the RMS deviation of each snapshot with respect to the corresponding initial structure. The temperature increased throughout the simulation at the rate of 2 K per ps. Each plot is for a different protein and the order from top to bottom is as in Figure 1. In each plot, the solid lines are for the native structures and the remaining three lines are for the decoys indicated by the keys. The values are in Å.

**protein in solution; however, the drawback of such simulations is that they are computationally very intensive because the system constituted by even a small protein in an explicit water bath is very large. In order to compensate for the consequent shortness of the explicit water simulations, up to 50 nanosecond simulations were performed in the absence of an explicit water bath. The AMBER force field ff99SB was selected for these simulations since it includes modified backbone torsion parameters to correct for a bias in favour of $\alpha$-helices that was detected for the AMBER ff99 force field[24]. In these simulations the effect of water on the protein is taken in a implicit manner with a generalized Born sol-**

vation model[7]. Eight simulations were done, two for each protein sequence, one being for its native fold and the second for one of the decoys explored previously.

In figure 3 are displayed the cumulative RMSD with respect to the initial structure of all the eight protein systems which show that all systems have achieved a good degree of convergence within the first 10 to 20 nanoseconds. In most cases, the decoys deviate more from the initial state than the native structures, which is to be expected since the native structures come from experimental data and are presumably much better "converged" to start with. However, it is very curious to note that the decoy for the 1QLX protein deviates much less from the initial structure than the corresponding native structure. Since the native structure of 1QLX is mainly $\alpha$ and the decoy simulated is mainly $\beta$ these results are in agreement with the studies that show that the ff99SB AMBER force field is not biased in favour of the alpha-helical secondary structure[24]. On the other hand, a very large deviation from the initial conformation is observed for the decoy of protein 1I0S, whose residues 1 to 104 have been made to have a mainly $\alpha$ fold, and whose residues 105 to 161 form a separate domain which has kept its original mainly $\beta$ fold. Close inspection of the structural changes shows that this large RMSD is due to the changes in the latter domain, which is relatively decoupled from the now mainly $\alpha$ domain of residues 1 to 104 and which unfolds completely leading to an unstructured structure, very different from its initial mainly $\beta$ fold.

In the previous section three measures were applied in order to differentiate the native structures from the decoys, namely, the average potential energy, the RMSD with respect to the average structure and the resistance to a heat pulse. In figure 4 the results for the first measure are displayed. Each point in figure 4 represents the average energy over 0.2 nanoseconds and the error bars are the standard deviations, that is, each point can be compared with values in the third line of the cells of table II. (The missing values in the figure have been lost due to hardware failure; in the case of the decoy of protein 1QLX, the missing first 20 nanoseconds have been compensated for by prolonging the simulation for

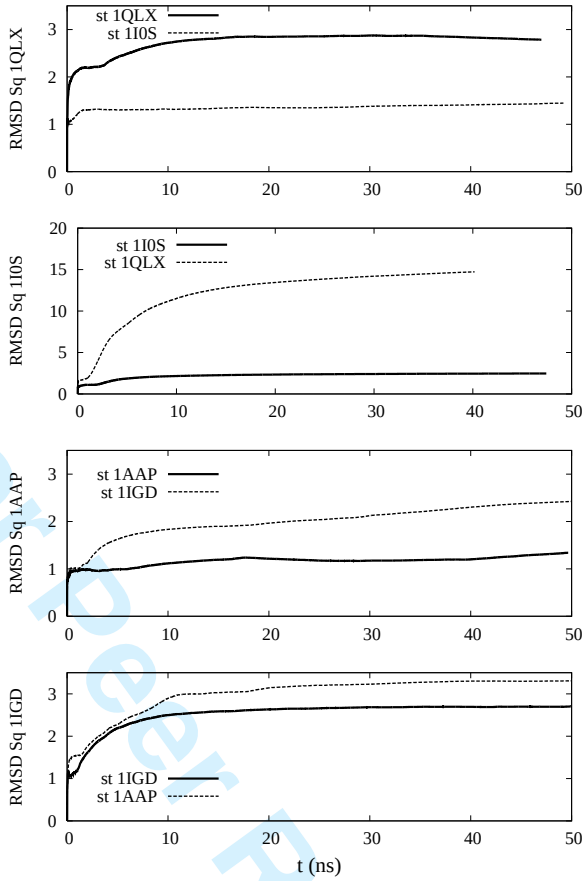



FIG. 3: Variation of the cumulative RMS deviation of the eight structures with respect to the initial structure. Each plot is for a different protein sequence and the order from top to bottom is as in Figure 1. In each plot, the solid lines are for the native structures and the dashed lines are for the decoys which can be identified by the keys. The values are in Å.

**another 20 nanoseconds). A first finding is that the average potential energies with the ff99SB AMBER force field[7,24] do not change appreciably throughout the 50 nanosecond simulation and a second finding is that the native structures have all consistently lower potential energies than the decoy, although this is distinctly so especially in the case of the mainly $\beta$ protein 1I0S. Indeed, for the proteins 1QLX and 1IGD the potential energies of the native structures and of the decoys have values that are relatively close and for the protein 1AAP the potential energies of the different conformers of the native structure and decoy actually fall within the same range.**

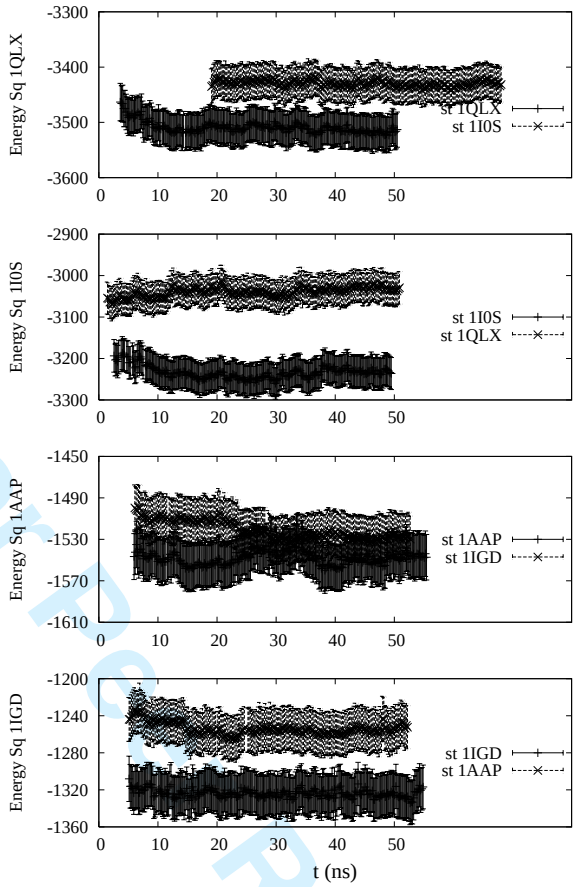



FIG. 4: Variation of the total potential energy of the eight structures with respect to time. The points are averages over 200 ps and the error bars represent the standard deviations in that interval. The values are in kcal/mol. Each plot is for a different protein sequence and the order from top to bottom is as in Figure 1. In each plot, the solid lines are for the native structures and the crosses lines are for the decoys which can be identified by the keys.

**The second measure used before to distinguish the native folds from the decoys are the values of the RMSD with respect to the average structure. In this case the average structures were computed with the snapshots stored during the 30 nanoseconds after the initial 20 nanoseconds, when the simulations have converged, according to figure 3. The top values in each cell of table IV shows the results. In order to compare with the data in table III values obtained in the last 0.2 nanoseconds of the 50 nanosecond simulation are also evaluated and displayed at the bottom of each cell of table IV.**

**As expected, the fluctuations, on average, are larger for the longer intervals.**

TABLE IV: **Average fluctuations in the absence of explicit water molecules(Å)**

| | **mainly $\alpha$** | **mainly $\beta$** | **disordered** | **$\alpha/\beta$** |
|---|---|---|---|---|
| | **Str 1QLX** | **Str 1I0S** | **Str 1AAP** | **Str 1IGD** |
| | **1.17 ± 0.24** | **1.00 ± 0.18** | | |
| | **0.77 ± 0.11** | **0.79 ± 0.09** | | |
| | **1.80 ± 0.24** | **0.94 ± 0.23** | | |
| | **1.20 ± 0.17** | **0.66 ± 0.08** | | |
| | | | **1.12 ± 0.24** | **1.27 ± 0.27** |
| | | | **0.71 ± 0.13** | **0.72 ± 0.14** |
| | | | **0.87 ± 0.16** | **0.87 ± 0.17** |
| | | | **0.78 ± 0.13** | **0.77 ± 0.12** |

The top values in each cell are the averages and standard deviations over 30 nanoseconds and the bottom values are the averages and standard deviations over 0.2 nanoseconds.

**On the other hand, the trend observed for the simulations with an explicit water bath according to which the native structures fluctuate less than the decoys, is not so clear in the simulations with implicit solvent since, on the one hand, both the native structure and the decoy of protein 1IGD have very similar degrees of fluctuation and since, on the other hand, there is one clear exception to that trend which is protein 1QLX, whose decoy actually fluctuates *less* than the native in the 30 nanosecond interval, while the decoys of proteins 1IGD and 1AAP fluctuate only marginally more than their corresponding native structure. Considering again that the degree of fluctuation is a measure of the entropy, the values in table IV, taken together with the results for the potential energies in figure 4, indicate that the native structures have generally lower free energies than the decoys, with the possible exception of proteins 1I0S and 1AAP.**

**The third measure to distinguish native folds from decoys was the resistance to a heat pulse. For this, the last structure sampled in the implicit solvent simulation was taken and heated from 298 K to 698 K, at a rate 25 times slower**

than previously, to allow for a more clear manifestation of the different relaxation times of the eight protein systems. Figure 5 shows that native structures are consistently more resistant to the heat pulse than the decoys, as was found in the explicit solvent simulations. However, while before the rule was that the larger structures took longer, that is, required higher temperatures, to deviate from the initial structure, the slower rate of the present heat pulse allows for a distinctive behaviour even among the different native structures, with the largest protein (1I0S, 161 amino acids), displaying a structural stability similar to that of a much smaller protein (1IGD, 61 amino acids) and with the human prion fragment (1AAP, 104 amino acids) being the most stable.

## IV. DISCUSSION

In this study four proteins were selected, each of which belongs to a different CATH class[20] and for each protein three decoys were built such that their folds and three dimensional structures were very different from the corresponding native fold, as can be assessed by the RMSD values that varied between 9.4 and 19.9 Å (see Table I). Three scoring measures were taken to pick out the decoys from the native states of the corresponding proteins, namely, the average potential energy at 298 K, the degree of fluctuation and the structural resistance to a heat pulse. Two sets of simulations were performed one with an explicit water bath but just 1 nanosecond, and a second set in which solvation was treated implicitly but with a duration of approximately 50 nanoseconds and for a smaller number of structures. The aim was to find the best simulation conditions and the best measures to distinguish the decoys from the native structures.

The difference in simulation length not withstanding we find, as do other authors[25], that the results obtained with explicit solvent were different from the results obtained with implicit solvent. Indeed, in the simulations with explicit solvent, the degree of fluctuation and the resistance to a heat pulse were clearly better measures than the average potential energy since only one decoy had an amplitude of fluctuation approximately equal to that of the corresponding

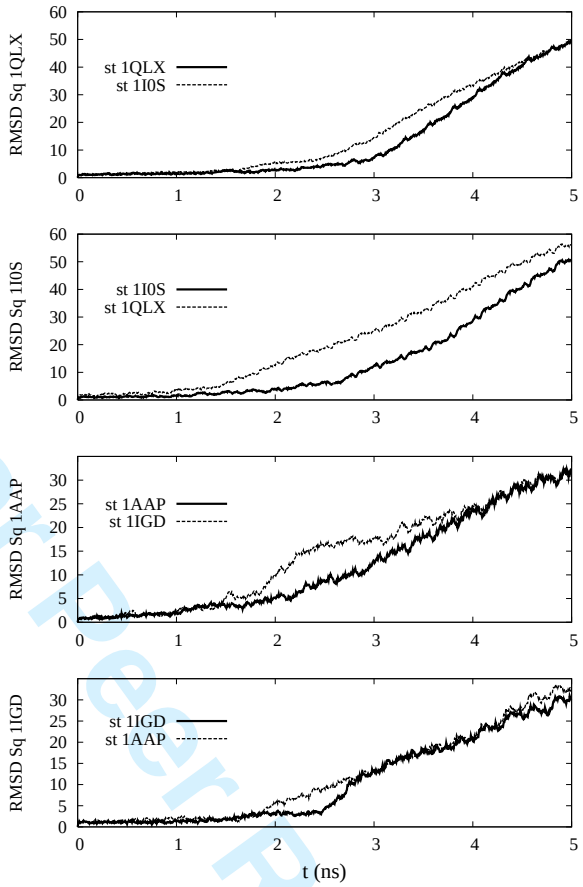



FIG. 5: Variation of the RMS deviation of each snapshot with respect to the corresponding initial structure. The temperature increased throughout the simulation at the rate of 2 K per 25 ps. Each plot is for a different protein and the order from top to bottom is as in Figure 1. In each plot, the solid lines are for the native structures and the dashed lines are for the decoys identified by the keys. The values are in Å.

**native state and only two decoys had a resistance to a heat pulse very similar to that of the corresponding native states. On the other hand, in the simulations with the AMBER ff99SB force field and implicit solvent[7,24] we find that the average potential energy is a good measure since its value for the native structures is generally lower than for the decoys, something that indicates the general validity of that force field for protein simulations. The degree of fluctuation of the protein, however, is less distinct in the implicit solvent simulations, being sometimes very similar for the decoys and the corresponding native structures, a result that can also be due to a more equilibrated statistical**

ensemble achieved by the longer implicit solvent simulations of all the eight protein systems tested. The overall conclusion is that of the three measures tested in this study, the resistance to a heat pulse is that which performs better as, in both sets of simulations, and for all proteins, it is able to pick out more consistently the native states from the decoys which tend to be destabilized faster (that is, at lower temperatures).

It is nevertheless curious that the decoys tested here, which are arguably the most unlikely structures that the respective sequences can adopt, cannot be more clearly distinguished from their native structures. There are studies, both theoretical[8–12,34–36] and experimental[26–30,32,33] that indicate that the native structure is not necessarily that which has the lowest free energy. Indeed, the first example was described in 1968 by Levinthal[1] who found two forms of an alkaline phosphatase at 317 K, one active and the other inactive, synthesized at different temperatures, in mutants of E. Coli. More recently, other proteins that can assume more than one structure in the same thermodynamical conditions have been found[28–30,32,33], the most notable of which are the prions[28,29]. Furthermore, since Levinthal first proposed that folding is achieved by a kinetic mechanism according to which proteins follow specific pathways[1], such a kinetic mechanism has been *directly* identified in a few cases[30,32,33] and, given that protein intermediates characterize these pathways, one can say that the experimental evidence for a kinetic mechanism for folding is indeed substantial[37,38] and may even include proteins that apparently follow a two-state process[38,39]. We think that the lack of a clear distinction between the decoys in this study and their native structures is in agreement with the latter studies and is an indication that the native structure is indeed a kinetic trap, rather than being defined by free energy minimization.



**Acknowledgments**

This work was funded by the Foundation for Science and Technology (FCT, Portugal) and by POCI 2010 and the European Community fund, FEDER. Many of the computer

simulations were performed at the Laboratory for Advanced Computing (LCA), University of Coimbra, Portugal. I would also like to thank one of the reviewers for inspiring a better way in which to present these results.

---

* Electronic address: lhansson@ualg.pt; URL: http://w3.ualg.pt/~lhansson

† Electronic address: pal@di.fct.unl.pt

[1] Levinthal, C., *J. Chim. Phys.* **65**, (1968) 44.

[2] Anfinsen, C.B., *Science* **181**, (1973) 223.

[3] Bryngelson, J.D., Onuchic J.N., Socci N.D. and Wolynes P.G., Proteins **21**, (1995) 167.

[4] Onuchic, J.N., Luthey-Schulten Z. and Wolynes P.G., *Ann. Rev. Phys. Chem.* **48**, (1997) 545.

[5] Wolynes, P.G., *Quart. Revs. Biophys.* **38**, (2005) 405.

[6] Karplus, M. and Kuriyan, J. *Proc. Natl. Acad. Sci. USA* **102**, (2005) 6679.

[7] Case D.A. *et al.* (1999) AMBER 6, University of California, San Francisco.

[8] Becker O.M. and Karplus M., *J. Chem. Phys.* **106**, (1997) 1495.

[9] Levy Y. and Becker O.M., *Phys. Rev. Lett.* **81**, (1998) 1126.

[10] Wales D.J., Miller M.A and Walsh T.R., *Nature* **394**, (1998) 758.

[11] Mortenson P.N. and Wales D.J., *J. Chem. Phys.* **114**, (2001) 6443.

[12] Krivov S.V. and Karplus M., *Proc. Natl. Acad. Sci. USA* **101**, (2004) 14766.

[13] Wroblewska L., Jagielska A. and Skolnick, J. *Biophys. J.* **94**, (2008) 3227.

[14] Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A. and Baker, D. *Proteins: Struc, Func & Gen* **52**, (2003) 76.

[15] Berman H.M. *et al.*, *Nucleic Acids Research* **28**, (2000) 235.

[16] Zahn R. *et al.*, *Proc. Natl. Acad. Sci. USA***97**, (2000) 145.

[17] Chiu H.-J., Johnson E., Schroder I. and Rees D.C., *Structure* **9**, (2001) 311.

[18] Hynes T.R. *et al.*, *Biochemistry* **29**, (1990) 10018.

[19] Gallagher T., Alexander P., Bryan P. and Gillilan G.L., *Biochemistry*, **33**, (1994) 4721.

[20] Orengo C.A. *et al.*, *Structure* **5**, (1997) 1093.

[21] All 16 minimum energy structures, four of which have the native fold and 12 of which are alternative, hybrid, structures are available, in PDB format, upon request.

[22] Phillips J.C. *et al.*, *J. Comp. Chem.* **26**, (2005) 1781.

[23] Humphrey W., Dalke A. and Schulten K., *J. Molec. Graphics* **14**, (1996) 33.

[24] Hornak V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. and Simmerling, C. (Simmerling, Carlos) *Proteins - Struc. Funct. and Bioinformatics* **65**, (2006), 712.

[25] Roe, D.R., Okur, A., Wickstrom, L., Hornak, V. and Simmerling, C. *J. Phys. Chem. B* **111** (2007) 1846.

[26] Mitra, R. K., Sinha, S.S. and Pal, S.K. *Langmuir* **23** (2007) 10224.

[27] Rodriguez-Larrea, D., Ibarra-Molero, B., de Maria, L., Borchert, T.V. and Sanchez-Ruiz, J.M. *Proteins* **70** (2008) 19.

[28] Prusiner S.B., *Science* **216**, (1982) 136.

[29] Prusiner S.B. and McCarty M., *Annu. Rev. Gen.* **40**, (2006) 25.

[30] Baker D., Sohl J.L. and Agard D.A., *Nature* **356**, (1992) 263.

[31] Rietveld, A.W.M. and Ferreira, S.T. *Biochemistry* **35** (1996) 7743.

[32] Sohl J.L., Jaswal S.S. and Agard D.A., *Nature* **395**, (1998) 817.

[33] Tsutsui Y., Liu L., Gershenson A. and Wintrode P.L., *Biochemistry* **45**, (2006) 6561.

[34] Cruzeiro-Hansson L. and Silva P.A.S., *J. Biol. Phys.***27**, (2001) S6.

[35] Cruzeiro, L., *J. Phys.: Condens. Matter* **17**, (2005) 7833.

[36] Cruzeiro, L., *J. Phys. Org. Chem.* **21**, (2008) 549.

[37] Englander S.W., *Annu. Revs. Biophys. Biomol. Struct.* **29**, (2000) 213.

[38] Brockwell D.J. and Radford S.E., *Curr. Op. Struct. Biol.* **17**, (2007) 30.

[39] Roder H. and Colón W., *Curr. Op. Struct. Biol.***7**, (1997) 15.