A molecular dynamics investigation of the kinetic bottlenecks of the hPin1 WW domain. I: simulations with the Sorenson/Head-Gordon model

FABIO CECCONI
Istituto dei Sistemi Complessi
(ISC-CNR)
(also INFM),
Via dei Taurini 19,
I-00185 Roma,
ITALY
Fabio.Cecconi@roma1.infn.it

CARLO GUARDIANI
Centro CSDC
(also INFN *)
Università di Firenze,
Via Sansone 1,
I-50019 Sesto Fiorentino,
ITALY
carlo.guardiani@unifi.it

ROBERTO LIVI
Dipartimento di Fisica
(also CSDC, INFN and INFM),
Università di Firenze,
Via Sansone 1,
I-50019 Sesto Fiorentino,
ITALY
Livi@fi.infn.it

Abstract: - The availability of a large amount of experimental data makes the Pin1 WW domain an ideal benchmark to test computational methods. The purpose of the present work is to identify the kinetic bottlenecks of the folding/unfolding pathway through Molecular Dynamics simulations. In this paper, the first of the series, we use the Sorenson/Head-Gordon model, based on the hydrophobicity properties of the chain residues. The unfolding simulation shows a highly cooperative mechanism that correctly identifies the contacts of loop I as the kinetic bottlenecks, in agreement with the Φ-value analysis performed by Gruebele *et al.*. The folding simulation, on the other hand, proves to be able to capture the essential topological features of the native fold even if the Kabsch distance from the PDB structure is still rather high. In the second paper of the series we report on simulations within the frame of the Go model, and the performances of the two models are compared and discussed.

Key-Words: - WW domains, Pin1 protein, kinetic bottlenecks, Go model, Sorenson/Head-Gordon model, Molecular Dynamics

1 Introduction

The WW domains are a family of fast-folding, compact, modular domains featuring a triple-stranded, antiparallel beta-sheet. In particular, the human Pin1 protein WW domain, due to the availability of a large amount of structural [1, 2], thermodynamical and kinetic [3] experimental data, represents an excellent benchmark to test computational methods.

The structure of this domain was resolved both through NMR [2] and X-ray diffraction [1] techniques. The protein is characterized by two hydrophobic clusters providing the largest contribution to the thermodynamic stability of the molecule [3]: cluster 1 (CL1) involves residues Leu7, Trp11, Tyr24 and Pro37; cluster two (CL2) comprizes Tyr23, Phe25 and Arg14. The stability of the molecule also derives from a network of hydrogen-bonds whose

*We gratefully acknowledge partial financial support from the INFN TO61 Experiment

central element is the highly conserved Asn26 located on strand β_2 and acting both as donor and acceptor in bonds with Pro9, Trp11, Ile28 and Thr29, thus linking strands β_1 and β_3. Another important feature of the domain, is the presence of two loops. Loop I (L1) plays a key role in substrate recognition [1] as it binds to the phosphate of the pS-P motif of the Proline-rich ligands. Moreover, the Φ-value analysis performed by Gruebele *et al.* [3] showed that the mutations of Ser16, Ser18 and Ser19 in loop I maximally destabilize the transition state, so that the formation of L1 appears to be the rate-limiting step in the folding/unfolding process. Loop II (L2), on the other hand, gives an important contribution to thermal stability, but it is involved in the formation of the transition state only at high temperatures [3]. Due to the ability of inducing conformational changes in Proline-rich, phosphorilated substrates, Pin1 is a potential regulator of the cell-cycle, and maybe involved in pathologies like Liddle's syndrome, muscular distrophy and Alzheimer's disease [4, 5]

The purpose of the present work is to identify the bottlenecks [6] in the folding process of WW domains through Molecular Dynamics simulations of thermal denaturation, using simplified protein models. The kinetic bottlenecks are related to the establishment of those specific interactions requiring the overcoming of large free-energy barriers. The formation of such interactions acts as a nucleus for the establishment of further contacts and accelerates the searching of the native state. The reliability of this method for identifying the critical amino-acids will be tested through a systematic comparison with the results collected by Gruebele *et al.* [3]. Two off-lattice, simplified models were used, namely the Go model [7] and the Sorenson/Head-Gordon (SHG) [8] model. Recent papers [9, 10] provide growing evidence that the Go model represents a useful tool for the characterization of the transition states. On the other hand, in this model the folding is driven by the native-state topology, so that the chemical properties of amino-acid residues are ignored and there is no *a priori* guarantee that the natural folding pathway will be followed. This motivates a detailed comparison of the performances of the Go and SHG models. The latter, in fact, being based on the hydrophobicity of amino-acid residues and on the secondary structural propensities of the dihedral angles, is more likely to reconstruct the natural sequence of events in the folding/unfolding pathways. Our simulations showed that the unfolding mechanisms reconstructed by the two models are consistent with each other and with the experimental data, even if the SHG model appears to be more reliable in the identification of kinetic bottlenecks. The SHG model, conversely, is less accurate in the prediction of the native structure. Yet it captured essential features of the native fold such as the overall topology, the hydrophobic clusters and more than 50 % of the native contacts. In the present paper we report on our unfolding simulations using the SHG model. In the second paper of the series [11], an account is given of the unfolding simulations using the Go model, and the performances of the two models are compared and discussed.

2 Methods

The SHG model is an off-lattice, minimal model describing the protein as a chain of beads of three flavours: hydrophobic (B), hydrophylic (L) and neutral (N) (see table in Ref [12]). The driving force responsible for the collapse onto a compact structure is the attraction between B-beads, whereas the repulsion between L and N beads determines the rearrangments of the compact structure into the native topology. The long-range interactions between residues which may be far apart in sequence space is modeled through the potential:

$$V_{LR} = \sum_{i,j \geq i+3} \epsilon_h S_1 \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - 2 S_2 \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \quad (1)$$

where $\epsilon_h = 2.1 Kcal\, mol^{-1}$ sets the energy scale and $\sigma = 4.0$ Å. The attractive interaction between hydrophobic residues is assured by setting $S_1 = S_2 = 1$ for BB pairs. A long-range repulsive potential describing the interaction between LL and LB pairs is obtained by setting $S_1 = 1/3$ and $S_2 = -1$. The interactions involving neutral residues are also repulsive and they are modeled as an excluded volume potential by setting $S_1 = 1$ and $S_2 = 0$.

The dihedral potential plays a key role in determining the secondary structure. Its analytical expression is as follows:

$$V_{dih} = \sum_{i=1}^{N-3}[A_i(1+\cos\phi_i) + B_i(1-\cos\phi) + C_i(1+\cos 3\phi_i) + D_i(1+\cos(\phi_i+\pi/4))]$$

where the coefficients A_i, B_i, C_i, D_i are chosen according to appropriate secondary structural propensities. Indeed, each dihedral, in the chain, is defined to be either helical (H: $A_i = 0$, $B_i = C_i = D_i = 1.2\epsilon_h$), extended (E: $A_i = 0.9\epsilon_h$, $C_i = 1.2\epsilon_h$, $B_i = D_i = 0$), or turn (T: $A_i = B_i = D_i = 0$, $C_i = 0.2\epsilon_h$). Based on the information stored in the PDB file 1NMV.pdb we chose the sequence: TTTTT-EEEEE-TTTTT-TEEEE-TTTTT-EETTT-EET. The SHG force-field is completed by two harmonic potentials with stiff constants that keep bond lenghts and bond angles approximately fixed to their equilibrium values. Being a minimal model, the SHG approach neglects important determinants of protein structure such as side-chain packing and hydrogen bonding. This leads to the opportunity of a sequence optimization procedure based on energy gap maximization [12, 13] to compensate the model limitations. Here, we used the sequence proposed in Ref. [12] LBBNN-BLBLB-NLNNN-LBBBB-LLNNL-BNBBL-LBNNL. We performed constant temperature MD simulations within the isokinetic scheme using reduced units. The temperature was measured in units of $\epsilon_h/R = 1070.96K$, time in units of $\tau = (M/\epsilon_h)^{1/2}\sigma = 4.44$ ps, ($\sigma = 4.0$ Å is the equilibrium length of Lennard-Jones interactions $M = 110$ Da), energy in units ϵ_h, specific heat in units $R = 1.9872 \times 10^{-3}\ Kcal\, mol^{-1}\, K^{-1}$ and the radius of gyration in units σ. The value of the energy scale ϵ_h was set to $2.13 Kcal\, mol^{-1}$ in order to reach a denaturation temperature compatible with experimental data $T = 332K$ [3]. The unfolding simulations started from a reference conformation (low temperature, $T = 0.01$) and we gradually heated the system to the value $T = 1$ in 50 temperature jumps. For each temperature we equilibrated the system over 6×10^6 time steps. Sampling of observables was performed for further 6×10^6 time steps of the dynamics. We monitored the departure from the reference state through the overlap Q, representing the fraction of native contacts still present. We also measured the parameters signalling the formation of the two hydrophobic clusters CL1 and CL2 typical of the WW domain

$$Q_{CLk} = \frac{\sum_{ij \in CLk} R_{ij}}{\sum_{ij \in CLk} r_{ij}}, k = 1, 2$$

where R_{ij} and r_{ij} are the native and current distances respectively, between residues i and j belonging to the same cluster. Small values of Q_{CLk} indicate that the cluster is ill-formed, because its residues are far apart. As a further reaction coordinate of the folding process, we also monitored the gyration radius and the $rmsd$ between the native and the reference conformations after the optimal superposition according to Kabsch's algorithm.

An accurate estimate of the density of states $\Omega(E, Q)$ was obtained through the weighted histogram method. The density of states were then used to compute energy and specific heat at each temperature. The knowledge of $\Omega(E, Q)$ can also be used to compute the probability $P_T(E, Q)$ that, at temperatute T, the protein is characterized by energy E and overlap Q. The sum of $P_T(E, Q)$ over all possible energies, yields the probability $P_T(Q)$ for the system to have a structural overlap Q at temperature T, which in turn, by reversing the Boltzmann's weight gives the potential of mean force along the reaction coordinate Q: $W_T(Q) = -RT\ln[P_T(Q)]$.

A detailed characterization of the folding/unfolding process was obtained by measuring the probability $P_{ij}(T)$ of native contact formation as a function of the temperature. The $P_{ij}(T)$ plots are typically characterized by a sigmoidal shape, keeping values close to 1 at low temperatures and decreasing to zero at high temperatures. The kinetic bottlenecks are those contacts whose $P_{ij}(T)$ plots exhibit an abrupt change in correspondence of the peaks and shoulders of the specific heat profile [6].

3 Results

The reference conformation denatured in the unfolding simulations was produced in a previous slow-cooling run, starting from a random coil structure of fragment 6-40 (chosen for the sake of consistency with the preliminary study reported in Ref [12]). The temperature was gradually reduced from $T = 1$ to $T = 10^{-2}$ in 50 steps each involving a thermalization stage of 6×10^6 time steps followed by a production stage of the same lenght, where the quantities of interest were sampled. After a final steepest-descent run we obtained a structure with a 4.74 Å $rmsd$ from the PDB conformation (Figures 1 and 2). The folded structure correctly displays the topology of a triple-stranded, antiparallel β-sheet, that however lacks the typical twist that in the PDB structure makes loop L2 almost perpendicular to loop L1. As a result, the folded structure is much more compact than the PDB conformation and shows a much larger number of native contacts (72 versus 41). The fact that 22 out of the 41 PDB contacts are also present in the folded structure is a clear indication of the good structural performance of the SHG simulation. The analysis of the contact pattern shows the importance of a stretch of four consecutive hydrophobic residues (Val22, Tyr23, Tyr24, Phe25) located on strand β_2 and thus lending themselves to the creation of a bridge between strands β_1 and β_3. It can be noticed that two of the mutations introduced in the sequence optimization procedure Lys13: L → B and Gln33: L → B are necessary for the formation of a network of β_1-β_2-β_3 hydrophobic contacts, representing the most important interactions stabilizing the structure produced by the folding algorithm. In the SHG model, thus, residues Tyr23, Tyr24, and Phe25 play the same role of the conserved Asn26 in the real protein, which lies at the center of a network of hydrogen bonds with Thr29, Ile28, Pro9 and Trp11. This observation sheds ligh t on the chemical foundations of the sequence optimization method, whose power basically relies on the possibility of replacing hydrogen bonds and salt bridges with effective hydrophobic interactions. A final criterion to assess the performance of the SHG protein model is to analyze its reliability to generate the two hydrophobic clusters CL1 and CL2 typical of the Pin1 WW domain. The values of the order parameters of the two clusters are $Q_{CL1} = 1.025$ and $Q_{CL2} = 1.046$: the values very close to 1 clearly show that the SHG simulation correctly forms both clusters. The PDB structure and the folded reference conformation are shown in Figures 1 and 2.

We now report on the analysis of thermodynamic and structural properties monitored during the unfolding simulations using the SHG model (see Figure 3 for the plots of energy and specific heat). All distances are measured with respect to the folded ref-

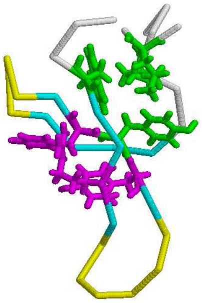

Figure 1: PDB structure of WW domain of Pin1 protein extracted from 1NMV.pdb file. The α-carbon trace and the side-chains of residues involved in the hydrophobic clusters are shown using the following color-code: β-strand residues, blue, loop residues, yellow, non-sheet residues not involved in hydrophobic clusters, white. Side-chains of residues participating in CL1 (Leu7, Trp11, Tyr24, Pro37) are shown in green with a stick representation, whereas those participating in CL2 (Arg14, Tyr23, Phe25) are represented through magenta sticks. The figure was drawn with the RASMOL program.

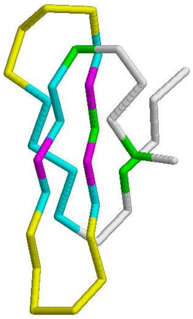

Figure 2: Reference Structure of the WW domain of Pin1 protein generated in the folding simulation using the SHG model. The α-carbon trace is shown using the following color-code: β-strand residues, blue, loop residues, yellow, non-sheet residues not involved in hydrophobic clusters, white. The residues involved in CL1 are shown in green, whereas those participating in CL2 are appear as magenta interruptions of the blue β-strand regions. The figure was drawn using the RASMOL program.

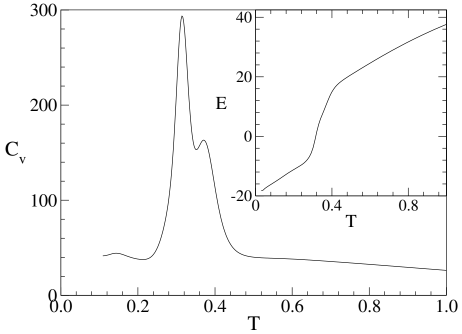

Figure 3: Specific heat as a function of temperature in the SHG unfolding. Inset: plot of energy versus temperature, both plots are obtained using the weighted histogram method (see Methods).

erence conformation. The specific heat plot (Fig 3) shows a sharp peak at $T = 0.31$ and a shoulder at $T = 0.37$. The existence of a sharp peak is a first indication of the cooperativity of the unfolding mechanism, where most contacts break down almost simultaneously in a narrow temperature range. As expected, the pattern of specific heat is consistent with that of the structural parameters, whose values change abruptly in the temperature range corresponding to the peak of the specific heat.

The sequence of breakdown events can be deduced by ranking the contacts according to temperature where the derivative of the sigmoidal $P_{ij}(T)$ plot is maximal. The unfolding pathway can be summarized as follows. At very low temperatures, before the peak of the specific heat, there occurs the disruption of contacts of cluster CL1. The next stage, corresponding to the increasing part of the specific heat peak, is characterized by the breakdown of β_2-β_3 contacts. There is also breakage of β_1-β_2 contacts in the regions of the strands most distant from loop L1. The peak of the specific heat plot, where the next step of the pathway takes place, is characterized by the contact disruption in the region of L1. This finding is important, because it is consistent with the experimental results of the Φ-value analysis by Gruebele *et al.* [3], which shows that the formation of L1 is the rate limiting step at physiological temperatures. The last stage of the unfolding process corresponds to the shoulder of the specific heat profile and involves the breakdown of $\beta_1 - \beta_2$ contacts in the intermediate portion of the strands and head-β_2 as well

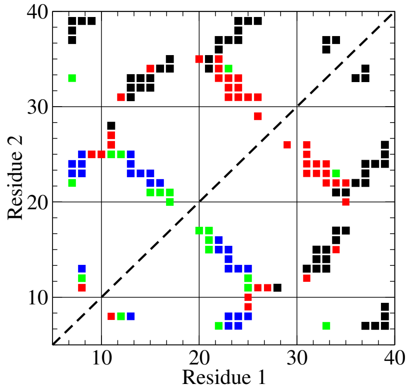

Figure 4: Color-coded contact map of the SHG model reference structure. The black squares represent those contacts broken at low temperature, before the peak of specific heat: these contacts connect β_1 to β_3 and β_2 to the $Tail$. The red squares indicate the contacts breaking down in correspondence of the increasing branch of the specific heat peak: the largest group of red squares refers to $\beta_2 - \beta_3$ contacts, while the smallest group refer to $\beta_1 - \beta_2$ contacts in the region of the strands most distant from loop L1. The green squares represent contacts disrupted at the top of the specific heat peak: the largest cluster of green squares represents the contacts of L1; finally, the blue squares represent the contacts broken at the shoulder of specific heat: the cluster on the left relates to $head - \beta_2$ contacts, while the cluster on the right relates to $\beta_1 - \beta_2$ contacts in the intermediate portion of the strands.

(we define *head* the N-terminal region Lys6-Gly10 and *tail* the C-terminal portion Asn36-Asn40). The unfolding pathway can be easily visualized by means of contact maps, where a color-code shows the contacts breaking down at different temperatures (Figure 4). The study of the SHG unfolding was completed by the analysis of the potential of mean force that represents an estimate of free energy. Around the unfolding temperature, the free energy profile shows a typical double well, whose most pronounced minima correspond to partially folded and unfolded structures. The double-well shape is the signature of an abrupt transition characterized by a massive rearrangement of the molecular structure.

4 Conclusions

We performed thermal unfolding simulations of the Pin1 protein WW domain with the purpose of identifying the kinetic bottlenecks in the folding/unfolding pathway. In this first paper of the series we report on the simulations using the simplified SHG model [8], where the folding is driven by the attractive interactions between hydrophobic residues and the secondary structures are formed by imposing an appropriate bias on the dihedral angles. The limitations of the model (for instance the lack of hydrogen-bonds) are compensated by a sequence optimization technique, based on energy gap maximization [12, 13].

The kinetic bottlenecks identified through the SHG unfolding simulation are located on loop L1. This finding is consistent with the Φ-value analysis performed by Gruebele [3]. The final structure produced by the simulation exhibits a rather high $rmsd$ from the PDB conformation, but it is characterized by the correct topology of a triple-stranded antiparallel β-sheet. Moreover, the folded conformation shares 22 of the 41 native contacts of the PDB structure, and it features well-formed hydrophobic clusters. Our work thus shows that simplified models, based on the physico-chemical properties of protein chains, are effective in reconstructing the folding/unfolding pathway and, in particular, in the identification of the kinetic bottlenecks, while still being able to capture the essential structural features of the native fold. In the second paper of this series the results of the SHG simulations will be compared with those of the Go model.

References:

[1] M.A. Verdecia, H.K. Huang, M.E. Bowmanm, K.P. Lu, T. Hunter, and J.P. Noel. Structural basis for phosphoserine-proline recognition by ww domains. *Nature Struct. Biol.*, 7:639–643, 2000.

[2] E. Bayer, S. Goettsch, J. W. Mueller, B. Griewel, E. Guiberman, L.M. Mayr, and P. Bayer. Structural analysis of the mitotic regulator hpin1 in solution: Insights into domain architecture and substrate binding. *J.Biol.Chem.*, 278:26183, 2003.

[3] M. Jaeger, H. Nguyen, J.C. Crane, J.W. Kelly, and M. Gruebele. The folding mechanism of a beta-sheet: the ww domain. *J. Mol. Biol.*, 311:373–393, 2001.

[4] L. Garnier, J.W. Wills, M.F. Verderame, and M. Sudol. Ww domains and retrovirus budding. *Nature*, 381:744–745, 1996.

[5] M. Sudol. Structure and function of the ww domain. *Prog. Biophys. Mol. Biol.*, 65:113–132, 1996.

[6] F. Cecconi, C. Micheletti, P. Carloni, and A. Maritan. The structural basis of antiviral drug resistance and role of folding pathways in hiv-1 protease. *Proteins Struct. Funct. Genet.*, 43:365–372, 2001.

[7] N. Go and H.A. Scheraga. On the use of classical statistical mechanics in the treatmentof polymer chain conformations. *Macromolecules*, 9:535–542, 1976.

[8] J.M. Sorenson and T. Head-Gordon. Matching simulation and experiment: a new simplified model for simulating protein folding. *J. Comp. Bio.*, 7:469–481, 2000.

[9] C. Micheletti, J.R. Banavar, A. Maritan, and F. Seno. Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett*, 82:3372–3375, 1999.

[10] O.V. Galzitskaya and A.V. Finkelstein. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA*, 96:11299–11304, 1999.

[11] F. Cecconi, C. Guardiani, and R. Livi. A molecular dynamics investigation of the kinetic bottlenecks of the hpin1 ww domain: simulations with the go model. Submitted to the 2006 WSEAS International Conference on *Mathematical Biology and Ecology*, Miami, Florida, USA, January 18-20, 2006.

[12] S. Brown, J. Fawzi, and T. Head-Gordon. Coarse-grained sequences for protein folding and design. *Proc. Natl. Acad. Sci. USA*, 100:10712–10717, 2003.

[13] J.M. Sorenson and T. Head-Gordon. Redesigning the hydrophobic core of a model β-sheet protein: destabilizing traps through a threading approach. *Proteins*, 37:582–591, 1999.